Comparative analysis of the insect mobile genetic element repertoire and its influence on genome size dynamics

A DISSERTATION PRESENTED
BY

MALTE PETERSEN

FROM
HAMBURG-BERGEDORF
TO
THE FACULTY OF MATHEMATICS AND NATURAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DOCTORAL DEGREE OF
DOCTOR RERUM NATURALIUM (DR. RER. NAT.)

RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
BONN, GERMANY
2019

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

Erstgutachter: Prof. Dr. Bernhard Misof
Zweitgutachter: PD Dr. Lars Podsiadlowski
Fachnahes Kommissionsmitglied: Prof. Dr. Dietmar Quandt
Fachfremdes Kommissionsmitglied: PD Dr. Torsten Wappler

Tag der Promotion: 26. August 2019
Jahr der Veröffentlichung: 2020

Thesis advisor: Bernhard Misof Malte Petersen

Comparative analysis of the insect mobile genetic element repertoire and its influence on genome size dynamics

Abstract

This thesis presents comparative genomics studies in insects as well as bioinformatics software development. Its empirical research part is focused mainly on mobile genetic elements, also termed transposable elements. The data basis contains datasets from public repositories, a rich and often underexplored source of information on genomic biodiversity. Transposable elements in particular are often neglected when the results of a genome sequencing study are published, although they make up a major part of virtually every eukaryotic genome.

After a general introduction in Chapter 1, I characterize and compare the transposable element repertoire of 73 arthropod species in Chapter 2 and find that it correlates to genome size in both abundance and diversity. In Chapter 3, I study the effect of transposable elements on the evolution of genome size in more detail and on an expanded dataset of 96 species. In Chapter 4, I present a software pipeline for delineating orthology among coding nucleotide sequences, an essential tool for many comparative and phylogenetic studies. Finally, Chapter 5 is a general conclusion.

Chapter 2

Transposable elements (TEs) are a major component of metazoan genomes and are associated with a variety of mechanisms that shape genome architecture and evolution. Despite the ever-growing number of insect genomes sequenced to date, our understanding of the diversity and evolution of insect TEs remains poor. Here, we present a standardized characterization and an order-level comparison of arthropod TE repertoires, encompassing 62 insect and 11 outgroup species. The insect TE repertoire contains TEs of almost every class previously described, and in some cases even TEs previously reported only from vertebrates and plants. Additionally, we identified a large fraction of unclassifiable TEs. We found high variation in TE content, ranging from less than 6 % in the antarctic midge (Diptera), the honey bee and the turnip sawfly (Hymenoptera) to more than 58 % in the malaria mosquito (Diptera) and the migratory locust (Orthoptera), and a possible relationship between the content and diversity of TEs and the genome size. While most insect orders exhibit a characteristic TE composition, we also observed intraordinal differences, e.g., in Diptera, Hymenoptera, and Hemiptera. Our findings

shed light on common patterns and reveal lineage-specific differences in content and evolution of TEs in insects. We anticipate our study to provide the basis for future comparative research on the insect TE repertoire.

Chapter 3

Genome size in insects displays inter-specific variation in excess of 130-fold, a range only paralleled in the metazoan phylum by amphibians. In general, these inter-specific differences seem to be best explained by differential rates of transposable element (TE) accumulation. In fact, we observe that TE accumulation rates are lineage-specific and that major insect clades have distinct TE age distributions. Given this observation, we hypothesize that evolutionarily younger insect lineages should have more TEs that are older than the insect lineage itself. To test this hypothesis, we infer ancient and lineage-specific TE insertions, and quantify genome size increase and decrease in 96 arthropod species from 18 major insect orders, spanning a geological age range of around 400 million years. Our analysis reveals that most insect lineages appear to have a specific rate of TE accumulation that is correlated with genome size, along with a distinct, clade-specific and TE class dependent TE age distribution. Additionally, lineage-specific rates of genome size reduction appear to counteract genome expansion through TE activity. Our results are inconsistent with a general "accordion" model of genome size dynamics in eukaryotes, therefore we suggest that TE management in insects is fundamentally different than in vertebrates. We propose that in the face of burst-like TE proliferation events, clade-specific rates of genome size reduction strongly influence the large variation in extant insect genome sizes.

Chapter 4

Orthology characterizes genes of different organisms that arose from a single ancestral gene via speciation, in contrast to paralogy, which is assigned to genes that arose via gene duplication. An accurate orthology assignment is a crucial step for comparative genomic studies. Orthologous genes in two organisms can be identified by applying a so-called reciprocal search strategy, given that complete information of the organisms' gene repertoire is available. In many investigations, however, only a fraction of the gene content of the organisms under study is examined (e.g., RNA sequencing). Here, identification of orthologous nucleotide or amino acid sequences can be achieved using a graph-based approach that maps nucleotide sequences to genes of known orthology. Existing implementations of this approach, however, suffer from algorithmic issues that may cause problems in downstream analyses.

We present a new software pipeline, Orthograph, that addresses and solves the above problems and implements useful features for a wide range of comparative genomic

and transcriptomic analyses. Orthograph applies a best reciprocal hit search strategy using profile hidden Markov models and maps nucleotide sequences to the globally best matching cluster of orthologous genes, thus enabling researchers to conveniently and reliably delineate orthologs and paralogs from transcriptomic and genomic sequence data. We demonstrate the performance of our approach on *de novo*-sequenced and assembled transcript libraries of 24 species of apoid wasps (Hymenoptera: Aculeata) as well as on published genomic datasets.

With Orthograph, we implemented a best reciprocal hit approach to reference-based orthology prediction for coding nucleotide sequences such as RNAseq data. Orthograph is flexible, easy to use, open source and freely available at `https://mptrsen.github.io/Orthograph`. Additionally, we release 24 *de novo*-sequenced and assembled transcript libraries of apoid wasp species.

Contents

List of Figures

List of Tables

To Detlef,
who died too soon,
and his unnamed successor,
who also died too soon.

Acknowledgments

I AM GRATEFUL to Bernhard Misof for the opportunity to undertake this PhD project as well as for his input and support over the five years this project took to complete. I would not have gotten this far without the discussions, insight, and motivation from you.

Likewise to Lars Podsiadlowski for his readiness to co-advise this thesis and many helpful discussions. I always appreciated your positivity.

I also thank Dietmar Quandt as well as Torsten Wappler for their readiness to join my commission and their time reading this thesis.

I am grateful to Christoph Mayer and Oliver Niehuis for many helpful discussions, their careful insight, and humor.

I appreciate the financial support of the Alexander Koenig Gesellschaft, which enabled me to attend several conferences, the Leibniz Association for funding the graduate school on genomic biodiversity research, and the German Research Foundation, which funded my position for the second half of my PhD project.

Thank you to these ZFMK people in no particular order for relaxed lunch breaks, regular beach volleyball sessions, playing squash, climbing walls, discussions or ice cream or coffee, helpful input and output, being an outlet when I needed to vent, satisfying my gaming drive, contributing to an exceptionally kind and welcoming atmosphere, or simply for being a nice person: Peter Grobe, Dirk Ahrens, Hans-Joachim Krammer, Matthias Geiger, Jan Philip Øyen, Julia Schwarzer, Sebastian Martin, Alexandros Vasilikopoulos, Panagiotis Provataris, Jonas Astrin, Claudia Etzbauer, Ameli Kirse, Victoria Moris, Karsten Stehr, Karoline Mauer, Juliane Romahn, Simon Käfer, Sofia Paraskevopoulou, Thomas Gerken, Tanja Ziesmann, Sandra Meid, Hamideh Fard, Dirk Rohwedder, and my great friends Jonas Eberle and Stefanie Heufelder. I apologize to whomever I may have forgotten here.

Gaby Nottebrock for being an awesome Master student under my supervision. You really made your project glow. Go boldly.

No thanks to Jeanne Wilbrandt and myself for wasting a year of both our time. It was beautiful while it lasted, though, and I learned a lot from you.

My family has always backed me unconditionally in all endeavours during my career. I can never return this. Thank you.

Finally, thanks to my wife Hannah for taking care of everything in the wrap-up stages of this work and enduring my absent-mindedness and lack of energy for anything else. Without you, I would have emerged a different person.

1

General Introduction

The genome, that is, the entirety of DNA of an organism is a composition of different functional complexes. It does not only contain genes, which are transcribed to messenger RNA and translated into the proteins that make up cells and, ultimately, all organisms; in fact, the human gene repertoire of around 23,000 genes makes up only around 2 % of the human genome (Makałowski, 2001) (Figure 1.1). More prominent components of the human genome include introns (non-coding sections of genes, around 26 %), but by far the most voluminous

chunk consists of repetitive elements: DNA segments that occur in sometimes many copies throughout the genome. More than half of the three billion base pairs (Gbp) of the human genome (52 %) is occupied by repetitive elements (Lander et al., 2001; de Koning et al., 2011). The major part of these repetitive elements in the human genome, also called repeats, is formed by transposable elements (45 % of the genome).

Transposable elements (TEs) are also known as "jumping genes" or "parasitic DNA". They were discovered in the 1940s by their defining property, the capability of movement within the genome (McClintock, 1950). By duplicating themselves through various mechanisms that depend on the TE type, TEs can reach copy numbers in the thousands (Petersen et al., 2019) and, like in the human genome (Figure 1.1), be a major contributor to the genome size. This genome "inflation" effect due to TE proliferation has been observed throughout eukaryotes in general (Chénais et al., 2012), and reiterated in vertebrates (Chalopin et al., 2015), arthropods (Petersen et al., 2019), and plants (Staton & Burke, 2015). In contrast, there are species with small genomes that carry a small TE load. This has been observed in plants (Ibarra-Laclette et al., 2013), nematodes (Burke et al., 2015), and insects (Kelley et al., 2014).

The genomes of mammals, such as human, and birds exhibit much less variation in size than, for example, the genomes of arthropods or amphibians (Gregory, 2005). In mammals, genome size varies around five-fold and in birds even only around two-fold, whereas in insects, the spread is around 240-fold (Figure 1.2, Table 1.1 on page 5). This immense variation surpasses that of amphibians, where some species have huge genomes of up to 118 Gbp, and is paralleled only by the group of bony fishes (Osteichthyes, excluding lungfishes), which exhibit a genome size spread of around 220-fold. Before the discovery of TEs and non-coding DNA, such as introns,

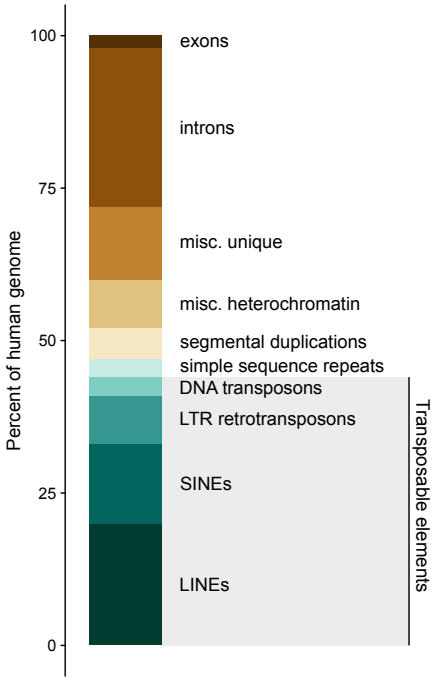

Figure 1.1: Composition of the human genome. Almost half of the three billion base pairs in the genome is attributed to transposable elements of various classes (DNA transposons, LTR retrotransposons, LINEs, SINEs). Data source: Lander et al. (2001)

in the genome, it was assumed that genome size should correlate with perceived organismic complexity, but the fact that amoeba have genomes with up to a staggering 670 Gbp (Parfrey et al., 2008) did not fit well with that assumption. This apparent contradiction was named the "C-value paradox" and later renamed to "C-value enigma" (Gregory, 2007), as still a connection between genome size and organismic complexity appears absent.

No matter the size, the genome needs to be maintained: repair mechanisms and transcription machinery as well as error correction use energy. The transcription and translation error rate increases with genome size (Zaher & Green, 2009), making more repairs necessary. Larger genome size has been linked to decreased development rate (White & McLaren, 2000) and increased oxygen requirements (Vinogradov, 1997; Gregory, 2002). In plants (Grime, 1983), invertebrates (Gregory, 2005), and vertebrates (Horner & Macgregor, 1983; Olmo & Odierna, 1982; Gregory, 2000), it has been shown that cell size increases with genome size (Dufresne & Jeffery, 2011).

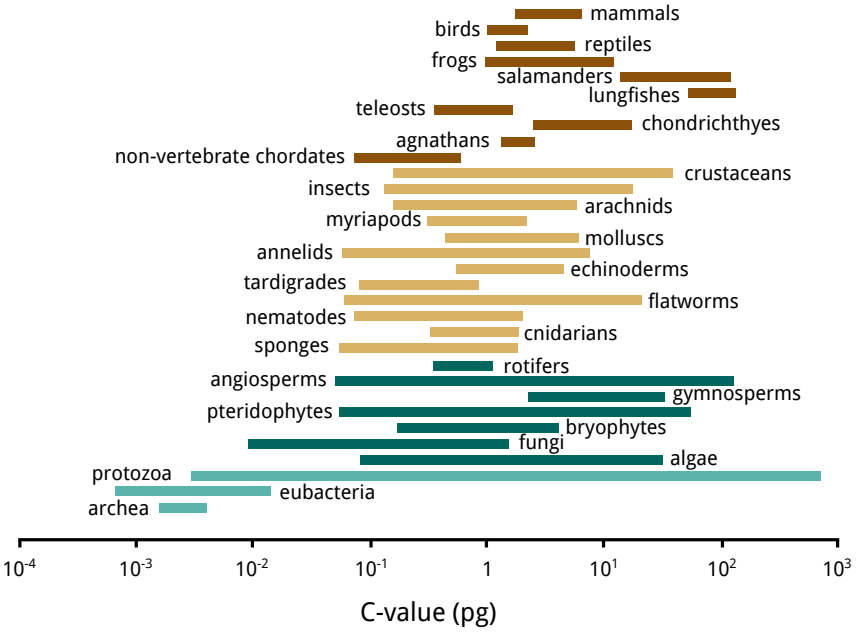

Figure 1.2: Genome size spread in eukaryotes and prokaryotes. The C-value is the amount of haploid nuclear DNA in picogram (pg); one pg DNA is approximately 978 Mbp. Colors are ordered for chordate animals (brown), invertebrate animals (yellow), plants, fungi and algae (green), and prokaryotes (teal). Figure modified from Gregory (2004).

Larger cells are also less efficient to maintain and proliferate, and they divide more slowly (Bennett et al., 1977). In summary, a large genome comes with a cost. What would be the benefit of having a large genome?

The more genetic material in the genome, the higher the likelihood of random mutations (Wielgoss et al., 2011). Mutations provide the basis for genotypic evolution: through natural selection, deleterious changes to the genome will be removed from populations over time, and beneficial changes — or innovations — prevail. A higher mutation rate is not always a negative property: it also brings with it a higher rate of beneficial mutations. Therefore, the genome is thought to reach an equilibrium between the incurred metabolic cost of sustaining a high

Table 1.1: Genome size spread in Metazoa. Values are in picogram DNA; one pg is approx. 978 mega-basepairs (Mbp). Data from the Genome Size Database (Gregory, 2018), `http://www.genomesize.com`, accessed 2018-05-07.

Phylum	Subphylum	Class	n	min	max	Δfold
Annelida		Oligochaeta	35	0.43	7.64	17.77
Annelida		Polychaeta	100	0.06	7.2	120
Arthropoda	Chelicerata	Arachnida	148	0.08	7.5	93.75
Arthropoda	Crustacea	Branchiopoda	68	0.16	2.91	18.19
Arthropoda	Crustacea	Copepoda	73	0.14	14.68	104.86
Arthropoda	Crustacea	Malacostraca	241	0.68	64.62	95.03
Arthropoda	Hexapoda	Insecta	1353	0.07	16.93	241.86
Chordata	Vertebrata	Amphibia	932	0.95	120.6	126.95
Chordata	Vertebrata	Aves	903	0.91	2.16	2.37
Chordata	Vertebrata	Chondrichthyes	199	1.51	17.05	11.29
Chordata	Vertebrata	Mammalia	816	1.63	8.4	5.15
Chordata	Vertebrata	Osteichthyes	1909	0.34	74.86	220.18
Chordata	Vertebrata	Reptilia	423	1.05	5.44	5.18
Mollusca		Bivalvia	108	0.65	5.4	8.31
Mollusca		Gastropoda	149	0.43	7.85	18.26
Nematoda		Secernentea	72	0.02	2.5	125

mutation rate (*i.e.*, the damage caused by deleterious mutations), and the cost of mechanisms that reduce the mutation rate (Bernstein et al., 1987; Altenberg, 2011).

1.1 Impact of transposable elements on the genome

The presence and activity of TEs can have disruptive influence on the genome architecture. By inserting at critical positions, TEs can disable genes (Kazazian et al., 1988). An insertion in regulatory sequence can change gene expression (Warnefors et al., 2010). TEs, by way of their repetitive nature, provide hotspots for ectopic (non-homologous) recombination (Lim, 1988; Gray, 2000; Fiston-Lavier et al., 2007), thus increasing the likelihood for segmental duplications, deletions, and inversions (Mathiopoulos et al., 1998; Remnant et al., 2013). On the one hand, TEs are obviously a source of potentially deleterious mutations. On the other hand, TEs can be "do-

mesticated" and genes exapted from TE sequence (Gahan, 2001; Daborn et al., 2002; Aminetzach et al., 2005; Chen & Li, 2007), conferring novel functions to the host. Such innovations can happen within a few hundred generations (Dolgin & Charlesworth, 2006; Struchiner et al., 2009; Kofler et al., 2015). As a famous example, the melanism in the British peppered moth — in which camouflage evolved that matches the birch trees blackened as a result of industrialisation — is caused by TEs (van't Hof et al., 2016). These observations document that TE activity can also have beneficial effects on the host genome (especially in times of stress (Chénais et al., 2012)), and should therefore not be entirely subdued.

To keep the TE population in check, defenses that remove or silence TEs have developed in host organisms. In many groups of organisms, a multi-layered network of epigenetic regulation mechanisms evolved in place to prevent TE activity at both the pre- and post-transcriptional stage. In plants, an epigenetic modification called DNA methylation prevents TEs from being transcribed and thus from transposing (Slotkin & Martienssen, 2007; Lisch, 2009). After transcription, proteins from the RNA interference (RNAi) pathway can disable messenger RNA and thereby silence TEs (Buchon & Vaury, 2006). Similarly, a class of non-coding RNA, so-called Piwi-interacting RNA (piRNA) protect the integrity of the genome, in particular in germline cells, by forming a complex with Piwi proteins, which can bind and cleave RNA (Zeng et al., 2011). This complex can recognize and silence target TEs in the RNA stage (Siomi et al., 2011; Mondal et al., 2018). Similar systems were identified in vertebrate genomes (Suzuki & Bird, 2008; Schübeler, 2015): DNA methylation is thought to be a genome defense mechanism in mammals as well (Yoder et al., 1997). Interestingly, vertebrate genomes are globally methylated, and in plant genomes, only gene bodies and TEs are methylated (Suzuki & Bird, 2008). Fun-

gal genomes exhibit an even more mosaic-like methylation pattern: here, only TEs are methylated and genes are not. In invertebrates, TEs tend to be unmethylated. The fruit fly *Drosophila melanogaster* does not even have the methyltransferase enzyme in its gene repertoire. Likewise, some butterfly species have lost RNAi pathway genes (Pauli et al., 2016). Thus, these genome defenses appear to be modular and complementary to one another. They are effective to a certain extent: permanently inactive TEs become genomic "cruft" and are degraded by random mutations over time and genetic drift like other parts of the genome that are not subject to selection (Szitenberg et al., 2016). As a result of these extensive silencing techniques, it is not surprising that most of the TE population in extant genomes is inactive (Yoder et al., 1997; Zilberman et al., 2007).

There are two major models to explain TE population dynamics in the genome: the equilibrium model and the burst model (Petrov et al., 2011; Kofler et al., 2012; Cridland et al., 2013; Blumenstiel et al., 2014). In the equilibrium model, the TE insertion rate is assumed to be more or less constant, and TEs are silenced and removed by purifying selection at a likewise constant rate (Charlesworth & Charlesworth, 1983). This way, TE insertion rate and DNA removal rate would cancel each other out, and the genome size remains stable. The equilibrium model provides a better fit for TE dynamics under the effects of purifying selection (Barrón et al., 2014) than the transposition burst model. The burst model, which is also termed the non-equilibrium model, predicts that TEs undergo periods of high transposition activity while otherwise proliferating at a constant but lower rate. Under the transposition burst model, the absence of a correlation between the TE age and frequency would be expected, which better explains the observed TE age distribution in insect genomes as well as the large genome size fluctuations during

insect evolution, given that TE abundance is a predictor for genome size (Alfsnes et al., 2017; Petersen et al., 2019).

1.2 Insects, phylogenetics, and comparative genomics

Insects are among the most speciose groups of organisms on earth (Figure 1.3 on page 9) and, since their appearance approximately 480 million years ago (Mya) (Misof et al., 2014), have conquered land, freshwater, and air (but not saltwater). Protected by their hard exoskeleton, insect representatives have invaded virtually all conceivable ecosystems including human habitations (Bertone et al., 2016). Insects are immensely diverse in morphology (Grimaldi & Engel, 2005) and often highly specialized towards a specific food source, habitat, or lifestyle. Bees, wasps, ants, and termites, for example, form eusocial communities with a complex caste system. As disease vectors, mosquitoes are responsible for more human deaths than all other animals combined (WHO, 2017; Linnell, 2011; Lamarque, 2009; De Maddalena et al., 2008; Kasturiratne et al., 2008; Packer et al., 2005). Beetles, cicadas, and grasshoppers are examples for an important source of food for livestock and humans alike as well as a pest with high economic impact (Oliveira et al., 2014). Obviously, insects play pivotal roles in most ecosystems of the planet. Insect population diversity and abundance, however, is declining (Vogel, 2017) as a result of widespread human influence (see below), with disastrous reverberations at all levels of the local food chains. In order to mount efficient conservation efforts, a thorough understanding of insect biology is required.

Despite their mega-diversity and ecological importance, insects are astonishingly understudied on the genomic level compared to other animals: As of 2018-07-06, there were 1,115 pub-

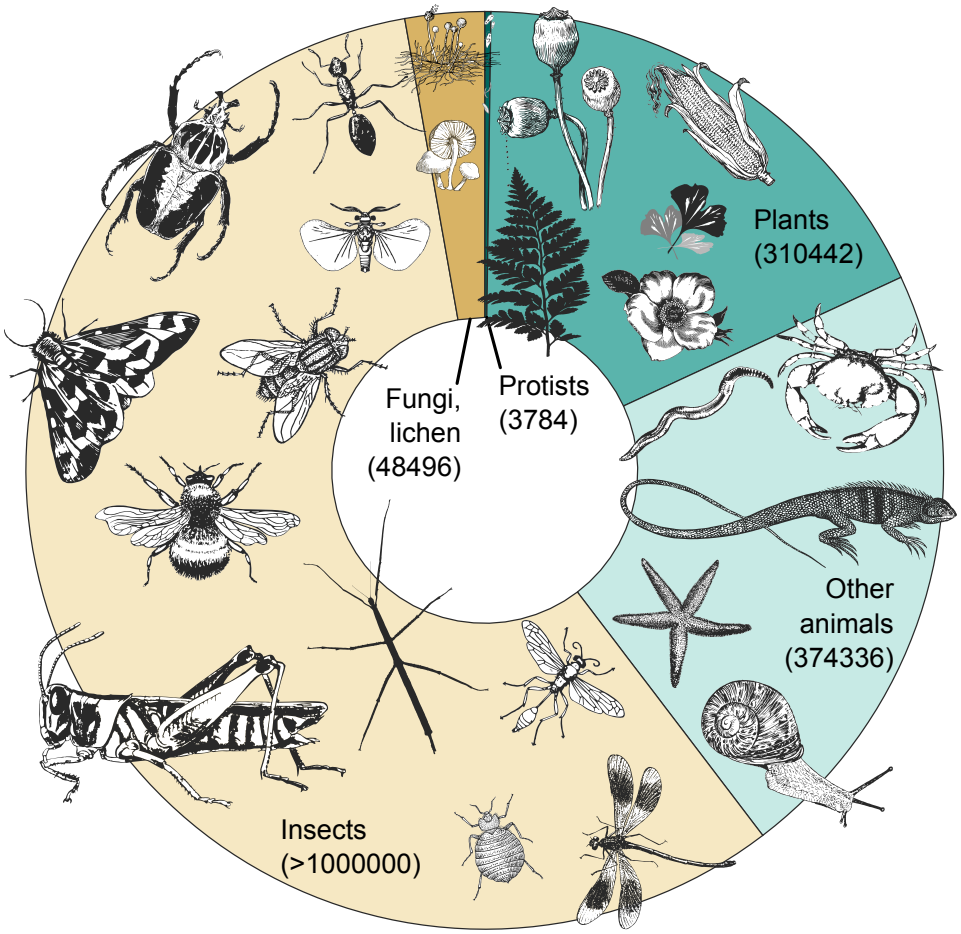

Figure 1.3: Insects contribute more than half of the currently known species diversity. Data source: IUCN (2018); pictures source: Pixabay, available under the Creative Commons Public Domain license (CC0).

lished non-insect animal genome sequences in the NCBI database (O'Leary et al., 2016) (about 0.3 % of the total non-insect species diversity; Figure 1.3) and 493 insect genomes (0.065 %; about five times less genomes per species). That number is growing by about 50 genomes per year (O'Leary et al., 2016), however at this rate, it will take more than 150 years to even sequence the genomes of 10 % of the insect biodiversity. During this time, many species will have gone extinct — as it stands, our genome sequencing efforts are losing the race against man-made biodiversity loss. Ever-improving, massively parallel sequencing technologies that appeared during the early

2000's (Behjati & Tarpey, 2013) have accelerated the pace and accuracy at which genomes can be sequenced, but it will not be enough. Of many insect species, we will never be able to obtain the genomic source code, and this will hamper our understanding of the roles these species occupied in the interaction network of their habitats.

This not only affects insects, but all kingdoms of life: we are currently experiencing a mass extinction event that parallels other episodes in earth's history with high rates of biodiversity decline (Pimm et al., 1995; Dirzo & Raven, 2003; Schipper et al., 2008; Barnosky et al., 2011; Dirzo et al., 2014). Other than the five previous major extinction events (Kolbert, 2014), it is anthropogenic in origin (Leakey & Lewin, 1996; Ceballos et al., 2015) and is associated with global warming (Cook et al., 2016; Wuebbles et al., 2017), large-scale deforestation (Wright, 2005), destruction of marine and freshwater habitats (Burkhead, 2012), and introduction of invasive species (Mooney & Cleland, 2001), all hallmarks of human influence. Put shortly, the rate at which species go extinct is alarming (Newbold et al., 2016; Ceballos et al., 2017; Hallmann et al., 2017), and our children will likely experience a world with less than half the biodiversity we know today. While this issue has raised the attention of country leaders and conservation policies are being put in place worldwide (Puntaru, 2017), this might not be enough to reverse the trend without sustaining irreparable damage to the ecosystems of the planet. To make matters worse, there are signs that the issue, despite its urgency, is fading from public awareness (Kusmanoff et al., 2017).

We cannot save what we do not know. Thus, conservation efforts require intimate knowledge of the systems they aim to preserve. The road towards understanding the biology and the interaction of species is, however, travelled on multiple levels. It is not enough to observe the

behaviour or the ecology of an animal to understand the impact of it being removed from its habitat. It is also not enough to describe functional morphology to gain insight on ecological implications. Neither is it sufficient to analyze the genes and draw conclusions based on their composition and structure. Profound understanding of any system can only be gained by studying it from multiple angles and with interdisciplinary approaches. One approach that can add to the knowledge about a species is to sequence and analyze its genome: the "source code of life" that defines, by a manifold of means, its appearance, features, behaviour and interactions with the environment. By comparing the composition, the functional networks, and other properties of one species' genome to that of other species, one can gain insights on the mechanisms of evolution that led to the mega-diversity of today's insects.

Comparative genomics studies — that is, investigations comparing genomic features of more than one species — are not only limited by the fact that for many species, there is no genomic sequence information available. Additionally, comparative analyses have to take the evolutionary history of the species into consideration (Dunn et al., 2018), usually in the form of a phylogenetic tree that conveys information on the species' relationships. An undisputed phylogenetic tree down to family level does not exist for insects so far. Misof et al. (2014) and the 1KITE project (`http://1kite.org`) have inferred a robust backbone phylogeny for most major insect orders from transcriptomic data (see below), and several publications have presented order-level phylogenies, for example for Coleoptera (McKenna et al., 2015), Hymenoptera (Peters et al., 2017; Branstetter et al., 2017), Lepidoptera (Breinholt et al., 2018), and Hemiptera (Johnson et al., 2018). However, accurately reconstructing species ancestry remains a challenge that has obstructed reliable comparative analyses in insects so far.

Among the approaches to infer a phylogenetic species tree, the most informed, because based on matrices with data points numbering in the millions, is reconstruction from amino acid or nucleotide sequences. Phylogenetic species tree reconstruction, however, regardless the method, is no simple task and relies on sophisticated methods on all stages of the analysis (proposed by Misof et al. (2014), for example). Perhaps most of all, selection of the phylogenetic markers (the features that distinguish species and define their level of relatedness) is paramount. Most modern genomic studies use single-copy genes that are found in all (or almost all) species. The implied assumption is that these genes share a common ancestry and are related via speciation events, that is, they are *orthologous* to one another (Koonin, 2005) and their phylogeny reflects the species phylogeny.

Several commonly used methods to identify orthologous genes rely on clustering nucleotide or amino acid sequences based on their similarity (Chen et al., 2007). Since orthologs tend to be more similar to each other than to all other genes in the genomes under comparison (Altenhoff et al., 2012), the orthology hypothesis can be tested via a bi-directional search for similarity applying a best reciprocal hit (BRH) criterion: Only if the genes in question form the best search hit in both directions (*i.e.*, the genes are more similar to each other than to all other genes) they can be assumed to be orthologous. By grouping genes from multiple species that share BRH relations, one can form clusters of orthologous genes or simply orthologous groups (OGs) (Altenhoff & Dessimoz, 2012). This approach has been implemented in several software packages for use in genomic datasets (Li et al., 2003; Tatusov et al., 2003; Berglund et al., 2008; Zdobnov et al., 2017). Obtaining a complete and accurate genome sequence is, however, associated with technical difficulties and a cost that depends on the genome size. For these reasons, many phy-

logenetic studies employ transcriptomes: the nucleotide sequences of transcripts present in the sample at the time of RNA fixation (Wang et al., 2009). Transcriptomes can be sequenced at a fixed cost, however, in contrast to genomes, transcriptomes are inherently incomplete with regard to the gene set, simply because not all genes are expressed all the time and may therefore be absent from the sequenced RNA sample. While this is usually not a problem for phylogenetic analyses (Wiens, 2006), it means that the above-mentioned methods to *de novo* infer orthology designed for genomes cannot not be used on transcriptomic data. To infer orthology among transcripts, one usually applies a reference-based strategy that maps transcripts to known OGs. In the study by Misof et al. (2014), we used a software that implements this approach (Ebersberger et al., 2009), however, during the analyses, it became obvious that it had several design issues that were non-trivial to fix. Therefore, I concepted and wrote a re-implementation of the reference-based BRH orthology inference approach that mitigates those issues while delivering equal performance (Petersen et al., 2017). The software, Orthograph, is described in chapter 4.

1.3 Research questions

Insects and arthropods differ at the genomic and phenotypic level from vertebrate representatives. Like all forms of life, they share general mechanisms of genetic and genomic functionality and universal elements such as genes and TEs. However, genome composition, architecture and structural dynamics in insects appear drastically different from vertebrates and plants. In particular, the insect TE repertoire has not been subject to a large-scale study that would enable comparisons both within and between orders. The TE content has been assessed and put into context with the genome size in studies focusing on mosquitoes (Neafsey et al., 2015) and

Drosophila fruit flies (Sessegolo et al., 2016)), but conclusions across orders are hampered by absent information on the TE content and composition of key species as well as by non-standard TE annotation methods that make comparisons difficult. Additionally, a pan-ordinal taxon sampling could facilitate the inference of the ancestral insect TE content.

In chapter 2, I characterize the insect TE repertoire in a comparative standardized study based on genomic sequence data from 73 insect and non-insect arthropod species encompassing all major orders. The results highlight differential TE abundance and composition in comparisons between and within insect orders. The study also demonstrates that the TE diversity in insects is much larger than previously thought. The correlation between TE content and genome size in insects is substantiated. Additionally, the chapter shows that the TE copy divergence distribution is highly diverse and that some TEs are recently active among insects.

Genome size in insects is subject to large fluctuations, and TE content is not only correlated to genome size, but also to the insect phylogeny. To illuminate the effects of TE activity on genome size dynamics in insects, chapter 3 presents an extended analysis on the genomes of 96 arthropod species and genome size estimate data from 613 arthropod species. The study infers a comprehensive, time-calibrated phylogeny for these species using information from published studies and mitochondrial nucleotide sequences. This phylogeny is used to infer ancestral genome sizes and the amount of genome expansion and contraction since the last ancestor as a result of TE activity in the genome. It is shown that after about 100 Mya, the amount of ancestral TEs converges to zero. The rates of TE-associated DNA gain and loss are correlated with the phylogeny. The chapter also discusses a trend towards genome size stability despite varying rates of DNA removal in insects.

In chapter 4, I present the software Orthograph as a tool for orthology assessment in transcriptomic — or other coding nucleotide sequence — datasets. Building on and incorporating existing software packages for sequence similarity search and comparison, it implements an algorithm that uses the BRH criterion to map transcripts to clusters of genes with known orthology. The software is written in a modular fashion to make it flexible and portable. The chapter discusses the issues that previous implementations of the BRH mapping strategy suffer from and shows that Orthograph overcomes them by employing a relational database backend system that enables it to compare millions of search results in short time. Orthograph represents a powerful, versatile, and future-proof application which has been used in seven co-authored studies to date (Mayer et al., 2016; Pauli et al., 2016; Bank et al., 2017; Dowling et al., 2017; Peters et al., 2017; Gillung et al., 2018; Johnson et al., 2018; Wipfler et al., 2019) that are in appendix A.

Appendix B lists an additional seven co-authored publications. Among others, Misof et al. (2014) (page 310) consummates the inference of a robust phylogenetic backbone tree of 145 insect species from all major insect orders. I was a major contributor to the study. The analysis includes node dating with a comprehensive calibration dataset of 37 fossils. The study presents a semi-automated and documented reproducible workflow that facilitates phylogeny inference from transcriptomic data of many species. It also provides the backbone phylogeny and node dates for the study in chapter 3 and the breeding ground for the development of Orthograph (chapter 4).

References

Alfsnes, K., Leinaas, H. P., & Hessen, D. O. (2017). Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecology and Evolution*, (pp. n/a–n/a).

Altenberg, L. (2011). An Evolutionary Reduction Principle for Mutation Rates at Multiple Loci. *Bulletin of Mathematical Biology*, 73(6), 1227–1270.

Altenhoff, A. M. & Dessimoz, C. (2012). Inferring orthology and paralogy. *Methods in molecular biology (Clifton, N.J.)*, 855, 259–279.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., & Dessimoz, C. (2012). Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, 8(5), e1002514.

Aminetzach, Y. T., Macpherson, J. M., & Petrov, D. A. (2005). Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in Drosophila. *Science*, 309(5735), 764–767.

Bank, S., Sann, M., Mayer, C., Meusemann, K., Donath, A., Podsiadlowski, L., Kozlov, A., Petersen, M., Krogmann, L., Meier, R., Rosa, P., Schmitt, T., Wurdack, M., Liu, S., Zhou, X.,

Misof, B., Peters, R. S., & Niehuis, O. (2017). Transcriptome and target DNA enrichment sequence data provide new insights into the phylogeny of vespid wasps (Hymenoptera: Aculeata: Vespidae). *Molecular Phylogenetics and Evolution*, 116, 213–226.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, 471(7336), 51–57.

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population Genomics of Transposable Elements in Drosophila. *Annual Review of Genetics*, 48(1), 561–581.

Behjati, S. & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood. Education and Practice Edition*, 98(6), 236–238.

Bennett, M. D., Lewis, K. R., & Harberd, D. J. (1977). The Time and Duration of Meiosis [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 277(955), 201–226.

Berglund, A.-C., Sjölund, E., Ostlund, G., & Sonnhammer, E. L. L. (2008). InParanoid 6: Eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, 36(Database issue), D263–266.

Bernstein, H., Hopf, F. A., & Michod, R. E. (1987). The Molecular Basis of the Evolution of Sex. In J. G. Scandalios & E. W. Caspari (Eds.), *Advances in Genetics*, volume 24 of *Molecular Genetics of Development* (pp. 323–370). Academic Press.

Bertone, M. A., Leong, M., Bayless, K. M., Malow, T. L. F., Dunn, R. R., & Trautwein, M. D. (2016). Arthropods of the great indoors: Characterizing diversity inside urban and suburban homes. *PeerJ*, 4, e1582.

Blumenstiel, J. P., Chen, X., He, M., & Bergman, C. M. (2014). An Age-of-Allele Test of Neutrality for Transposable Element Insertions. *Genetics*, 196(2), 523–538.

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., Gates, M. W., Kula, R. R., & Brady, S. G. (2017). Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. *Current Biology*, 27(7), 1019–1025.

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., & Kawahara, A. Y. (2018). Resolving Relationships among the Megadiverse Butterflies and Moths with a Novel Pipeline for Anchored Phylogenomics. *Systematic Biology*, 67(1), 78–93.

Buchon, N. & Vaury, C. (2006). RNAi: A defensive RNA-silencing against viruses and transposable elements. *Heredity*, 96(2), 195–202.

Burke, M., Scholl, E. H., Bird, D. M., Schaff, J. E., Colman, S. D., Crowell, R., Diener, S., Gordon, O., Graham, S., Wang, X., Windham, E., Wright, G. M., & Opperman, C. H. (2015). The plant parasite Pratylenchus coffeaecarries a minimal nematode genome. *Nematology*, 17(6), 621–637.

Burkhead, N. M. (2012). Extinction Rates in North American Freshwater Fishes, 1900–2010. *BioScience*, 62(9), 798–808.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, 1(5), e1400253.

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, (pp. 201704949).

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biology and Evolution*, 7(2), 567–580.

Charlesworth, B. & Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetical Research*, 42(01), 1.

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, 2(4), e383.

Chen, S. & Li, X. (2007). Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes. *BMC Evol Biol*, 7(1), 46.

Chénais, B., Caruso, A., Hiard, S., & Casse, N. (2012). The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1), 7–15.

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., Carlton, J. S., Lewandowsky, S., Skuce, A. G., Green, S. A., Nuccitelli, D., Jacobs, P., Richardson,

M., Winkler, B., Painting, R., & Rice, K. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11(4), 048002.

Cridland, J. M., Macdonald, S. J., Long, A. D., & Thornton, K. R. (2013). Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. *Molecular Biology and Evolution*, 30(10), 2311–2327.

Daborn, P. J., Yen, J. L., Bogwitz, M. R., Goff, G. L., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., Feyereisen, R., Wilson, T. G., & ffrench-Constant, R. H. (2002). A Single P450 Allele Associated with Insecticide Resistance in Drosophila. *Science*, 297(5590), 2253–2256.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics*, 7(12), e1002384.

De Maddalena, A., Gabriotti, V., & Heim, W. (2008). *Sharks: The Perfect Predators*. Auckland Park, South Africa: Jacana Media. OCLC: ocn256666681.

Dirzo, R. & Raven, P. H. (2003). Global State of Biodiversity and Loss. *Annual Review of Environment and Resources*, 28(1), 137–167.

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J. B., & Collen, B. (2014). Defaunation in the Anthropocene. *Science*, 345(6195), 401–406.

Dolgin, E. S. & Charlesworth, B. (2006). The Fate of Transposable Elements in Asexual Populations. *Genetics*, 174(2), 817–827.

Dowling, D., Pauli, T., Donath, A., Meusemann, K., Podsiadlowski, L., Petersen, M., Peters, R. S., Mayer, C., Liu, S., Zhou, X., Misof, B., & Niehuis, O. (2017). Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects. *Genome Biology and Evolution*, (pp. evw281).

Dufresne, F. & Jeffery, N. (2011). A guided tour of large genome size in animals: What we know and where we are heading. *Chromosome Research*, 19(7), 925–938.

Dunn, C. W., Zapata, F., Munro, C., Siebert, S., & Hejnol, A. (2018). Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences*, (pp. 201707515).

Ebersberger, I., Strauss, S., & Von Haeseler, A. (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, 9(1), 157.

Fiston-Lavier, A.-S., Anxolabehere, D., & Quesneville, H. (2007). A model of segmental duplication formation in Drosophila melanogaster. *Genome Research*, 17(10), 1458–1470.

Gahan, L. J. (2001). Identification of a Gene Associated with Bt Resistance in Heliothis virescens. *Science*, 293(5531), 857–860.

Gillung, J. P., Winterton, S. L., Bayless, K. M., Khouri, Z., Borowiec, M. L., Yeates, D. K., Kimsey, L. S., Meusemann, K., Misof, B., Shin, S., Zhou, X., Mayer, C., Petersen, M., & Wiegmann, B. M. (2018). Bias in big-data phylogenetics: Anchored phylogenomics unravels the evolution of spider flies (Acroceridae) and reveals discordance between nucleotides and amino acids. *Molecular Biology and Evolution*.

Gray, Y. H. (2000). It takes two transposons to tango: Transposable-element-mediated chromosomal rearrangements. *Trends in genetics: TIG*, 16(10), 461–468.

Gregory, T. R. (2000). Nucleotypic effects without nuclei: Genome size and erythrocyte size in mammals. *Genome*, 43(5), 895–901.

Gregory, T. R. (2002). Genome size and developmental complexity. *Genetica*, 115(1), 131–146.

Gregory, T. R. (2004). Macroevolution, hierarchy theory, and the C-value enigma. *Paleobiology*, 30(2), 179–202.

Gregory, T. R., Ed. (2005). *The Evolution of the Genome*. Burlington, MA: Elsevier Academic. OCLC: ocm57727263.

Gregory, T. R. (2007). Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biological Reviews*, 76(1), 65–101.

Gregory, T. R. (2018). Animal Genome Size Database.

Grimaldi, D. A. & Engel, M. S. (2005). *Evolution of the Insects*. Cambridge [U.K.] ; New York: Cambridge University Press.

Grime, J. (1983). Prediction of weed and crop response to climate based upon measurements of nuclear DNA content. *Aspects of Applied Biology*, 4, 87–98.

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., Stenmans, W., Müller, A., Sumser, H., Hörren, T., Goulson, D., & de Kroon, H. (18-Oct-2017). More than

75 percent decline over 27 years in total flying insect biomass in protected areas. *PLOS ONE*, 12(10), e0185809.

Horner, A. & Macgregor, H. (1983). C value and cell volume: Their significance in the evolution and development of amphibians. *Journal of Cell Science*, 63, 135–46.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A. J., Juárez, M. J. A., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S. C., Himmelbauer, H., Minoche, A. E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Pérez, S. A., Ortega-Estrada, M. d. J., Cervantes-Luevano, J. I., Michael, T. P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V. A., & Herrera-Estrella, L. (2013). Architecture and evolution of a minute plant genome. *Nature*, 498(7452), 94–98.

IUCN (2018). *The IUCN Red List of Threatened Species*. Technical Report 2018-1, IUCN.

Johnson, K. P., Dietrich, C. H., Friedrich, F., Beutel, R. G., Wipfler, B., Peters, R. S., Allen, J. M., Petersen, M., Donath, A., Walden, K. K. O., Kozlov, A. M., Podsiadlowski, L., Mayer, C., Meusemann, K., Vasilikopoulos, A., Waterhouse, R. M., Cameron, S. L., Weirauch, C., Swanson, D. R., Percy, D. M., Hardy, N. B., Terry, I., Liu, S., Zhou, X., Misof, B., Robertson, H. M., & Yoshizawa, K. (2018). Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences*, 115(50), 12775–12780.

Kasturiratne, A., Wickremasinghe, A. R., de Silva, N., Gunawardena, N. K., Pathmeswaran, A., Premaratna, R., Savioli, L., Lalloo, D. G., & de Silva, H. J. (04-Nov-2008). The Global

Burden of Snakebite: A Literature Analysis and Modelling Based on Regional Estimates of Envenoming and Deaths. *PLOS Medicine*, 5(11), e218.

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160), 164–166.

Kelley, J. L., Peyton, J. T., Fiston-Lavier, A.-S., Teets, N. M., Yee, M.-C., Johnston, J. S., Bustamante, C. D., Lee, R. E., & Denlinger, D. L. (2014). Compact Genome of the Antarctic Midge Is Likely an Adaptation to an Extreme Environment. *Nature Communications*, 5.

Kofler, R., Betancourt, A. J., & Schlötterer, C. (2012). Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila melanogaster. *PLoS Genetics*, 8(1), e1002487.

Kofler, R., Nolte, V., & Schlötterer, C. (2015). Tempo and Mode of Transposable Element Activity in Drosophila. *PLOS Genetics*, 11(7), e1005406.

Kolbert, E. (2014). *The Sixth Extinction: An Unnatural History*. New York: Henry Holt and Company, first edition edition.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39, 309–338.

Kusmanoff, A. M., Fidler, F., Gordon, A., & Bekessy, S. A. (2017). Decline of 'biodiversity' in conservation policy discourse in Australia. *Environmental Science & Policy*, 77, 160–165.

Lamarque, F., Ed. (2009). *Human-Wildlife Conflict in Africa: Causes, Consequences and Management Strategies*. Number 157 in FAO Forestry Paper. Rome: Food and Agriculture Organization of the United Nations. OCLC: 837661033.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P.,

Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., & International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.

Leakey, R. E. & Lewin, R. (1996). *The Sixth Extinction: Patterns of Life and the Future of Humankind*. New York: Anchor Books. OCLC: 36002376.

Li, L., Stoeckert, C., & Roos, D. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9), 2178–2189.

Lim, J. K. (1988). Intrachromosomal rearrangements mediated by hobo transposons in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 85(23), 9153–9157.

Linnell, J. D. C. (2011). Can we separate the sinners from the scapegoats? *Animal Conservation*, 14(6), 602–603.

Lisch, D. (2009). Epigenetic Regulation of Transposable Elements in Plants. *Annual Review of Plant Biology*, 60(1), 43–66.

Makałowski, W. (2001). The human genome structure and organization. *Acta Biochimica Polonica*, 48(3), 587–598.

Mathiopoulos, K. D., della Torre, A., Predazzi, V., Petrarca, V., & Coluzzi, M. (1998). Cloning of inversion breakpoints in the Anopheles gambiae complex traces a transposable element at the inversion junction. *Proceedings of the National Academy of Sciences*, 95(21), 12444–12449.

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design. *Molecular Biology and Evolution*, 33(7), 1875–1886.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6), 344–355.

McKenna, D. D., Wild, A. L., Kanda, K., Bellamy, C. L., G., B. R., Caterino Michael S., Farnum Charles W., Hawks David C., Ivie Michael A., Jameson Mary Liz, Leschen Richard a. B., Marvaldi Adriana E., Mchugh Joseph V., Newton Alfred F., Robertson James A., Thayer Margaret K., Whiting Michael F., Lawrence John F., Ślipiński Adam, Maddison David R., & Farrell Brian D. (2015). The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Systematic Entomology*, 40(4), 835–880.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von, R. B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K., & Zhou, X. (2014). Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science*, 346, 763–7.

Mondal, M., Klimov, P., & Flynt, A. S. (2018). Rewired RNAi-mediated genome surveillance in house dust mites. *PLOS Genetics*, 14(1), e1007183.

Mooney, H. A. & Cleland, E. E. (2001). The evolutionary impact of invasive species. *Proceedings of the National Academy of Sciences*, 98(10), 5446–5451.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arca, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., Burt, A., Chan, C. S., Chauve, C., Chiu, J. C., Christensen, M., Costantini, C., Davidson, V. L. M., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S. B., Guelbeogo, W. M., Hall, A. B., Han, M. V., Hlaing, T., Hughes, D. S. T., Jenkins, A. M., Jiang, X., Jungreis, I., Kakani, E. G., Kamali, M., Kemppainen, P., Kennedy, R. C., Kirmitzoglou, I. K., Koekemoer, L. L., Laban, N., Langridge, N., Lawniczak, M. K. N., Lirakis, M., Lobo, N. F., Lowy, E., MacCallum, R. M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S. N., Moore, W., Murphy, K. A., Naumenko, A. N., Nolan, T., Novoa, E. M., O'Loughlin, S., Oringanje, C., Oshaghi, M. A., Pakpour, N., Papathanos, P. A., Peery, A. N., Povelones, M., Prakash, A., Price, D. P., Rajaraman, A., Reimer, L. J., Rinker, D. C., Rokas, A., Russell, T. L., Sagnon, N., Sharakhova, M. V., Shea, T., Simao, F. A., Simard, F., Slotman, M. A., Somboon, P., Stegniy, V., Struchiner, C. J., Thomas, G. W. C., Tojo, M., Topalis, P., Tubio, J. M. C., Unger, M. F., Vontas, J., Walton, C., Wilding, C. S., Willis, J. H., Wu, Y.-C., Yan, G., Zdobnov, E. M., Zhou, X., Catteruccia, F., Christophides, G. K., Collins, F. H., Cornman, R. S., Crisanti, A., Donnelly, M. J., Emrich, S. J., Fontaine, M. C., Gelbart, W., Hahn, M. W., Hansen, I. A., Howell, P. I., Kafatos,

F. C., Kellis, M., Lawson, D., Louis, C., Luckhart, S., Muskavitch, M. A. T., Ribeiro, J. M., Riehle, M. A., Sharakhov, I. V., Tu, Z., Zwiebel, L. J., & Besansky, N. J. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, 347(6217), 1258522–1258522.

Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., Palma, A. D., Ferrier, S., Hill, S. L. L., Hoskins, A. J., Lysenko, I., Phillips, H. R. P., Burton, V. J., Chng, C. W. T., Emerson, S., Gao, D., Pask-Hale, G., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmons, B. I., Whitmee, S., Zhang, H., Scharlemann, J. P. W., & Purvis, A. (2016). Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science*, 353(6296), 288–291.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733–745.

Oliveira, C. M., Auad, A. M., Mendes, S. M., & Frizzas, M. R. (2014). Crop losses and the economic impact of insect pests on Brazilian agriculture. *Crop Protection*, 56, 50–54.

Olmo, E. & Odierna, G. (1982). Relationships between DNA content and cell morphometric parameters in reptiles. *Basic and Applied Histochemistry*, 26(1), 27–34.

Packer, C., Ikanda, D., Kissui, B., & Kushnir, H. (2005). Conservation biology: Lion attacks on humans in Tanzania. *Nature*, 436(7053), 927–928.

Parfrey, L. W., Lahr, D. J. G., & Katz, L. A. (2008). The dynamic nature of eukaryotic genomes. *Molecular Biology and Evolution*, 25(4), 787–794.

Pauli, T., Vedder, L., Dowling, D., Petersen, M., Meusemann, K., Donath, A., Peters, R. S., Podsiadlowski, L., Mayer, C., Liu, S., Zhou, X., Heger, P., Wiehe, T., Hering, L., Mayer, G., Misof, B., & Niehuis, O. (2016). Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects. *BMC Genomics*, 17, 861.

Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P. A., Heraty, J., Kjer, K. M., Klopfstein, S., Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B., & Niehuis, O. (2017). Evolutionary History of the Hymenoptera. *Current Biology*.

Petersen, M., Armisén, D., Gibbs, R. A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., & Misof, B. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evolutionary Biology*, 19(1).

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R. S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., & Niehuis, O. (2017). Orthograph: A versatile tool

for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, 18, 111.

Petrov, D. A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K., & Gonzalez, J. (2011). Population Genomics of Transposable Elements in Drosophila melanogaster. *Molecular Biology and Evolution*, 28(5), 1633–1644.

Pimm, S. L., Russell, G. J., Gittleman, J. L., & Brooks, T. M. (1995). The Future of Biodiversity. *Science*, 269(5222), 347–350.

Puntaru, C. (2017). The politics of biodiversity conservation.

Remnant, E. J., Good, R. T., Schmidt, J. M., Lumb, C., Robin, C., Daborn, P. J., & Batterham, P. (2013). Gene duplication in the major insecticide target site, Rdl, in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 110(36), 14705–14710.

Schipper, J., Chanson, J. S., Chiozza, F., Cox, N. A., Hoffmann, M., Katariya, V., Lamoreux, J., Rodrigues, A. S. L., Stuart, S. N., Temple, H. J., Baillie, J., Boitani, L., Lacher, T. E., Mittermeier, R. A., Smith, A. T., Absolon, D., Aguiar, J. M., Amori, G., Bakkour, N., Baldi, R., Berridge, R. J., Bielby, J., Black, P. A., Blanc, J. J., Brooks, T. M., Burton, J. A., Butynski, T. M., Catullo, G., Chapman, R., Cokeliss, Z., Collen, B., Conroy, J., Cooke, J. G., da Fonseca, G. A. B., Derocher, A. E., Dublin, H. T., Duckworth, J. W., Emmons, L., Emslie, R. H., Festa-Bianchet, M., Foster, M., Foster, S., Garshelis, D. L., Gates, C., Gimenez-Dixon, M., Gonzalez, S., Gonzalez-Maya, J. F., Good, T. C., Hammerson, G., Hammond, P. S., Happold, D., Happold, M., Hare, J., Harris, R. B., Hawkins, C. E., Haywood, M., Heaney, L. R., Hedges,

S., Helgen, K. M., Hilton-Taylor, C., Hussain, S. A., Ishii, N., Jefferson, T. A., Jenkins, R. K. B., Johnston, C. H., Keith, M., Kingdon, J., Knox, D. H., Kovacs, K. M., Langhammer, P., Leus, K., Lewison, R., Lichtenstein, G., Lowry, L. F., Macavoy, Z., Mace, G. M., Mallon, D. P., Masi, M., McKnight, M. W., Medellín, R. A., Medici, P., Mills, G., Moehlman, P. D., Molur, S., Mora, A., Nowell, K., Oates, J. F., Olech, W., Oliver, W. R. L., Oprea, M., Patterson, B. D., Perrin, W. F., Polidoro, B. A., Pollock, C., Powel, A., Protas, Y., Racey, P., Ragle, J., Ramani, P., Rathbun, G., Reeves, R. R., Reilly, S. B., Reynolds, J. E., Rondinini, C., Rosell-Ambal, R. G., Rulli, M., Rylands, A. B., Savini, S., Schank, C. J., Sechrest, W., Self-Sullivan, C., Shoemaker, A., Sillero-Zubiri, C., Silva, N. D., Smith, D. E., Srinivasulu, C., Stephenson, P. J., van Strien, N., Talukdar, B. K., Taylor, B. L., Timmins, R., Tirira, D. G., Tognelli, M. F., Tsytsulina, K., Veiga, L. M., Vié, J.-C., Williamson, E. A., Wyatt, S. A., Xie, Y., & Young, B. E. (2008). The Status of the World's Land and Marine Mammals: Diversity, Threat, and Knowledge. *Science*, 322(5899), 225–230.

Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, 517, 321–326.

Sessegolo, C., Burlet, N., & Haudry, A. (2016). Strong Phylogenetic Inertia on Genome Size and Transposable Element Content among 26 Species of Flies. *Biology Letters*, 12(8), 20160407.

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: The vanguard of genome defence. *Nature Reviews Molecular Cell Biology*, 12(4), 246–258.

Slotkin, R. K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4), 272–285.

Staton, S. E. & Burke, J. M. (2015). Transposome: A toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics*, 31(11), 1827–1829.

Struchiner, C. J., Massad, E., Tu, Z., & Ribeiro, J. M. C. (2009). The tempo and mode of evolution of transposable elements as revealed by molecular phylogenies reconstructed from mosquito genomes. *Evolution*, 63(12), 3136–3146.

Suzuki, M. M. & Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6), 465–476.

Szitenberg, A., Cha, S., Opperman, C. H., Bird, D. M., Blaxter, M. L., & Lunt, D. H. (2016). Genetic drift, not life history or RNAi, determine long term evolution of transposable elements. *Genome Biology and Evolution*, (pp. evw208).

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41.

van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., & Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534(7605), 102–105.

Vinogradov, A. E. (1997). Nucleotypic Effect in Homeotherms: Body-Mass Independent Resting Metabolic Rate of Passerine Birds is Related to Genome Size. *Evolution*, 51(1), 220–225.

Vogel, G. (2017). Where have all the insects gone? *Science*.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), 57–63.

Warnefors, M., Pereira, V., & Eyre-Walker, A. (2010). Transposable Elements: Insertion Pattern and Impact on Gene Expression Evolution in Hominids. *Molecular Biology and Evolution*, 27(8), 1955–1962.

White, M. & McLaren, I. (2000). Copepod development rates in relation to genome size and 18S rDNA copy number. *Genome*, 43(5), 750–755.

WHO (2017). *Vector-Borne Diseases*. Fact sheet, WHO.

Wielgoss, S., Barrick, J. E., Tenaillon, O., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., Lenski, R. E., & Schneider, D. (2011). Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia coli. *G3 (Bethesda, Md.)*, 1(3), 183–186.

Wiens, J. J. (2006). Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics*, 39(1), 34–42.

Wipfler, B., Letsch, H., Frandsen, P. B., Kapli, P., Mayer, C., Bartel, D., Buckley, T. R., Donath, A., Edgerly-Rooks, J. S., Fujita, M., Liu, S., Machida, R., Mashimo, Y., Misof, B., Niehuis, O., Peters, R. S., Petersen, M., Podsiadlowski, L., Schütte, K., Shimizu, S., Uchifune, T., Wilbrandt, J., Yan, E., Zhou, X., & Simon, S. (2019). Evolutionary history of Polyneoptera

and its implications for our understanding of early winged insects. *Proceedings of the National Academy of Sciences*, (pp. 201817794).

Wright, S. J. (2005). Tropical forests in a changing environment. *Trends in Ecology & Evolution*, 20(10), 553–560.

Wuebbles, D., Fahey, D., Hibbard, K., DeAngelo, B., Doherty, S., Hayhoe, K., Horton, R., Kossin, J., Taylor, P., Waple, A., Weaver, C., Wuebbles, D., Fahey, D., Hibbard, K., Dokken, D., Stewart, B., & Maycock, T. (2017). *Executive Summary. Climate Science Special Report: Fourth National Climate Assessment, Volume I*. Technical report, U.S. Global Change Research Program.

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8), 335–340.

Zaher, H. S. & Green, R. (2009). Fidelity at the molecular level: Lessons from protein synthesis. *Cell*, 136(4), 746–762.

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., Seppey, M., Loetscher, A., & Kriventseva, E. V. (2017). OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, 45(D1), D744–D749.

Zeng, L., Zhang, Q., Yan, K., & Zhou, M.-M. (2011). Structural insights into piRNA recognition by the human PIWI-like 1 PAZ domain. *Proteins: Structure, Function, and Bioinformatics*, 79(6), 2004–2009.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, 39(1), 61–69.

2

Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects

This chapter has been published in: Petersen, M., Armisén, D., Gibbs, R.A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., Misof, B. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evol Biol*, 19. doi:10.1186/s12862-018-1324-9

Author contributions to the original article:

BM, MP, and ON conceived the study. MP performed all analyses. BM and MP interpreted the results and wrote the manuscript draft. AK, DA, GM, LH, and ON collected specimens and performed laboratory procedures including RNA/DNA extraction. RAG and SR co-ordinated, sequenced, assembled and made available genome reference sequences of species within the i5K pilot.

2.1 Introduction

Repetitive elements, including transposable elements (TEs), are a major sequence component of eukaryote genomes. In vertebrate genomes, for example, the TE content varies from 6 % in the pufferfish *Tetraodon nigroviridis* to more than 55 % in the zebrafish *Danio rerio* (Chalopin et al., 2015). More than 45 % of the human genome (de Koning et al., 2011) consist of TEs. In plants, TEs are even more prevalent: up to 90 % of the maize (*Zea mays*) genome is covered by TEs (SanMiguel et al., 1996). In insects, the genomic portion of TEs ranges from as little as 1 % in the antarctic midge (Kelley et al., 2014) to as large as 65 % in the migratory locust (Wang et al., 2014).

TEs are known as "jumping genes" and traditionally viewed as selfish parasitic nucleotide sequence elements propagating in genomes with mainly deleterious or at least neutral effects

on host fitness (Mackay, 1989; Pasyukova, 2004) (reviewed in Barrón et al. (2014)). Due to their propagation in the genome, TEs are thought to have a considerable influence on the evolution of the host's genome architecture. By transposing into, for example, host genes or regulatory sequences, TEs can disrupt coding sequences or gene regulation, and/or provide hot spots for ectopic (non-homologous) recombination that may induce chromosomal rearrangements in the host genome such as deletions, duplications, inversions, and translocations (Burns & Boeke, 2012). For example, the shrinkage of the Y chromosome in the fruit fly *Drosophila melanogaster*, which consists mostly of TEs, is thought to be caused by such intrachromosomal rearrangements induced by ectopic recombination (Adams, 2000; Kent et al., 2017). As such potent agents for mutation, TEs are also responsible for cancer and genetic diseases in humans and other organisms (Vorechovsky, 2009; Chenais, 2015; Hancks & Kazazian, 2016).

Despite the potential deleterious effects of their activity on gene regulation, there is growing evidence that TEs can also be drivers of genomic innovation that confer selective advantages to the host (Casola et al., 2007; González et al., 2008). For example, it is well documented that the frequent cleavage and rearrangement of DNA strands induced by TE insertions provides a source of sequence variation to the host genome, or that by a process called molecular domestication of TEs, host genomes derive new functional genes and regulatory networks (Feschotte, 2008; Böhne et al., 2008; Santos et al., 2014). Furthermore, many exons have been *de novo*-recruited from TE insertions in coding sequences of the human genome (Zhang & Chasin, 2006). In insects, TE insertions have played a pivotal role in the acquisition of insecticide resistance (Chen & Li, 2007; Itokawa et al., 2010; Gahan, 2001), as well as in the rewiring of a regu-

latory network that provides dosage compensation (Ellison & Bachtrog, 2013), or the evolution of climate adaptation (González et al., 2010; Kim et al., 2014).

TEs are classified depending on their mode of transposition. Class I TEs, also known as retrotransposons, transpose via an RNA-mediated mechanism that can be circumscribed as "copy-and-paste". They are further subdivided into long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons. Non-LTR retrotransposons include long and short interspersed nuclear elements (LINEs and SINEs) (Malik et al., 1999; Eickbush & Jamburuthugoda, 2008). Whereas LTR retrotransposons and LINEs encode a reverse transcriptase, the non-autonomous SINEs rely on the transcriptional machinery of autonomous elements, such as LINEs, for mobility. Frequently found LTR retrotransposon families in eukaryote genomes include Ty3/Gypsy, which was originally described in *Arabidopsis thaliana* (Marin & Llorens, 2000), Ty1/Copia (Flavell, 1992), as well as BEL/Pao (de la Chaux & Wagner, 2011).

In Class II TEs, also termed DNA transposons, the transposition is DNA-based and does not require an RNA intermediate. Autonomous DNA transposons encode a transposase enzyme and move via a "cut-and-paste" mechanism. During replication, terminal inverted repeat (TIR) transposons and Crypton-type elements cleave both DNA strands (Wicker et al., 2007). Helitrons, also known as rolling-circle (RC) transposons due to their characteristic mode of transposition (Kapitonov & Jurka, 2001), and the self-synthesizing Maverick/Polinton elements (Krupovic & Koonin, 2016) cleave a single DNA strand in the process of replication. Both Helitron and Maverick/Polinton elements occur in autonomous and non-autonomous versions (Kapitonov & Jurka, 2006, 2007), the latter of which do not encode all proteins necessary for transposition. Helitrons are the only Class II transposons that do not cause a flanking target site

duplication when they transpose. Class II also encompasses other non-autonomous DNA transposons such as miniature inverted TEs (MITEs) (Shirasawa et al., 2012), which exploit and rely on the transposase mechanisms of autonomous DNA transposons to replicate.

Previous reports on insect genomes describe the composition of TE families in insect genomes as a mixture of insect specific TEs and TEs common to metazoa (Feschotte & Pritham, 2007; Maumus et al., 2015; Chuong et al., 2016). Overall, surprisingly little effort has been put into characterizing TE sequence families and TE compositions in insect genomes in large-scale comparative analyses encompassing multiple taxonomic orders to paint a picture of the insect TE repertoire. Dedicated comparative analyses of TE composition have been conducted on species of mosquitoes (Neafsey et al., 2015), of drosophilid flies (Sessegolo et al., 2016), and of *Macrosiphini* (aphids) (Bouallègue et al., 2017). Despite these efforts in characterizing TEs in insect genomes, still little is known about the diversity of TEs in insect genomes, owed in part to the huge insect species diversity and to the lack of a standardized analysis that allows comparisons across taxonomic orders. While this lack of knowledge is due to the low availability of sequenced insect genomes in the past, efforts such as the i5k initiative (Robinson et al., 2011) have helped to increase the number of genome sequences from previously unsampled insect taxa. With this denser sampling of insect genomic diversity available, it now seems possible to comprehensively investigate the TE diversity among major insect lineages.

Here, we present the first exhaustive analysis of the distribution of TE classes in a sample representing half of the currently classified insect (hexapod *sensu* Misof *et al.* (Misof et al., 2014)) orders and using standardized comparative methods implemented in recently developed software packages. Our results show similarities in TE family diversity and abundance among the

investigated insect genomes, but also profound differences in TE activity even among closely related species.

2.2 Materials and methods

2.2.1 Genomic data sets

We downloaded genome assemblies of 42 arthropod species from NCBI GenBank at `ftp.ncbi.nlm.nih.gov/genomes` (last accessed 2014-11-26; supplementary table C.3) as well as the genome assemblies of 31 additional species from the i5k FTP server at ftp.hgsc.bcm.edu/I5K-pilot (last accessed 2016-07-08; supplementary table C.3). Our taxon sampling includes 21 dipterans, four lepidopterans, one trichopteran, five coleopterans, one strepsipteran, 14 hymenopterans, one psocodean, six hemipterans, one thysanopteran, one blattodean, one isopteran, one orthopteran, one ephemeropteran, one odonate, one archaeognathan, and one dipluran. As outgroups we included three crustaceans, one myriapod, six chelicerates, and one onychophoran.

2.2.2 Construction of species-specific repeat libraries and TE annotation in the genomes

We compiled species-specific TE libraries using automated annotation methods. RepeatModeler Open-1.0.8 (Smit et al., 2015) was employed to cluster repetitive *k*-mers in the assembled genomes and infer consensus sequences. These consensus sequences were classified using a reference-based similarity search in RepBase Update 20140131 (Jurka et al., 2005). The entries in the resulting repeat libraries were then searched for using nucleotide BLAST in the NCBI

nr database (downloaded 2016-03-17 from ftp.ncbi.nlm.nih.gov/blast/db) to verify that the included consensus sequences are indeed TEs and not annotation artifacts. Repeat sequences that were annotated as "unknown" and that resulted in a BLAST hit for known TE proteins such as reverse transcriptase, transposase, integrase, or known TE domains such as gag/pol/env, were kept and considered unknown TE nucleotide sequences; but all other "unknown" sequences were not considered TE sequences and therefore removed. The filter patterns are included in the data package available at the Dryad repository at the URI `https://doi.org/10.5061/dryad.55p667b`. The filtered repeat library was combined with the Metazoa-specific section of RepBase version 20140131 and subsequently used with RepeatMasker 4.0.5 (Smit et al., 2015) to annotate TEs in the genome assemblies.

2.2.3 Validation of Alu presence

To exemplarily validate our annotation, we selected the SINE Alu, which was previously only identified in primates (Kriegs et al., 2007). We retrieved a Hidden Markov model (HMM) profile for the AluJo subfamily from the repeat database Dfam (Hubley et al., 2015) and used the HMM to search for Alu copies in the genome assemblies. We extracted the hit nucleotide subsequences from the assemblies and inferred a multiple nucleotide sequence alignment with the canonical Alu nucleotide sequence from Repbase (Jurka et al., 2005).

2.2.4 Genomic TE coverage and correlation with genome size

We used the tool "one code to find them all" (Bailly-Bechet et al., 2014) on the RepeatMasker output tables to calculate the genomic proportion of annotated TEs. "One code to find them

all" is able to merge entries belonging to fragmented TE copies to produce a more accurate estimate of the genomic TE content and especially the copy numbers. To test for a relationship between genome assembly size and TE content, we applied a linear regression model and tested for correlation using the Spearman rank sum method. To see whether the genomes of holometabolous insects are different than the genomes of hemimetabolous insects in TE content, we tested for an effect of the taxa using their mode of metamorphosis as a three-class factor: Holometabola (all holometabolous insect species), non-Eumetabola (all non-holometabolous hexapod species, with the exception of Hemiptera, Thysanoptera, and Psocodea; (Beutel, 2013)), and Acercaria (Hemiptera, Thysanoptera, and Psocodea). We also tested for a potential phylogenetic effect on the correlation between genome size and TE content with the phylogenetic independent contrasts (PIC) method proposed by Felsenstein (Felsenstein, 1985) using the ape package (Paradis et al., 2004) within R (R Core Team, 2017)

2.2.5 Kimura distance-based TE age distribution

We used intra-family TE nucleotide sequence divergence as a proxy for intra-family TE age distributions. Sequence divergence was calculated as intra-family Kimura distances (rates of transitions and transversions) using the specialized helper scripts from the RepeatMasker 4.0.5 package. The tools compute the Kimura distance between each annotated TE copy and the consensus sequence of the respective TE family, and provide the data in tabular format for processing. When plotted (Fig. 2.5), a peak in the distribution shows the genomic coverage of the TE copies with that specific Kimura distance to the repeat family consensus. Thus, a large peak with high Kimura distance would indicate a group of TE copies with high sequence divergence due to

genetic drift or other processes. The respective TE copies are likely older than copies associated with a peak at low Kimura distance. We used the Kimura distances without correction for CpG pairs since TE DNA methylation is clearly absent in holometabolous insects and insufficiently described in hemimetabolous insects (Glastad et al., 2014). All TE age distribution landscapes were inferred from the data obtained by annotating the genomes with *de novo*-generated species-specific repeat libraries.

2.3 Results

2.3.1 Diversity of TE content in arthropod genomes

TE content varies greatly among the analyzed species (Fig. 2.1, supplemental table C.2) and differs even between species belonging the same order. In the insect order Diptera, for example, the TE content varies from around 55 % in the yellow fever mosquito *Aedes aegypti* to less than 1 % in *Belgica antarctica*. Even among closely related *Drosophila* species, the TE content ranges from 40 % (in *D. ananassae*) to 10 % (in *D. miranda* and *D. simulans*). The highest TE content (60 %) was found in the large genome (6.5 Gbp) of the migratory locust *Locusta migratoria* (Orthoptera), while the smallest known insect genome, that of the antarctic midge *B. antarctica* (Diptera, 99 Mbp), was found to contain less than 1 % TEs. The TE content of the majority of the genomes was spread around a median of 24.4 % with a standard deviation of 12.5 %.

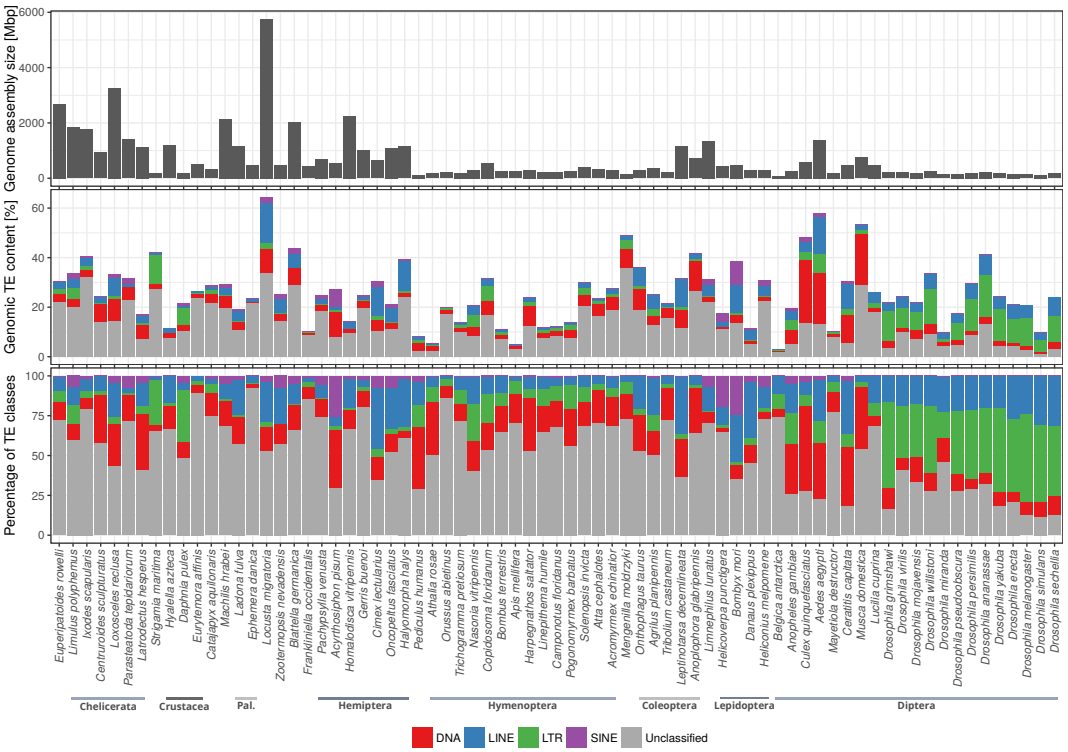

Figure 2.1: Genome assembly size, total amount and relative proportion of DNA transposons, LTR, LINE and SINE retrotransposons in arthropod genomes and a representative of Onychophora as an outgroup. Also shown is the genomic proportion of unclassified/uncharacterized repetitive elements. Pal., Palaeoptera

2.3.2 Relative contribution of different TE types to arthropod genome sequences

We assessed the relative contribution of the major TE groups (LTR, LINE, SINE retrotransposons, and DNA transposons) to the arthropod genome composition (Fig. 2.1). In most species, "unclassified" elements, which need further characterization, represent the largest fraction. They contribute up to 93 % of the total TE coverage in the mayfly *Ephemera danica* or the copepod *Eurytemora affinis*. Unsurprisingly, in most investigated *Drosophila* species the unclassifiable elements comprise less than 25 % and in *D. simulans* only 11 % of the entire TE

content, likely because the *Drosophila* genomes are well annotated and most of their content is known (in fact, many TEs were first found in representatives of *Drosophila*). Disregarding these unclassified TE sequences, LTR retrotransposons dominate the TE content in representatives of Diptera, in some cases contributing around 50 % (e.g., in *D. simulans*). In Hymenoptera, on the other hand, DNA transposons are more prevalent, such as 35.25 % in Jerdon's jumping ant *Harpegnathos saltator*. LINE retrotransposons are represented with up to 39.3 % in Hemiptera and Psocodea (*Acyrthosiphon pisum* and *Cimex lectularius*), with the exception of the human body louse *Pediculus humanus*, where DNA transposons contribute 44.43 % of the known TE content. SINE retrotransposons were found in all insect orders, but they contributed less than 10 % of the genomic TE content in any taxon in our sampling, with the exception of *Helicoverpa punctigera* (18.48 %), *Bombyx mori* (26.38 %), and *A. pisum* (27.11 %). In some lineages, such as Hymenoptera and most dipterans, SINEs contribute less than 1 % to the TE content, whereas in Hemiptera and Lepidoptera the SINE coverage ranges from 0.08 % to 26.38 % (Hemiptera) and 3.35 % to 26.38 % (Lepidoptera). Note that these numbers are likely higher and many more DNA, LTR, LINE, and SINE elements may be obscured by the large "unclassified" portion.

2.3.3 Contribution of TEs to arthropod genome size

We assessed the TE content, that is, the ratio of TE versus non-TE nucleotides in the genome assembly, in 62 hexapod (insects *sensu* Misof et al. (2014)) species as well as an outgroup of 10 non-insect arthropods and a representative of Onychophora (velvet worms). We tested whether there was a relationship between TE content and genome assembly size, and found a positive correlation (Fig. 2.2 and supplemental table C.2). This correlation is statistically significant

(Spearman's rank sum test, $\varrho = 0.495, p \lll 0.005$). Genome size is significantly smaller in holometabolous insects than in non-holometabolous insects (one-way ANOVA, $p = 0.0001$). Using the ape package v. 4.1 (Paradis et al., 2004) for R (R Core Team, 2017), we tested for correlation between TE content and genome size using phylogenetically independent contrasts (PIC, Felsenstein (1985)). The test confirmed a significant positive correlation (Pearson product-moment correlation, $\varrho = 0.497, p = 0.0001$, corrected for phylogeny using PIC) between TE content and genome size. Additionally, genome size is correlated with TE diversity, that is, the number of different TE superfamilies found in a genome (Spearman, $\varrho = 0.712, p \lll 0.005$); this is also true under PIC (Pearson, $\varrho = 0.527, p \lll 0.005$; Fig. C.1).

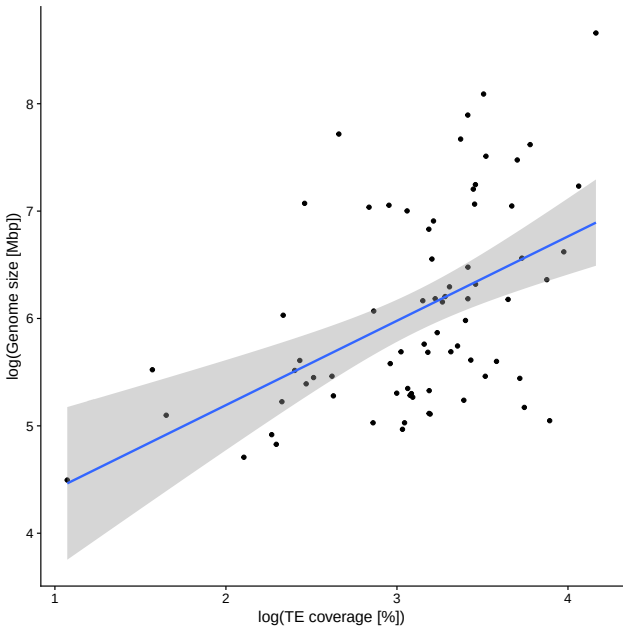

Figure 2.2: TE content in 73 arthropod genomes is positively correlated to genome assembly size (Spearman rank correlation test, $\varrho = 0.495, p \lll 0.005$). This correlation is also supported under phylogenetically independent contrasts (Felsenstein, 1985) (Pearson product moment correlation, $\varrho = 0.497, p = 0.0001225$). Dots: Individual measurements; blue line: linear regression; grey area: confidence interval.

2.3.4 Distribution of TE superfamilies in arthropods

We identified almost all known TE superfamilies in at least one insect species, and many were found to be widespread and present in all investigated species (Fig. 2.3, note that in this figure, TE families were summarized in superfamilies). Especially diverse and ubiquitous are DNA transposon superfamilies, which represent 22 out of 70 identified TE superfamilies. The most widespread (present in all investigated species) DNA transposons belong to the superfamilies Academ, Chapaev and other superfamilies in the CMC complex, Crypton, Dada, Ginger, hAT (Blackjack, Charlie, *etc.*), Kolobok, Maverick, Harbinger, PiggyBac, Helitron (RC), Sola, TcMar (Mariner, Tigger, *etc.*), and the P element superfamily. LINE non-LTR retrotransposons are similarly ubiquitous, though not as diverse. Among the most widespread LINEs are TEs belonging to the superfamilies CR1, Jockey, L1, L2, LOA, Penelope, R1, R2, and RTE. Of the LTR retrotransposons, the most widespread are in the superfamilies Copia, DIRS, Gypsy, Ngaro, and Pao as well as endogenous retrovirus particles (ERV). SINE elements are diverse, but show a more patchy distribution, with only the tRNA-derived superfamily present in all investigated species. We found elements belonging to the ID superfamily in almost all species except the Asian long horned beetle, *Anoplophora glabripennis*, and the B4 element absent from eight species. All other SINE superfamilies are absent in at least 13 species. Elements from the Alu superfamily were found in 48 arthropod genomes, for example in the silkworm *Bombyx mori* (Fig. 2.4, all Alu alignments are shown in Additional File 1).

On average, the analyzed species harbor a mean of 54.8 different TE superfamilies, with the locust *L. migratoria* exhibiting the greatest diversity (61 different TE superfamilies), followed

by the tick *Ixodes scapularis* (60), the velvet worm *Euperipatoides rowelli* (59), and the dragonfly *Ladona fulva* (59). Overall, Chelicerata have the highest average TE superfamily diversity (56.7). The greatest diversity among the multi-representative hexapod orders was found in Hemiptera (55.7). The mega-diverse insect orders Diptera, Hymenoptera, and Coleoptera display a relatively low diversity of TE superfamilies (48.5, 51.8, and 51.8, respectively). The lowest diversity was found in *A. aegypti*, with only 41 TE superfamilies.

2.3.5 Lineage-specific TE presence and absence in insect orders

We found lineage-specific TE diversity within most insect orders. For example, the LINE superfamily Odin is absent in all Hymenoptera studied, whereas Proto2 was found in all Hymenoptera except in the ant *H. saltator* and in all Diptera except in *C. quinquefasciatus*. Similarly, the Harbinger DNA element superfamily was found in all Lepidoptera except for the silkworm *B. mori*. Also within Palaeoptera (*i.e.*, mayflies, damselflies, and dragonflies), the Harbinger superfamily is absent in *E. danica*, but present in all other representatives of Palaeoptera. These clade-specific absences of a TE superfamily may be the result of lineage-specific TE extinction events during the evolution of the different insect orders. Note that since a superfamily can encompass multiple different TEs, the absence of a specific superfamily can either result from independent losses of multiple TEs belonging to that superfamily, or a single loss if there only was a single TE of that superfamily in the genome.

We also found TE superfamilies represented only in a single species of an insect clade. For example, the DNA element superfamily Zisupton was found only in the wasp *Copidosoma floridanum*, but not in other Hymenoptera, and the DNA element Novosib was found only in *B.*

mori, but not in other Lepidoptera. Within Coleoptera, only the Colorado potato beetle, *Leptinotarsa decemlineata* harbors the LINE superfamily Odin. Likewise, we found the Odin superfamily among Lepidoptera only in the noctuid *Helicoverpa punctigera*. We found the LINE superfamily Proto1 only in *Pediculus humanus* and in no other species. These examples of clade or lineage specific occurrence of TEs, which are absent from other species of the same order (or the entire taxon sampling), could be the result of a horizontal transfer from food species or a bacterial/viral infection.

2.3.6 Lineage-specific TE activity during arthropod evolution

We further analyzed sequence divergence measured by Kimura distance within each species-specific TE content (Fig. 2.5; note that for these plots, we omitted the large fraction of unclassified elements). Within Diptera, the most striking feature is that almost all investigated drosophilids show a large spike of LTR retroelement proliferation between Kimura distance 0 and around 0.08. This spike is only absent in *D. miranda*, but bi-modal in *D. pseudoobscura*, with a second peak around Kimura distance 0.15. This second peak, however, does not coincide with the age of inversion breakpoints on the third chromosome of *D. pseudoobscura*, which are only a million years old and have been associated with TE activity (Wallace et al., 2011). A bi-modal distribution was not observed in any other fly species. On the contrary, all mosquito species exhibit a large proportion of DNA transposons which show a divergence between Kimura distance 0.02 and around 0.3. This divergence is also present in the calyptrate flies *Musca domestica*, *Ceratitis capitata*, and *Lucilia cuprina*, but absent in all acalyptrate flies, including representatives of the *Drosophila* family. Likely, the LTR proliferation in

drosophilids as well as the DNA transposon expansion in mosquitos and other flies was the result of a lineage-specific invasion and subsequent propagation into the different dipteran genomes.

In the calyptrate flies, Helitron elements are highly abundant, representing 28 % of the genome in the house fly *M. domestica* and 7 % in the blow fly *Lucilia cuprina*. These rolling circle elements are not as abundant in acalyptrate flies, except for the drosophilids *D. mojavensis*, *D. virilis*, *D. miranda*, and *D. pseudoobscura* (again with a bi-modal distribution). In the barley midge, *Mayetiola destructor*, DNA transposons occur across almost all Kimura distances between 0.02 and 0.45. The same holds true for LTR retrotransposons, although these show an increased expansion in the older age categories at Kimura distances between 0.37 and 0.44. LINEs and SINEs as well as Helitron elements show little occurrence in Diptera. In *B. antarctica*, LINE elements are the most prominent and exhibit a distribution across all Kimura distances up to 0.4. This may be a result of the overall low TE concentration in the small *B. antarctica* genome (less than 1 %) that introduces stochastic noise.

In Lepidoptera, we found a relatively recent SINE expansion event around Kimura distance 0.03 to 0.05. In fact, Lepidoptera and Trichoptera are the only holometabolous insect orders with a substantial SINE portion of up to 9 % in the silk worm *B. mori* (mean: 3.8 %). We observed that in the postman butterfly, *Heliconius melpomene*, the SINE fraction also appears with a divergence between Kimura distances 0.1 to around 0.31. Additionally, we found high LINE content in the monarch butterfly *Danaus plexippus* with a divergence ranging from Kimura distances 0 to 0.47 and a substantial fraction around Kimura distance 0.09.

In all Coleoptera species, we found substantial LINE and DNA content with a divergence around Kimura distance 0.1. In the beetle species *Onthophagus taurus*, *Agrilus planipennis*, and *L. decemlineata*, this fraction consists mostly of LINE copies, while in *T. castaneum* and *A. glabripennis* DNA elements make up the major fraction. In all Coleoptera species, the amount of SINEs and Helitrons is small (cf. Fig. 2.1). Interestingly, *Mengenilla moldrzyki*, a representative of Strepsiptera, which was previously determined to be the sister group of Coleoptera (Niehuis et al., 2012), shows more similarity in TE divergence distribution to Hymenoptera than to Coleoptera, with a large fraction of DNA elements covering Kimura distances 0.05 to around 0.3 and relatively small contributions from LINEs.

In apocritan Hymenoptera (*i.e.*, those with a wasp waist), the DNA element divergence distribution exhibits a peak around Kimura distance 0.01 to 0.05. In fact, the TE divergence distribution looks very similar among the ants and differs mostly in absolute coverage, except in *Camponotus floridanus*, which shows no such distinct peak. Instead, in *C. floridanus*, we found DNA elements and LTR elements with a relatively homogeneous coverage distribution between Kimura distances 0.03 and 0.4. *C. floridanus* is also the only hymenopteran species with a noticeable SINE proportion; this fraction's peak divergence is around Kimura distance 0.05. The relatively TE-poor genome of the honey bee, *Apis mellifera* contains a large fraction of Helitron elements with a Kimura distance between 0.1 and 0.35, as does *Nasonia vitripennis* with peak coverage around Kimura distance 0.15. These species-specific Helitron appearances are likely the result of an infection from a parasite or virus, as has been demonstrated in Lepidoptera (Coates, 2015). In the (non-apocritan) parasitic wood wasp, *O. abietinus*, the divergence distribution is similar to that in ants, with a dominant DNA transposon coverage around Kimura distance

0.05. The turnip sawfly, *A. rosae* has a large, zero-divergence fraction of DNA elements, LINEs and LTR retrotransposons followed by a bi-modal divergence distribution of DNA elements.

When examining Hemiptera, Thysanoptera, and Psocodea, the DNA element fraction with high divergence (peak Kimura distance 0.25) sets the psocodean *P. humanus* apart from Hemiptera and Thysanoptera. Additionally, *P. humanus* exhibits a large peak of LTR element coverage with a low divergence (Kimura distance 0). In Hemiptera and Thysanoptera, we found DNA elements with a high coverage around Kimura distance 0.05 instead of around 0.3, like in *P. humanus*, or only in miniscule amounts, such as in *Halyomorpha halys*. Interestingly, the three bug species *H. halys*, *Oncopeltus fasciatus*, and *Cimex lectularius* show a strikingly similar TE divergence distribution which differs from that in other species of Hemiptera. In these species, the TE landscape is characterized by a wide-ranging distribution of LINE divergence with peak coverage around Kimura distance 0.07. Further, they exhibit a shallow, but consistent proportion of SINE coverage with a divergence distribution between Kimura distance 0 and around 0.3. The other species of Hemiptera and Thysanoptera show no clear pattern of similarity. In the flower thrips *Frankliniella occidentalis* (Thysanoptera) as well as in the water strider *Gerris buenoi* and the cicadellid *Homalodisca vitripennis*, (Hemiptera), the Helitron elements show a distinct coverage between Kimura distances 0 and 0.3, with peak coverage at around 0.05 to 0.1 (*F. occidentalis*, *G. buenoi*) and 0.2 (*H. vitripennis*). In both *F. occidentalis* and *G. buenoi*, the divergence distribution is slightly bi-modal. In *H. vitripennis*, LINEs and DNA elements exhibit a divergence distribution with high coverage at Kimura distances 0.02 to around 0.45. SINEs and LTR element coverage is only slightly visible. This is in stark contrast to the findings in the pea aphid *Acyrthosiphon pisum*, where SINEs make up the majority of the TE

content and exhibit a broad spectrum of Kimura distances from 0 to 0.3, with peak coverage at around Kimura distance 0.05. Additionally, we found DNA elements in a similar distribution, but showing no clear peak. Instead, LINEs and LTR elements are distinctly absent from the *A. pisum* genome, possibly as a result of a lineage-specific extinction event.

The TE landscape in Polyneoptera is dominated by LINEs, which in the cockroach *Blattella germanica* have a peak coverage at around Kimura distance 0.04. In the termite *Zootermopsis nevadensis*, the peak LINE coverage is between Kimura distances 0.2 and 0.4. In the locust *L. migratoria*, LINE coverage shows a broad divergence distribution. Low-divergence LINEs show peak coverage at around Kimura distance 0.05. All three Polyneoptera species have a small, but consistent fraction of low-divergence SINE coverage with peak coverage between Kimura distances 0 to 0.05 as well as a broad, but shallow distribution of DNA element divergence.

LINEs also dominate the TE landscape in Paleoptera. The mayfly *E. danica* additionally exhibits a population of LTR elements with medium divergence in the genome. In the dragonfly *L. fulva*, we found DNA elements of similar coverage and divergence as the LTR elements. Both TE types have almost no low-divergence elements in *L. fulva*. In the early divergent apterygote hexapod orders Diplura (represented by the species *Catajapyx aquilonaris*) and Archaeognatha (*Machilis hrabei*), DNA elements are abundant with a broad divergence spectrum and low-divergence peak coverage. Additionally, we found other TE types with high coverage in low divergence regions in the genome of *C. aquilonaris* as well as SINE peak coverage at slightly higher divergence in *M. hrabei*.

The non-insect outgroup species also exhibit a highly heterogeneous TE copy divergence spectrum. In all species, we found high coverage of varying TE types with low divergence. All

chelicerate genomes contain mostly DNA transposons, with LINEs and SINEs contributing a fraction in the spider *Parasteatoda tepidariorum* and the tick *I. scapularis*. The only available myriapod genome, that of the centipede *Strigamia maritima*, is dominated by LTR elements with high coverage in a low-divergence spectrum, but also LTR elements that exhibit a higher Kimura distance. We found the same in the crustacean *Daphnia pulex*, but the TE divergence distribution in the other crustacean species was different and consisted of more DNA transposons in the copepod *E. affinis*, or LINEs in the amphipod *Hyalella azteca*.

2.4 Discussion

We used species-specific TE libraries to assess the genomic retrotransposable and transposable element content in sequenced and assembled genomes of arthropod species, including most extant insect orders.

2.4.1 TE content contributes to genome size in arthropods

TEs and other types of DNA repeats are an omnipresent part of metazoan, plant, as well as fungal genomes and are found in variable proportions in sequenced genomes of different species. In vertebrates and plants, studies have shown that TE content is a predictor for genome size (Chalopin et al., 2015; Staton & Burke, 2015). For insects, this has also been reported in clade-specific studies such as those on mosquitoes (Neafsey et al., 2015) and *Drosophila* fruit flies (Sessegolo et al., 2016). These observations lend further support to the hypothesis that genome size is also correlated with TE content in insects on a pan-ordinal scale.

Our analysis shows that both genome size and TE content are highly variable among the investigated insect genomes, even in comparative contexts with low variation in genome size. While non-holometabolous hexapods have a significantly smaller genome than holometabolous insects, the TE content is not significantly different. Still, we found that TE content contributes significantly to genome size in hexapods as a whole. These results are in line with prior studies on insects with a more limited taxon sampling reporting a clade-specific correlation between TE content and genome size (Vieira et al., 1999, 2002; Kidwell & Lisch, 2000; Honeybee Genome Sequencing Consortium, 2006; Bosco et al., 2007; Sessegolo et al., 2016), and expand that finding to larger taxon sampling covering most major insect orders. These findings further support the hypothesis that TEs are a major factor in the dynamics of genome size evolution in Eukaryotes. While differential TE activity apparently contributes to genome size variation (Petrov, 2001; Kidwell, 2002; Ågren & Wright, 2011), whole genome duplications, such as suggested by integer-sized genome size variations in some representatives of Hymenoptera (Li et al., 2018), segmental duplications, deletions, and other repeat proliferation (Parfrey et al., 2008) could contribute as well. This variety of influencing factors potentially explains the range of dispersion in the correlation.

The high range of dispersion in the correlation of TE content and genome size is most likely also amplified by heterogeneous underestimates of the genomic TE coverage. Most of the genomes were sequenced and assembled using different methods, and with insufficient sequencing depth and/or older assembly methods; the data are therefore almost certainly incomplete with respect to repeat-rich regions. Assembly errors and artifacts also add a possible error margin, as assemblers cannot reconstruct repeat regions that are longer than the insert size accurately

from short reads (Schatz et al., 2010; Sambaturu, 2014; Chaisson et al., 2015; Peona et al., 2018) and most available genomes were sequenced using short read technology only. Additionally, RepeatMasker is known to underestimate the genomic repeat content (de Koning et al., 2011). By combining RepeatModeler to infer the species-specific repeat libraries and RepeatMasker to annotate the species-specific repeat libraries in the genome assemblies, our methods are purposefully conservative and may have missed some TE types, or ancient and highly divergent copies.

This underestimation of the TE content notwithstanding, we found many TE families that were previously thought to be restricted to, for example, mammals, such as the SINE family Alu (Kriegs et al., 2007) and the LINE family L1 (Liu, 2003), or to fungi, such as Tad1 (Cambareri et al., 1994). Essentially, most known superfamilies were found in the investigated insect genomes (*cf.* Fig. 2.3) and additionally, we identified highly abundant unclassifiable TEs in all insect species. These observations suggest that the insect mobilome (the entirety of mobile DNA elements) is more diverse than the well characterized vertebrate mobilome (Chalopin et al., 2015) and requires more exhaustive characterization. We were able to reach these conclusions by relying on two essential non-standard analyses. First, our annotation strategy of *de novo* repeat library construction and classification according to the RepBase database was more specific to each genome than the default RepeatMasker analysis using only the RepBase reference library. The latter approach is usually done when releasing a new genome assembly to the public. The second difference between our approach and the conventional application of the RepBase library was that we used the entire Metazoa-specific section of RepBase instead of restricting our search to Insecta. This broader scope allowed us to annotate TEs that were previously unknown from insects, and that would otherwise have been overlooked. Additionally, by removing re-

sults that matched non-TE sequences in the NCBI database, our annotation becomes more robust against false positives. The enormous previously overlooked diversity of TEs in insects does not seem to be surprising given the geological age and species richness of this clade. Insects originated more than 450 million years ago (Misof et al., 2014) and represent over 80 % of the described metazoan species (Grimaldi & Engel, 2005). Further investigations will also show whether there is a connection between TE diversity or abundance and clade-specific genetic and genomic traits, such as the sex determination system (*e.g.*, butterflies have Z and W chromosomes instead of X and Y (Traut & Marec, 1997)) or the composition of telomeres, which have been shown in *D. melanogaster* to exhibit a high density of TEs (Levis et al., 1993), whereas telomeres in other insects consist mostly of simple repeats. It remains to be analyzed in detail, however, whether insect TE diversity evolved independently within insects or is the result of multiple TE introgression into insect genomes.

Our results show that virtually all known TE classes are present in all investigated insect genomes. However, a large part of the TEs we identified remains unclassifiable despite the diversity of metazoan TEs in the reference library RepBase. This abundance of unclassifiable TEs suggests that the insect TE repertoire requires more exhaustive characterization and that our understanding of the insect mobilome is far from complete.

It has been hypothesized that population-level processes might contribute to TE content differences and genome size variation in vertebrates (Lynch & Conery, 2003). In insects, it has been shown that TE activity also varies on the population level, for example in the genomes of *Drosophila* spp. (Perrat et al., 2013; Li et al., 2013; Blumenstiel et al., 2014) or in the genome of the British peppered moth *Biston betularia*, in which a tandemly repeated TE confers an

adaptive advantage in response to short-term environmental changes (van't Hof et al., 2016). The TE activity within populations is expected to leave footprints in the nucleotide sequence diversity of TEs in the genome as recent bursts of TEs should be detectable by a large number of TE sequences with low sequence divergence.

To explain TE proliferation dynamics, two different models of TE activity have been proposed: the equilibrium model and the burst model. In the equilibrium model, TE proliferation and elimination rates are more or less constant and cancel each other out at a level that is different for each genome (Charlesworth & Charlesworth, 1983). In this model, differential TE elimination rate contributes to genome size variation when TE activity is constant. This model predicts that in species with a slow rate of DNA loss, genome size tends to increase (Petrov et al., 2011; Sun et al., 2011). In the burst model, TEs do not proliferate at a constant rate, but rather in high copy rate bursts following a period of inactivity (Blumenstiel et al., 2014). These bursts can be TE family specific. Our analysis of TE landscape diversity (see below), supports the burst hypothesis. In almost every species we analyzed, there is a high proportion of abundant TE sequences with low sequence divergence and the most abundant TEs are different even among closely related species. It was hypothesized that TE bursts enabled by periods of reduced efficiency in counteracting host defense mechanisms such as TE silencing (Le Rouzic & Capy, 2006; Rebollo et al., 2010) have resulted in differential TE contribution to genome size.

2.4.2 TE landscape diversity in arthropods

In vertebrates, it is possible to trace lineage-specific contributions of different TE types (Chalopin et al., 2015). In insects, however, the TE composition shows a statistically significant correlation

to genome size, but a high range of dispersion. Instead, we can show that major differences both in TE abundance and diversity exist between species of the same lineage (Fig. 2.3). Using the Kimura nucleotide sequence distance, we observe distinct variation, but also similarities, in TE composition and activity between insect orders and among species of the same order. The number of recently active elements can be highly variable, such as LTR retrotransposons in fruit flies or DNA transposons in ants (Fig. 2.5). On the other hand, the shape of the TE coverage distributions can be fairly similar among species of the same order; this is particularly visible in Hymenoptera and Diptera. These findings suggest lineage-specific similarities in TE elimination mechanisms; possibly shared efficacies in the piRNA pathway that silences TEs during transcription in metazoans (*e.g.*, in *Drosophila* (Le Thomas et al., 2013; Yamashiro & Siomi, 2017), *B. mori* (Matsumoto et al., 2016), *Caenorhabditis elegans* (Zhang et al., 2018), and mouse (Tóth et al., 2016). Another possible explanation would be recent horizontal transfers from, for example, parasite to host species (see below).

2.4.3 Can we infer an ancestral arthropod mobilome in the face of massive horizontal TE transfer?

In a purely vertical mode of TE transmission, the genome of the last common ancestor (LCA) of insects — or arthropods — can be assumed to possess a superset of the TE superfamilies present in extant insect species. As many TE families appear to have been lost due to lineage-specific TE extinction events, the ancestral TE repertoire may have been even more extensive compared with the TE repertoire of extant species and might have included almost all known metazoan TE superfamilies such as the CMC complex, Ginger, Helitron, Mavericks, Jockey,

L1, Penelope, R1, DIRS, Ngaro, and Pao. Many SINEs found in extant insects were most likely part of the ancestral mobilome as well, for example Alu, which was previously thought to be restricted to primates (Deininger, 2011), and MIR.

The mobilome in extant species, however, appears to be the product of both vertical and horizontal transmission. In contrast to a vertical mode of transmission, horizontal gene transfers, common phenomenona among prokaryotes (and making a prokaryote species phylogeny nigh meaningless) and widely occurring in plants, are rather rare in vertebrates (Syvanen, 2012; Wallau et al., 2012), but have been described in Lepidoptera (Sormacheva et al., 2012) and other insects (Nakabachi, 2015). Recently, a study uncovered large-scale horizontal transfer of TEs (horizontal transposon transfer, HTT) among insects (Peccoud et al., 2017) and makes this mechanism even more likely to be the source of inter-lineage similarities in insect genomic TE composition. In the presence of massive HTT, the ancestral mobilome might be impossible to infer because the effects of HTT overshadow the result of vertical TE transfer. It remains to be analyzed in detail whether the high diversity of the insect mobilomes can be better explained by massive HTT events.

2.5 Conclusions

The present study provides an overview of the diversity and evolution of TEs in the genomes of major lineages of extant insects. The results show that there is large intra- and inter-lineage variation in both TE content and composition. This, and the highly variable age distribution of individual TE superfamilies, indicate a lineage-specific burst-like mode of TE proliferation in insect genomes. In addition to the complex composition patterns that can differ even among

species of the same genus, there is a large fraction of TEs that remain unclassified, but often make up the major part of the genomic TE content, indicating that the insect mobilome is far from completely characterized. This study provides a solid baseline for future comparative genomics research. The functional implications of lineage-specific TE activity for the evolution of genome architecture will be the focus of future investigations.

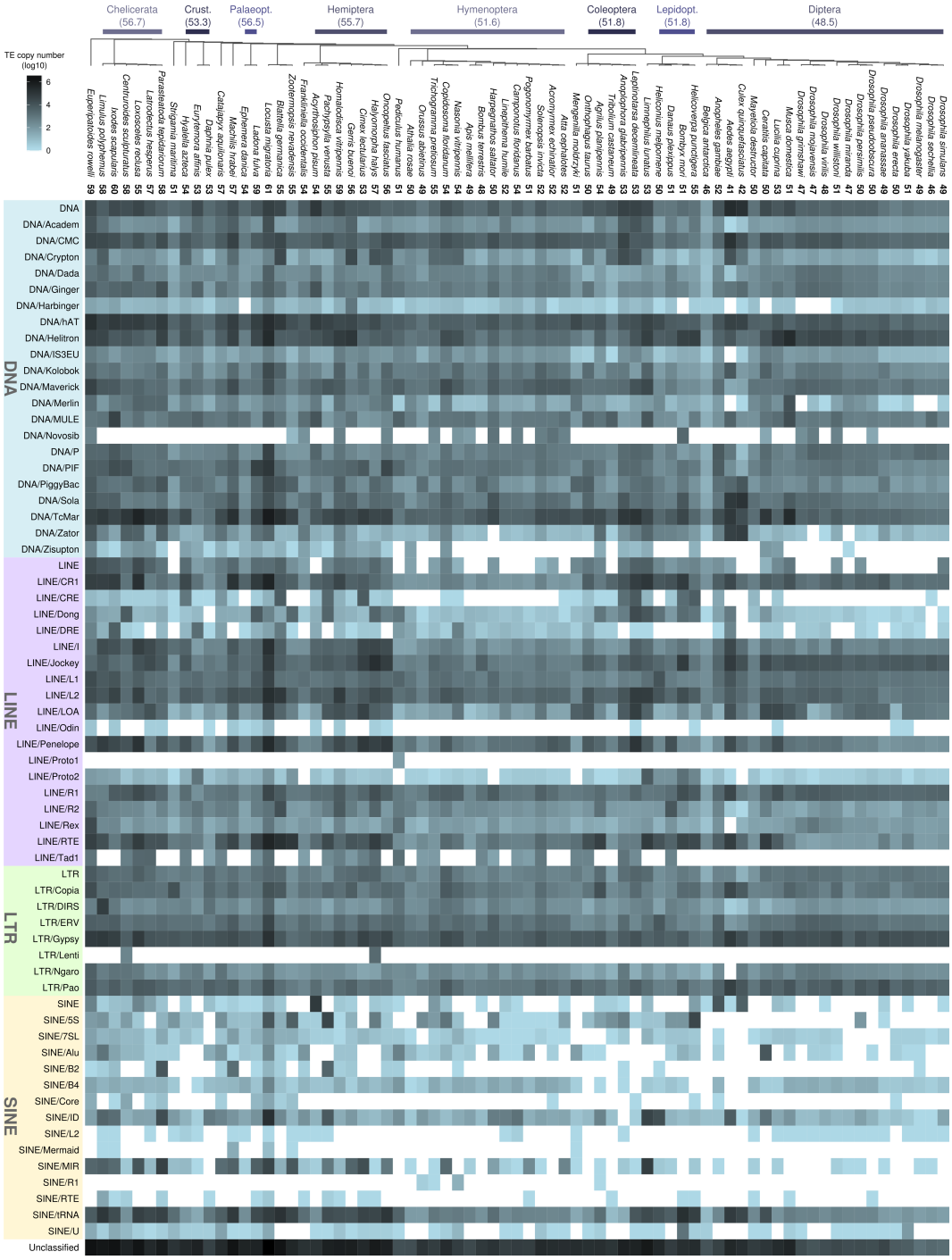

Figure 2.3: TE diversity in arthropod genomes: Many known TE superfamilies were identified in almost all insect species. Presence of TE superfamilies is shown as filled cells with the color gradient showing the TE copy number (log11). Empty cells represent absence of TE superfamilies. The numbers after each species name show the number of different TE superfamilies; numbers in parentheses below clade names denote the average number of TE superfamilies in the corresponding taxon.

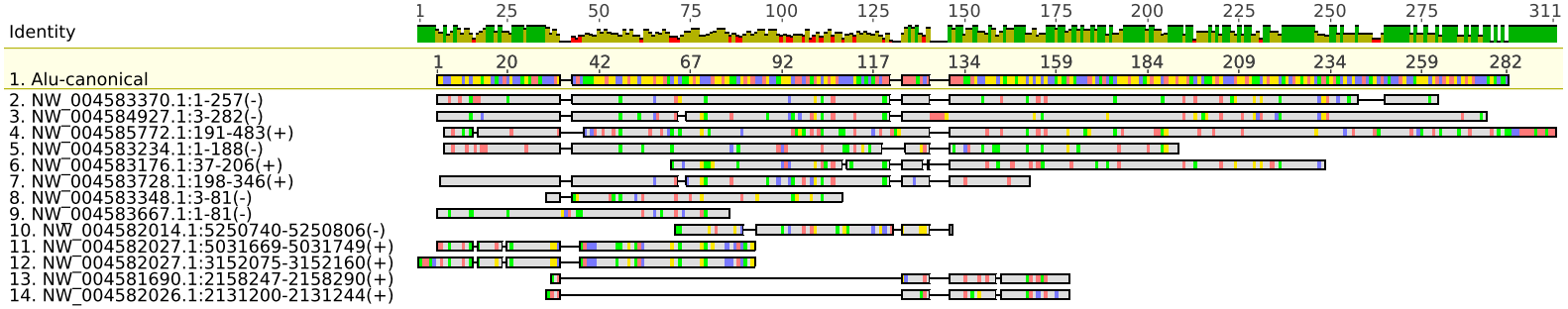

Figure 2.4: The Alu element found in *Bombyx mori*: Alignment of the canonical Alu sequence from Repbase with HMM hits in the *B. mori* genome assembly. Grey areas in the sequences are identical to the canonical Alu sequence. The sequence names follow the pattern "identifier:start-end(strand)" Image created using Geneious version 7.1 created by Biomatters. Available from `https://www.geneious.com`

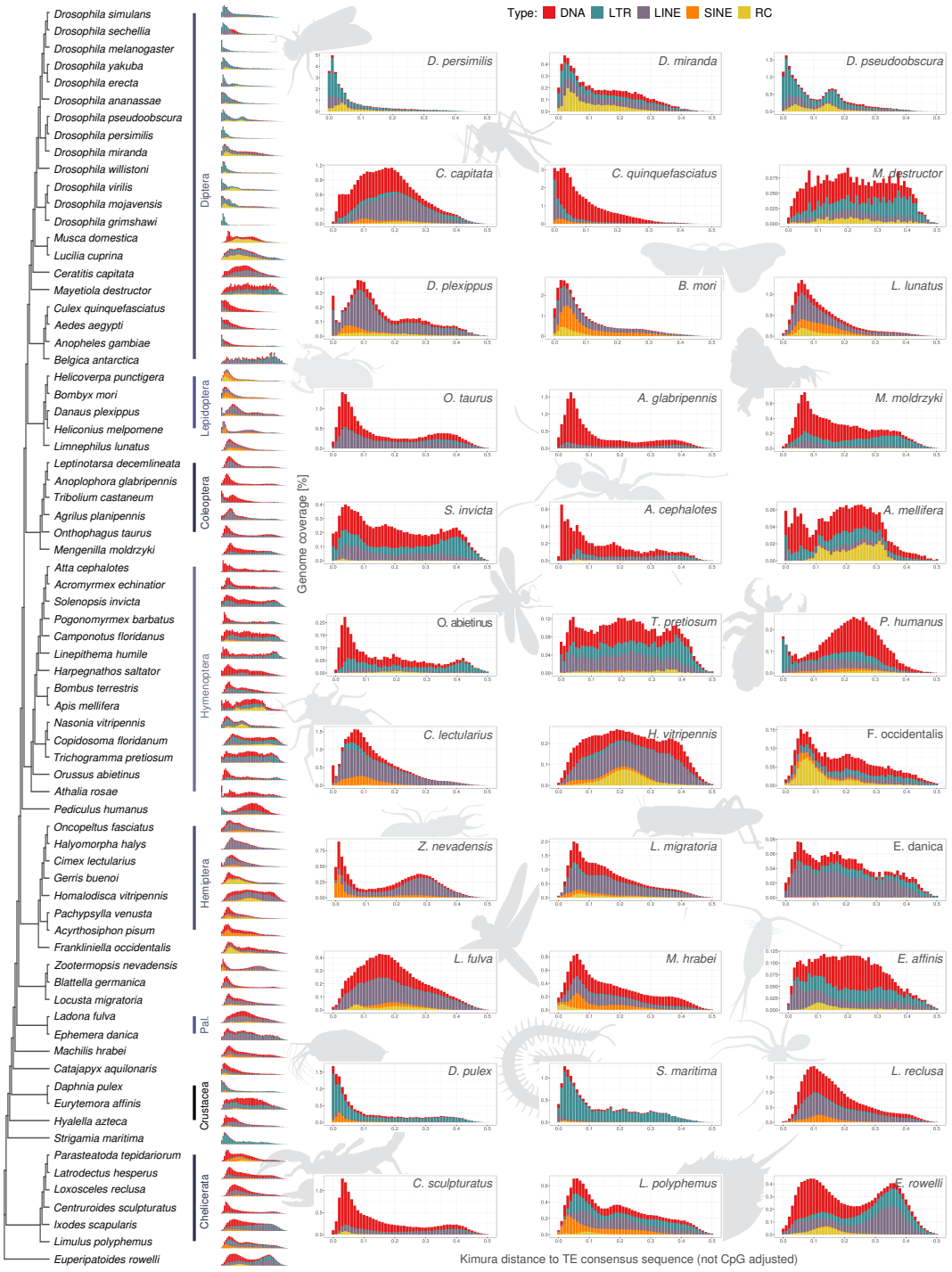

Figure 2.5: Cladogram with repeat landscape plots. The larger plots are selected representatives. The further to the left a peak in the distribution is, the younger the corresponding TE fraction generally is (low TE intra-family sequence divergence). In most orders, the TE divergence distribution is similar, such as in Diptera or Hymenoptera. The large fraction of unclassified elements was omitted for these plots. Pal., Palaeoptera

References

Ågren, J. A. & Wright, S. I. (2011). Co-evolution between transposable elements and their hosts: A major factor in genome size evolution? *Chromosome Research*, 19(6), 777–786.

Adams, M. D. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461), 2185–2195.

Bailly-Bechet, M., Haudry, A., & Lerat, E. (2014). "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, 5, 13.

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population Genomics of Transposable Elements in Drosophila. *Annual Review of Genetics*, 48(1), 561–581.

Beutel, R. (2013). *Insect Morphology and Phylogeny*. De Gruyter Graduate. Berlin ; New York: De Gruyter.

Blumenstiel, J. P., Chen, X., He, M., & Bergman, C. M. (2014). An Age-of-Allele Test of Neutrality for Transposable Element Insertions. *Genetics*, 196(2), 523–538.

Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., & Volff, J.-N. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research*, 16(1), 203–215.

Bosco, G., Campbell, P., Leiva-Neto, J. T., & Markow, T. A. (2007). Analysis of Drosophila Species Genome Size and Satellite DNA Content Reveals Significant Differences Among Strains as Well as Between Species. *Genetics*, 177(3), 1277–1290.

Bouallègue, M., Filée, J., Kharrat, I., Mezghani-Khemakhem, M., Rouault, J.-D., Makni, M., & Capy, P. (2017). Diversity and evolution of mariner-like elements in aphid genomes. *BMC Genomics*, 18(1).

Burns, K. H. & Boeke, J. D. (2012). Human Transposon Tectonics. *Cell*, 149(4), 740–752.

Cambareri, E., Helber, J., & Kinsey, J. (1994). Tadl-1 an active LINE-like element of Neurospora crassa. *MGG Molecular & General Genetics*, 242(6).

Casola, C., Lawing, A. M., Betrán, E., & Feschotte, C. (2007). PIF-like Transposons are Common in Drosophila and Have Been Repeatedly Domesticated to Generate New Host Genes. *Molecular Biology and Evolution*, 24(8), 1872–1888.

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, 16(11), 627–640.

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biology and Evolution*, 7(2), 567–580.

Charlesworth, B. & Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetical Research*, 42(01), 1.

Chen, S. & Li, X. (2007). Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes. *BMC Evol Biol*, 7(1), 46.

Chenais, B. (2015). Transposable elements in cancer and other human diseases. *Current Cancer Drug Targets*, 15(3), 227–242.

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, 18(2), 71–86.

Coates, B. S. (2015). Horizontal transfer of a non-autonomous Helitron among insect and viral genomes. *BMC Genomics*, 16(1), 137.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics*, 7(12), e1002384.

de la Chaux, N. & Wagner, A. (2011). BEL/Pao retrotransposons in metazoan genomes. *BMC Evol Biol*, 11(1), 154.

Deininger, P. (2011). Alu elements: Know the SINEs. *Genome Biology*, 12(12), 236.

Eickbush, T. H. & Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Research*, 134(1-2), 221–234.

Ellison, C. E. & Bachtrog, D. (2013). Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. *Science*, 342(6160), 846–850.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1), 1–15.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, 9(5), 397–405.

Feschotte, C. & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual review of genetics*, 41, 331–368.

Flavell, A. J. (1992). Ty1-copia group retrotransposons and the evolution of retroelements in the eukaryotes. *Genetica*, 86(1-3), 203–214.

Gahan, L. J. (2001). Identification of a Gene Associated with Bt Resistance in Heliothis virescens. *Science*, 293(5531), 857–860.

Glastad, K. M., Hunt, B. G., & Goodisman, M. A. (2014). Evolutionary insights into DNA methylation in insects. *Current Opinion in Insect Science*, 1, 25–30.

González, J., Karasov, T. L., Messer, P. W., & Petrov, D. A. (2010). Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila. *PLoS Genetics*, 6(4), e1000905.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., & Petrov, D. A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biology*, 6(10), e251.

Grimaldi, D. A. & Engel, M. S. (2005). *Evolution of the Insects*. Cambridge [U.K.] ; New York: Cambridge University Press.

Hancks, D. C. & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, 7(1).

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114), 931–949.

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2015). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, (pp. gkv1272).

Itokawa, K., Komagata, O., Kasai, S., Okamura, Y., Masada, M., & Tomita, T. (2010). Genomic structures of Cyp9m10 in pyrethroid resistant and susceptible strains of Culex quinquefasciatus. *Insect Biochemistry and Molecular Biology*, 40(9), 631–640.

Jurka, J., Kapitonov, V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4), 462–467.

Kapitonov, V. V. & Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 98(15), 8714–8719.

Kapitonov, V. V. & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 103(12), 4540–4545.

Kapitonov, V. V. & Jurka, J. (2007). Helitrons on a roll: Eukaryotic rolling-circle transposons. *Trends in Genetics*, 23(10), 521–529.

Kelley, J. L., Peyton, J. T., Fiston-Lavier, A.-S., Teets, N. M., Yee, M.-C., Johnston, J. S., Bustamante, C. D., Lee, R. E., & Denlinger, D. L. (2014). Compact Genome of the Antarctic Midge Is Likely an Adaptation to an Extreme Environment. *Nature Communications*, 5.

Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Phil. Trans. R. Soc. B*, 372(1736), 20160458.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1), 49–63.

Kidwell, M. G. & Lisch, D. R. (2000). Transposable elements and host genome evolution. *Trends in Ecology & Evolution*, 15(3), 95–99.

Kim, Y. B., Oh, J. H., McIver, L. J., Rashkovetsky, E., Michalak, K., Garner, H. R., Kang, L., Nevo, E., Korol, A. B., & Michalak, P. (2014). Divergence of Drosophila melanogaster repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel. *Proceedings of the National Academy of Sciences*, 111(29), 10630–10635.

Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., & Schmitz, J. (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends in Genetics*, 23(4), 158–161.

Krupovic, M. & Koonin, E. V. (2016). Self-synthesizing transposons: Unexpected key players in the evolution of viruses and defense systems. *Current Opinion in Microbiology*, 31, 25–33.

Le Rouzic, A. & Capy, P. (2006). Theoretical Approaches to the Dynamics of Transposable Elements in Genomes Populations, and Species. In *Transposons and the Dynamic Genome* (pp. 1–19). Springer Science Business Media.

Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., Perkins, E. M., Hur, J. K., Aravin, A. A., & Toth, K. F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes & Development*, 27(4), 390–399.

Levis, R. W., Ganesan, R., Houtchens, K., Tolar, L. A., & Sheen, F.-m. (1993). Transposons in place of telomeric repeats at a Drosophila telomere. *Cell*, 75(6), 1083–1093.

Li, W., Prazak, L., Chatterjee, N., Grüninger, S., Krug, L., Theodorou, D., & Dubnau, J. (2013). Activation of transposable elements during aging and neuronal decline in Drosophila. *Nature Neuroscience*, 16(5), 529–531.

Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., & Barker, M. S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences*, 115(18), 4713–4718.

Liu, G. (2003). Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome. *Genome Research*, 13(3), 358–368.

Lynch, M. & Conery, J. S. (2003). The evolutionary demography of duplicate genes. In *Genome Evolution* (pp. 35–44). Springer Nature.

Mackay, T. F. C. (1989). Transposable elements and fitness in Drosophila melanogaster. *Genome*, 31(1), 284–295.

Malik, H. S., Burke, W. D., & Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution*, 16(6), 793–805.

Marin, I. & Llorens, C. (2000). Ty3/Gypsy Retrotransposons: Description of New Arabidopsis thaliana Elements and Evolutionary Perspectives Derived from Comparative Genomic Data. *Molecular Biology and Evolution*, 17(7), 1040–1049.

Matsumoto, N., Nishimasu, H., Sakakibara, K., Nishida, K. M., Hirano, T., Ishitani, R., Siomi, H., Siomi, M. C., & Nureki, O. (2016). Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. *Cell*, 167(2), 484–497.e9.

Maumus, F., Fiston-Lavier, A.-S., & Quesneville, H. (2015). Impact of Transposable Elements on Insect Genomes and Biology. *Current Opinion in Insect Science*, 7, 30–36.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von, R. B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K., & Zhou, X. (2014). Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science*, 346, 763–7.

Nakabachi, A. (2015). Horizontal gene transfers in insects. *Current Opinion in Insect Science*, 7, 24–29.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arca, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., Burt, A., Chan, C. S., Chauve, C., Chiu, J. C., Christensen, M., Costantini, C., Davidson, V. L. M., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S. B., Guelbeogo, W. M., Hall, A. B., Han, M. V., Hlaing, T., Hughes, D. S. T., Jenkins, A. M., Jiang, X., Jungreis, I., Kakani, E. G., Kamali, M., Kemppainen, P., Kennedy, R. C., Kirmitzoglou, I. K., Koekemoer, L. L., Laban, N., Langridge, N., Lawniczak, M. K. N., Lirakis, M., Lobo, N. F., Lowy, E., MacCallum, R. M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S. N., Moore, W., Murphy, K. A., Naumenko, A. N., Nolan, T., Novoa, E. M., O'Loughlin, S., Oringanje, C., Oshaghi, M. A., Pakpour, N., Papathanos, P. A., Peery, A. N., Povelones, M., Prakash, A., Price, D. P., Rajaraman, A., Reimer, L. J., Rinker, D. C., Rokas, A., Russell, T. L., Sagnon, N., Sharakhova, M. V., Shea, T., Simao, F. A., Simard, F., Slotman, M. A., Somboon, P., Stegniy, V., Struchiner, C. J., Thomas, G. W. C., Tojo, M., Topalis, P., Tubio, J. M. C., Unger, M. F., Vontas, J., Walton, C., Wilding, C. S., Willis, J. H., Wu, Y.-C., Yan, G., Zdobnov, E. M., Zhou, X., Catteruccia, F., Christophides, G. K., Collins, F. H., Cornman, R. S., Crisanti, A., Donnelly, M. J., Emrich, S. J., Fontaine, M. C., Gelbart, W., Hahn, M. W., Hansen, I. A., Howell, P. I., Kafatos, F. C., Kellis, M., Lawson, D., Louis, C., Luckhart, S., Muskavitch, M. A. T., Ribeiro, J. M., Riehle, M. A., Sharakhov, I. V., Tu, Z., Zwiebel, L. J., & Besansky, N. J. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, 347(6217), 1258522–1258522.

Niehuis, O., Hartig, G., Grath, S., Pohl, H., Lehmann, J., Tafer, H., Donath, A., Krauss, V.,

Eisenhardt, C., Hertel, J., Petersen, M., Mayer, C., Meusemann, K., Peters, R. S., Stadler, P. F., Beutel, R. G., Bornberg-Bauer, E., McKenna, D. D., & Misof, B. (2012). Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera. *Current Biology*, 22(14), 1309–1313.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289–290.

Parfrey, L. W., Lahr, D. J. G., & Katz, L. A. (2008). The dynamic nature of eukaryotic genomes. *Molecular Biology and Evolution*, 25(4), 787–794.

Pasyukova, E. G. (2004). Accumulation of Transposable Elements in the Genome of Drosophila melanogaster is Associated with a Decrease in Fitness. *Journal of Heredity*, 95(4), 284–290.

Peccoud, J., Loiseau, V., Cordaux, R., & Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. *Proceedings of the National Academy of Sciences*, 114(18), 4721–4726.

Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are "complete" genome assemblies?-An avian perspective. *Molecular Ecology Resources*.

Perrat, P. N., DasGupta, S., Wang, J., Theurkauf, W., Weng, Z., Rosbash, M., & Waddell, S. (2013). Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. *Science*, 340(6128), 91–95.

Petersen, M., Armisén, D., Gibbs, R. A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., & Misof, B. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evolutionary Biology*, 19(1).

Petrov, D. A. (2001). Evolution of genome size: New approaches to an old problem. *Trends in Genetics*, 17(1), 23–28.

Petrov, D. A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K., & Gonzalez, J. (2011). Population Genomics of Transposable Elements in Drosophila melanogaster. *Molecular Biology and Evolution*, 28(5), 1633–1644.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rebollo, R., Horard, B., Hubert, B., & Vieira, C. (2010). Jumping genes and epigenetics: Towards new species. *Gene*, 454(1-2), 1–7.

Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., Lawson, D., Okamuro, J., Robertson, H. M., Schneider, D. J., et al. (2011). Creating a buzz about insect genomes. *Science*, 331(6023), 1386–1386.

Sambaturu, N. (2014). *Towards Handling Repeats in Genome Assembly*. PhD thesis, National University of Singapore, Singapore.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., & Bennetzen, J. L. (1996). Nested

Retrotransposons in the Intergenic Regions of the Maize Genome. *Science*, 274(5288), 765–768.

Santos, M. E., Braasch, I., Boileau, N., Meyer, B. S., Sauteur, L., Böhne, A., Belting, H. G., Affolter, M., & Salzburger, W. (2014). The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. *Nature Communications*, 5.

Schatz, M., Delcher, A., & Salzberg, S. (2010). Assembly of large genomes using second-generation sequencing. *Genome Res*, 20, 1165–73.

Sessegolo, C., Burlet, N., & Haudry, A. (2016). Strong Phylogenetic Inertia on Genome Size and Transposable Element Content among 26 Species of Flies. *Biology Letters*, 12(8), 20160407.

Shirasawa, K., Hirakawa, H., Tabata, S., Hasegawa, M., Kiyoshima, H., Suzuki, S., Sasamoto, S., Watanabe, A., Fujishiro, T., & Isobe, S. (2012). Characterization of active miniature inverted-repeat transposable elements in the peanut genome. *Theor Appl Genet*, 124(8), 1429–1438.

Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0.

Sormacheva, I., Smyshlyaev, G., Mayorov, V., Blinov, A., Novikov, A., & Novikova, O. (2012). Vertical Evolution and Horizontal Transfer of CR1 Non-LTR Retrotransposons and Tc1/mariner DNA Transposons in Lepidoptera Species. *Molecular Biology and Evolution*, 29(12), 3685–3702.

Staton, S. E. & Burke, J. M. (2015). Evolutionary Transitions in the Asteraceae Coincide with Marked Shifts in Transposable Element Abundance. *BMC Genomics*, 16(1).

Sun, C., Shepard, D. B., Chong, R. A., Arriaza, J. L., Hall, K., Castoe, T. A., Feschotte, C., Pollock, D. D., & Mueller, R. L. (2011). LTR Retrotransposons Contribute to Genomic Gigantism in Plethodontid Salamanders. *Genome Biology and Evolution*, 4(2), 168–183.

Syvanen, M. (2012). Evolutionary Implications of Horizontal Gene Transfer. *Annu. Rev. Genet.*, 46(1), 341–358.

Tóth, K. F., Pezic, D., Stuwe, E., & Webster, A. (2016). The piRNA Pathway Guards the Germline Genome Against Transposable Elements. In D. Wilhelm & P. Bernard (Eds.), *Non-Coding RNA and the Reproductive System*, volume 886 (pp. 51–77). Dordrecht: Springer Netherlands.

Traut, W. & Marec, F. (1997). Sex Chromosome Differentiation in Some Species of Lepidoptera (Insecta). *Chromosome Research*, 5(5), 283–291.

van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., & Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534(7605), 102–105.

Vieira, C., Lepetit, D., Dumont, S., & Biémont, C. (1999). Wake up of transposable elements following Drosophila simulans worldwide colonization. *Molecular Biology and Evolution*, 16(9), 1251–1255.

Vieira, C., Nardon, C., Arpin, C., Lepetit, D., & Biemont, C. (2002). Evolution of Genome Size in Drosophila. Is the Invader's Genome Being Invaded by Transposable Elements? *Molecular Biology and Evolution*, 19(7), 1154–1161.

Vorechovsky, I. (2009). Transposable elements in disease-associated cryptic exons. *Hum Genet*, 127(2), 135–154.

Wallace, A. G., Detweiler, D., & Schaeffer, S. W. (2011). Evolutionary history of the third chromosome gene arrangements of Drosophila pseudoobscura inferred from inversion breakpoints. *Molecular Biology and Evolution*, 28(8), 2219–2229.

Wallau, G. L., Ortiz, M. F., & Loreto, E. L. S. (2012). Horizontal Transposon Transfer in Eukarya: Detection, Bias, and Perspectives. *Genome Biology and Evolution*, 4(8), 801–811.

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., Hao, S., Chen, B., Ma, Z., Yu, D., Xiong, Z., Zhu, Y., Fan, D., Han, L., Wang, B., Chen, Y., Wang, J., Yang, L., Zhao, W., Feng, Y., Chen, G., Lian, J., Li, Q., Huang, Z., Yao, X., Lv, N., Zhang, G., Li, Y., Wang, J., Wang, J., Zhu, B., & Kang, L. (2014). The Locust Genome Provides Insight into Swarm Formation and Long-Distance Flight. *Nature Communications*, 5.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 8(12), 973–982.

Yamashiro, H. & Siomi, M. C. (2017). PIWI-Interacting RNA in Drosophila: Biogenesis Transposon Regulation, and Beyond. *Chemical Reviews*, 118(8), 4404–4421.

Zhang, D., Tu, S., Stubna, M., Wu, W.-S., Huang, W.-C., Weng, Z., & Lee, H.-C. (2018). The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. *Science*

(New York, N.Y.), 359(6375), 587–592.

Zhang, X. H.-F. & Chasin, L. A. (2006). Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proceedings of the National Academy of Sciences*, 103(36), 13427–13432.

3

Dynamics of genome size evolution in insects

This chapter is intended for publication in *PNAS*.

Authors: Petersen, M., Nottebrock G., Misof, B.

Author contributions: Analyses: MP, GN, figures: MP, GN, manuscript design and writing: MP, GN, BM

3.1 Introduction

Genome size variation is an important aspect of eukaryote genome evolution (Gregory, 2005; Petrov, 2001) and seems positively correlated with cell size (Dufresne & Jeffery, 2011), and body size in invertebrates (Gregory & Johnston, 2008). It has also been reported that genome size is positively correlated with cell division time (Bennett et al., 1977) and developmental rate (White & McLaren, 2000). Genome size variation, however, does not appear to be correlated with organismic complexity (the so-called c-value enigma (Gregory & Johnston, 2008)). It is currently unclear how genome size variation and the evolution of phenotypic traits are correlated.

Genome size can expand because of, for example, whole or partial genome duplications, or the accumulation of transposable elements (Bennetzen et al., 2005; Piegu et al., 2006; Vitte et al., 2007; Kelly et al., 2015; Nystedt et al., 2013; Blass et al., 2012; Neafsey & Palumbi, 2003; Sun et al., 2012; Sato & Nishida, 2010; Marburger et al., 2018; Kapusta et al., 2017). Genome size expansion and contraction have been found to be correlated with the frequency of transposable elements (TEs) in the genome (Sotero-Caio et al., 2017; Kapusta et al., 2017; Petrov et al., 1996, 2000). TEs play a pivotal role in functional adaptation and genome evolution in general that is not yet fully understood (reviewed in Maumus et al. (2015); Arkhipova (2018)).

For mammals and birds, an "accordion" model of genome size dynamics has been proposed Kapusta et al. (2017) that predicts a more or less stable genome size over time. This genome size stability in the presence of TE activity has also been termed the equilibrium model by Charlesworth & Charlesworth (1983). According to this model, genome expansion due to TE activity is counteracted by DNA removal, resulting in little fluctuations in genome size. In

arthropods, however, where most taxonomic orders are much older than the mammal and bird lineages (Misof et al., 2014), genome size varies strongly (Alfsnes et al., 2017; Petersen et al., 2019); this variation has been connected to, but is not entirely explained by TE content. Additionally, the TE landscape suggests a more burst-like activity profile in arthropods (Petersen et al., 2019).

In the present study, we exploit the growing number of sequenced arthropod genomes and assess genomic DNA gain and loss caused by TE activity. Our results are inconsistent with an "accordion" or equilibrium model of genome size dynamics, therefore we propose that genome size in insects is governed by different, more burst-like dynamics than in mammals and birds.

3.2 Materials and Methods

Ancestral genome size estimation

To determine the age of TE copies in the insect genomes, we first estimated clade-specific ancestral genome sizes with the approach described in the following. We sourced the Animal Genome Size Database (Gregory, 2018) (`http://genomesize.com`, accessed 2018-05-01) to obtain genome size estimates for 1,514 arthropod species. Additionally, we exploited the BOLD database (Ratnasingham & Hebert, 2007) (`http://www.boldsystems.org`, accessed 2018-03-19) to obtain 1,571,820 COI barcode nucleotide sequences for 105,397 arthropod species. Of these, we identified 605 species that were represented in both databases. We included our own genome size estimates for eight additional species (Supplemental Table D.2), bringing the total number of species to 613.

For the dipluran *Catajapyx aquilonaris*, which was not represented in the BOLD database, we added COI data by retrieving a COI sequence from the closely related species *Gollumjapyx smeagol* (GenBank ID DQ993154.1) and using it as query in a BLAST search in the genome assembly provided by the i5k initiative (source see Supplemental Table D.1). We received two hits and used the longer one (scaffold1312 47_cov1551 positions 852-1522) as query in a BLAST search in GenBank. The reciprocal search hit the mitochondrial genome of *Japyx solifugus* (Accession AY771989.1), another closely related species, which confirmed that our hit was indeed the COI sequence of *C. aquilonaris*.

To obtain a phylogenetic tree with branch length estimates, we first compiled a constraint tree topology from the literature. The arthropod order topology was taken from Misof et al. (2014). We computed multiple sequence alignments from the COI sequences for each order separately using MAFFT v7.309. We removed redundant sequences and sequences that could not be translated without having stop codons, and inferred ML phylogenies for each order using RAxML v8.2.11. We manually corrected these order-specific topologies using published phylogenies (reference list in Supplemental Table D.8). For those species without placement from the literature, we used the COI topology under majority-rule consensus in case there were more than one COI sequence. We combined the order-specific trees into a large tree and used that topology as constraint to estimate branch lengths using RAxML.

The resulting tree was rendered ultrametric by a short Python script using the ETE toolkit (Huerta-Cepas et al., 2016) (see Supplemental Material). To time-calibrate the phylogeny, we selected calibration points from (Misof et al., 2014) (Table D.7) and used the `chronos` function from the ape package in R to adjust the branch lengths. We used the upper and lower bounds

of the 95 % confidence interval as minimum and maximum node ages. We set λ to 2 and used the discrete model. We used the `fastAnc` ML implementation in the phytools package (Revell, 2012) in R to infer ancestral genome sizes using the Ornstein-Uhlenbeck model for each node along the tree including 95 % confidence intervals.

Genomic datasets and genome size estimates

The genome assembly accession numbers and data sources of 96 arthropod species are listed in Supplemental Table D.1. The genome assemblies were downloaded from NCBI (69 species) or from the i5k FTP server (27 species). Genome size estimates were obtained from the Animal Genome Size Database (Gregory (2018), `http://genomesize.com`), measured in our own lab using flow cytometry (FCM), or estimated using a *k*-mer peak method adopted from Hozza et al. (2015). Our own estimate results are listed in Supplemental Table D.2.

Transposable element annotation

We used a pipeline for repeat annotation from Petersen et al. (2019) that employs RepeatModeler 1.0.10 (Smit et al., 2015b) to infer a species-specific repeat consensus library from each genome assembly, and RepeatMasker 4.0.5 (Smit et al., 2015a) to annotate TE copies in the genome assemblies. The annotation by RepeatModeler includes a substantial fraction of "Unknown" elements, so the pipeline employs an intermediate filtering step to exclude false positives. We used NCBI BLAST to search the consensus libraries in the NCBI non-redundant nucleotide database, and removed all "Unknown" consensus sequences that did not result in a hit on a known TE protein. We also removed annotations shorter than 50 nucleotides.

To infer accurate TE content estimates and TE ages from the RepeatMasker annotation results, we developed a Perl program that uses the Kimura distance of each TE copy from the TE consensus sequence and the order-specific nucleotide substitution rate to infer the age of each TE copy in million years (My). The Perl code is available at this study's Github repository. We used a time-calibrated phylogeny of insects (Misof et al., 2014) and multiple sequence alignments of 1,478 protein-coding genes from 144 arthropod species (Misof et al., 2014) to infer order-specific nucleotide substitution rates by using the weighted arithmetic mean of substitution rates (see equation (D.1) in Supplemental material, page 377). The program also takes into account that TE annotations sometimes overlap each other by distributing the count of affected nucleotide positions as fractions evenly among overlapping TE copies. While this approach results in decimal instead of integer TE lengths, it provides a better estimate of the amount of nucleotides covered by TEs as it handles each element equally. The corrected lengths were only used in the TE content counts, not in the age estimations.

DNA gain and loss

Using the time-calibrated phylogeny and the TE age inferences, we classified TE copies into clade-specific and ancient (a TE copy was classified as ancient if it was older than the most recent split of the clade it was found in to the sister clade, otherwise as lineage-specific). For Chelicerata and Myriapoda, we took clade divergence times from Misof et al. (2014) (all divergence times are listed in Table D.4). We calculated the amount of DNA gained by TE proliferation as the amount of clade-specific TEs.

With the ancestral genome sizes and the inferred amounts of DNA gained by TE activity, we computed the amount of ancestral DNA in the extant genomes by subtracting the amounts of lineage-specific TE material from the extant genome assembly sizes. This allowed us to infer the amount of ancestral DNA that was lost since the common ancestor of the clade for each arthropod lineage by subtracting the amount of ancestral DNA from the estimated ancestral genome size.

We computed the DNA loss coefficient k (1) according to Lindblad-Toh et al. (2005) as $E = \mathcal{A}e^{-kt}$, where E is the amount of extant ancestral sequence in the genome assembly, $\mathcal{A}$ the total ancestral assembly size, and t the time since the split from the last ancestor.

$$k = \frac{\ln \frac{\mathcal{A}}{E}}{t} \tag{3.1}$$

3.3 Results

Ancestral genome size reconstruction

We reconstructed ancestral genome sizes (see Methods) of 613 arthropod species with published phylogenetic relationships (refs. listed in Supplemental Table D.8), amended with branch lengths inferred from COI barcode sequences, and genome size estimates for extant species obtained from the genome size database (Gregory, 2018). We inferred an ancestral genome size for the root node of hexapods (node 1) between 782 and 1943 Mbp (95 % confidence interval) (Figure 3.1). This inferred genome size is well above the maximum of many holometabolous clades such as Diptera (node 2, 272 to 545 Mbp), Lepidoptera (node 3, 318 to 738 Mbp), or Hy-

menoptera (node 4, 303 to 633 Mbp) (Table 3.1), but within the range of hemimetabolan orders, except for Orthoptera (node 5), of which the ancestral size was inferred to be between 3,677 and 9,473 Mbp. This is not surprising given the genome sizes of extant representatives of Orthoptera between 2 Gbp in *Acheta domesticus* and 16.5 Gbp in *Podisma pedestris*.

Table 3.1: Inferred ancestral genome sizes for major arthropod orders using the Ornstein-Uhlenbeck model. Ancestral size refers to the median, upper and lower bounds refer to the bounds of the 95% confidence interval (CI). All values in Mbp.

Clade	Node	Anc. size [Mbp]	Lower bound	Upper bound
Diptera	2	340.65	251.41	461.57
Diptera:Telmatogeton+Chironomus	17	242.47	165.97	354.25
Diptera:Drosophila	18	268.96	201.7	358.65
Diptera:Aedes	21	850.17	590.84	1223.32
Mecoptera		411.94	281.49	602.85
Lepidoptera	3	478.6	333.79	686.23
Lepidoptera:Papilionidae	6	368.1	236.45	573.04
Lepidoptera:Drepanidae	7	379.0	246.92	581.73
Lepidoptera:Geometridae	8	591.34	400.93	872.18
Lepidoptera:Notodontidae	9	427.97	279.38	655.57
Lepidoptera:Erebidae	10	742.04	519.66	1059.59
Lepidoptera:Euchaetes+Lymantria	20	828.79	599.19	1146.37
Trichoptera		527.27	339.93	817.85
Neuropterida		512.92	273.37	962.38

Table 3.1 –continued

Clade	Node	Anc. size	Lower bound	Upper bound
Coleoptera		565.59	385.45	829.92
Coleoptera:Callosobruchus	12	1037.25	694.8	1548.49
Coleoptera:Carabidae	13	321.01	207.3	497.11
Coleoptera:Tribolium	16	303.52	198.64	463.78
Coleoptera:Tenebrionidae	11	467.36	303.29	720.18
Coleoptera:Dermestidae	19	903.72	559.72	1459.12
Strepsiptera	14	192.46	105.68	350.49
Hymenoptera	4	418.05	280.94	622.08
Hymenoptera:base	15	242.59	139.94	420.54
Condylognatha		741.39	471.34	1166.16
Psocodea		705.64	472.19	1054.53
Hemiptera		778.22	496.61	1219.52
Sternorrhyncha		676.04	416.06	1098.48
Heteroptera		1056.97	642.94	1737.61
Auchenorryncha		1404.28	785.17	2511.56
Thysanoptera		514.07	268.12	985.63
Polyneoptera		1623.44	996.76	2644.11
Blattodea+Isoptera		1685.6	969.63	2930.26
Isoptera		1391.74	843.35	2296.74

Table 3.1 –continued

Clade	Node	Anc. size	Lower bound	Upper bound
Phasmatodea		1770.0	960.52	3261.7
Orthoptera	5	2659.61	1580.63	4475.16
Odonata		917.39	562.65	1495.78
Ephemeroptera		887.61	538.95	1461.83
Palaeoptera		887.61	538.95	1461.83
Diplura		990.59	504.5	1945.06
Archaeognatha		1630.8	828.79	3208.93
Ellipura		990.59	504.5	1945.06
Collembola		990.59	504.5	1945.06
Zygentoma		1002.05	629.0	1596.35
Hexapoda	1	990.59	504.5	1945.06
Crustacea		2043.45	1122.58	3719.73
Copepoda+Branchiopoda		1364.13	773.59	2405.49
Branchiopoda		973.39	524.83	1805.31
Copepoda		1256.67	714.34	2210.73
Malacostraca		3186.87	1774.25	5724.16

In some orders, we inferred a dynamic pattern of genome size evolution. For example within Lepidoptera, Papilionidae (node 6) have the smallest inferred ancestral genome size among Lepidoptera (median size: 358 Mbp), or the two sister groups Drepanidae and Geometridae (nodes

7 and 8) differ in their inferred ancestral genome size more than twofold (*Drepana* species (Drepanidae), median size: 318 Mbp; *Scopula limboundata* (Geometridae), median size: 870 Mbp), as is also the case for the two sister groups Notodontidae (node 9, *e.g.*, *Nadata gibbosa*; median size: 352 Mbp) and Erebidae (node 10, *e.g.*, *Malacosoma disstria*; median size: 636 Mbp). A similar dynamics in inferred ancestral genome size variation is visible in Coleoptera: Tenebrionidae (node 11, such as species of the genus *Tribolium*) have small inferred ancestral genomes (median size: 235 Mbp) in contrast to species of the Cleridae/Silvanidae/Chrysomelidae/Curculionidae complex with large inferred ancestral genomes (*e.g.*, *Callosobruchus*, node 12; median: 836 Mbp). Within Carabidae (node 13), species of the genus *Carabus* have inferred ancestral genome sizes of about 245 Mbp, in stark contrast to the other carabid species such as *Calosoma scrutator* (1,017 Mbp).

The smallest extant and ancestral inferred genome size was found in Strepsiptera (node 14) (inferred genome size of the most recent common ancestor (MRCA): 104 to 349 Mbp, with the extant *Xenos vesparum* having 127 Mbp). The ancestral holometabolan genome was inferred to be between 390 and 751 Mbp in size. Holometabola, however, appear to have undergone several events of genome size contraction according to our reconstructions (smaller than the holometabolan ancestor). Examples of smaller genome sizes than the holometabolan ancestor include the MRCA of basal Hymenoptera such as *Macrocentrus cingulum*, *Aphidius colemani*, and *Aphidius ervi* (node 15) with an inferred ancestral genome size between 140 and 420 Mbp; similarly, we inferred a genome size of 199 to 464 Mbp for the MRCA of the *Tribolium* beetle genus (node 16). Likewise, we inferred smaller genomes for the MRCA of the nematoceran flies *Telmatogeton japonicus* and *Chironomus plumosus* (Diptera, node 17, 166 to 354 Mbp) and of the

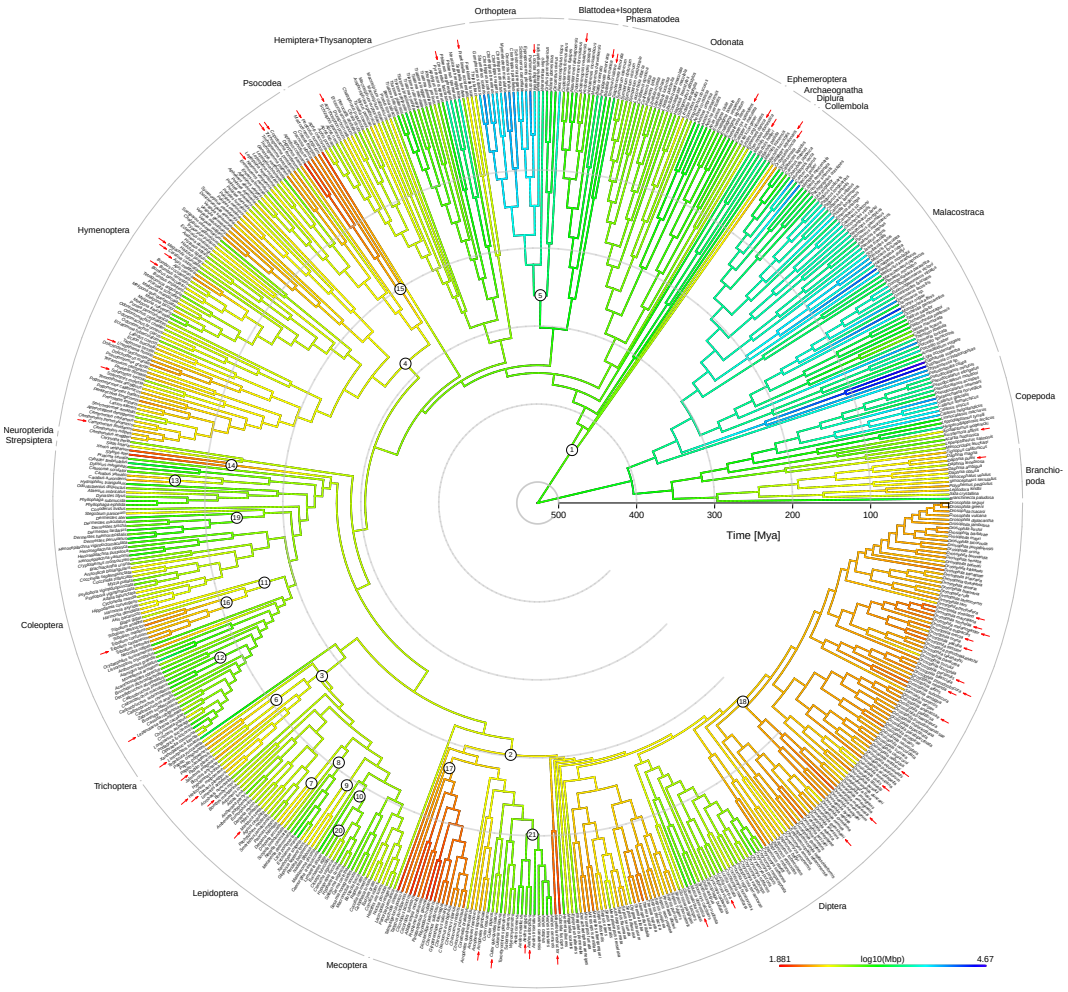

Figure 3.1: Ancestral genome size reconstruction reveals highly dynamic insect genomes. Chronogram based on published phylogenies and branch lengths estimated from COI nucleotide sequences. The branch coloration represents the inferred ancestral genome size (red: small, green: medium, blue: large). Red arrows denote species that are included in the TE age analysis. This figure is also appended in A2 poster format to the end of this thesis.

Drosophila group (node 18, 202 to 359 Mbp). Our inferences also include examples of genome expansion (larger than the holometabolan ancestor). For example, the MRCA of dermestid beetles (node 19) had an inferred genome size between 560 and 1,459 Mbp); the MRCA genome of *Lymantria dispar* and *Euchaetes egle* (Lepidoptera, node 20) was inferred to be between 520 and

1,060 Mbp, and the MRCA genome of the *Aedes* mosquito genus (node 21) was inferred to be between 591 and 1,224 Mbp. These results also contradict prior suggestions that Holometabola generally have smaller genomes than Hemimetabola, *e.g.*, by Hanrahan & Johnston (2011).

Transposable elements contribute to genome size variation

We investigated the correlation between genome size and TE content. In order to do so, we annotated TEs in 96 arthropod genomes using a pipeline that combines RepeatModeler (Smit et al., 2015b) and RepeatMasker (Smit et al., 2015a) and found that genome size correlates with TE content (PIC (Felsenstein, 1985), Pearson's product moment correlation, $p = 0.0003484$). We also inferred the age in million years for each TE copy using order-specific nucleotide substitution rates and the Kimura 2-parameter distance of each TE copy from the TE family consensus reported by RepeatMasker (see Methods). The median ages (Figure 3.2, page 105) of all TE classes within species are significantly correlated with the divergence times of the respective species, but only LINEs show this correlation also when applying PIC (Kendall's rank correlation, $p = 0.04$).

Using divergence times from the dated phylogeny based on the literature (see above), we classified the TE content into lineage-specific copies (younger than the age of the lineage, *i.e.*, the split of the species with its last common ancestor) and ancient copies (older than the lineage). In 36 out of 53 insect species that were represented in the dated tree, we found more than 99 % lineage-specific TEs (Figure 3.3 on page 106, Supplemental Figure D.1). Notable exceptions included the termite *Zootermopsis* with 86.3 % ancestral TEs, the bumblebee *Bombus terrestris* with 81.1 % ancestral TEs, and the dragonfly *Calopteryx splendens* with 82 % ancestral TEs. The

closely related *Drosophila* species displayed great variation in the ancestral TE fraction, ranging from 0.85 % in *D. mojavensis* to 47.4 % in *D. simulans*. We tested for phylogenetic signal using Blomberg's K and found low signal in the ancestral TE fraction ($K = 0.1, p = 0.9$). The fraction of ancestral TEs is significantly correlated with the age of the lineage under phylogenetically independent contrasts (PIC) (Kendall, $p = 2e - 10$).

We inferred the total amount of DNA gained and lost in each lineage by first calculating the fraction of lineage-specific TE derived DNA, *i.e.*, the amount of DNA that was gained by TE activity since the split from its sister species. We subtracted the amount of lineage-specific TEs (DNA gained since the split of the sister-species present in our tree) from the assembly size of each species. To compute the amount of DNA lost, we subtracted the amount of ancestral DNA from the inferred ancestral genome size of each species. This analysis revealed highly dynamic genome sizes among species and clades (Figure 3.4 on page 107, Table D.5). In 75 out of 89 species (we omitted the chelicerates and myriapods which were not represented in the dated phylogeny), the amount of DNA loss exceeds the amount of DNA gained through the accumulation of TEs. These 75 species include five dipterans, in particular two representatives of *Aedes* mosquitoes, but no representatives of *Drosophila* or other closely related species. The ratio of gain to loss ranged between 0.2 in the fly *Rhagoletis zephyria* to 5.1 in the butterfly *Calycopis cecrops*.

We inferred the largest absolute values of DNA gain (3.7 Gbp) and loss (7.1 Gbp) in the locust *Locusta migratoria* with 5.8 Gbp, the largest studied genome. It is followed by the amphipod *Hyalella azteca*, which was inferred to have lost 5.4 Gbp, but gained only 136 Mbp. In general, crustaceans appear to have lost large absolute amounts of DNA, however the average

ratio of DNA gain to DNA loss (0.11) is estimated to be lower compared to hexapods (0.88). The ratio of DNA gain to DNA loss was not significantly different in holometabolous and hemimetabolous insects (phylogenetic ANOVA, $p = 0.5$).

With the inferred DNA sequence gains and losses, we calculated the DNA loss coefficient according to Kapusta et al. (2017). The DNA loss coefficient, k, is calculated from the amount of DNA gained and lost since the last ancestor (difference between the extant genome size and the ancestral size in terms of lineage-specific DNA; see Methods). We assume that the DNA loss coefficient is constant over time and describes the loss of DNA sequence over time within a genome of a particular species. It has to be kept in mind that DNA loss is counterbalanced by DNA gain due to TE propagation within a genome. We found an extremely high DNA loss coefficient in the strepsipteran *Mengenilla moldrzyki* with a small genome (156 Mbp, 48.5 % TEs; $k = 0.0173$). We found the lowest DNA loss coefficients in the two mosquitoes *A. aegypti* (1,871 Mbp, 61.2 % TEs, $k \approx 0$) and *A. albopictus* (2,247 Mbp, 55.6 % TEs, $k \approx 0$), both of which have large genomes and a high TE content.

Interestingly, genome assembly size and DNA loss coefficient are negatively correlated (Kendall, PIC, $p = 0.001$) in contrast to a weak positive correlation between TE content and DNA loss coefficient (Pearson, PIC, $p = 0.03$). However, using PIC there is no correlation at all (Supplemental Figure D.2), neither among all species nor when subsampling the dataset to Holometabola/Hemimetabola or by taxonomic order. Instead, genome size appears to remain more or less constant (albeit with a large spread) despite changing coefficients of DNA loss. This is in stark contrast to the findings by Kapusta et al. (2017) who also reported a negative correlation between DNA loss coefficient and genome size in birds and mammals, but a significant

positive correlation supported by PIC between TE content and DNA loss coefficient (Supplemental Figure D.3).

3.4 Discussion

We present the most comprehensive analysis of genome size dynamics in arthropods, focusing on gain and loss of TE-derived DNA. In arthropods, and particularly in hexapods, genome size variation, which reaches an amplitude of up to 1,600 % (Figure 3.1), which substantially exceeds variation in mammals and birds (Kapusta et al., 2017). Kapusta et al. (2017) proposed an explanation for the relatively invariant genome sizes within mammals and birds, which can be observed despite the active propagation of TEs in these genomes, namely an "accordion" model of genome size evolution. The "accordion" model of genome size dynamics assumes that DNA gain, for example through massive lineage-specific TE propagation, is counteracted by DNA loss, for example, via ectopic recombination and other mechanisms and subsequent removal by repair mechanisms. This process is supposed to maintain a genome size equilibrium. Mechanistically, the "accordion" model proposes that TE insertions lead to DNA gain, but also generate targets for ectopic recombination which can induce DNA loss. Kapusta et al. (2017) further show that there is empirical evidence in mammalian and bird genomes of frequent macrodeletions compatible with the action of ectopic recombination. Given the proposed mechanistic explanation of the "accordion" model, it should also apply to arthropod genomes. In fact, we inferred a similar balance of DNA loss and gain within the major insect orders: Large genome sizes are correlated with high TE content (Figure D.3) and high amounts of DNA loss (Figure D.2), but the "accordion" model does not explain the large periodic shifts in genome size be-

tween the major insect orders. Instead, insect genomes appear to cope with TE influx in an entirely different manner than vertebrate genomes. Where in mammals, a high rate of DNA loss leads to a smaller extant genome size, in insects the genome size remains more or less constant (within the large range of dispersion) according to our DNA loss coefficient inferences (Figure D.2). These results suggest that in insect genomes, even a high rate of DNA loss is barely able to cope with the high rate of DNA influx due to TE activity and and potentially transfection keep the genome size stable – we did not observe a stable trend towards genome shrinkage in insects. However, the ancestral genome size reconstruction suggests that there have been periods of genome contraction during the evolution of arthropods which are not modeled using a constant coefficient of DNA loss. To better infer the pattern of DNA loss over the ~450 million years of insect evolution would require a variable DNA loss coefficient and a model that can take into account ancestral genome sizes and DNA gain/loss states at multiple points in the phylogeny.

Genome size reduction in vertebrates has been implicated in the metabolic requirements of powered flight (Wright et al., 2014); this is indicated by the fact that birds with higher metabolic rates, such as hummingbirds, have smaller genomes than flightless birds (Gregory, 2005). In insects, we would expect a similar rate of DNA removal over time if powered flight should play a role. However, we observe a different situation: in flightless arthropods, genome size shows a trend to increase with the DNA loss coefficient, while in insects capable of flight, the trend is downwards (Figure D.4). Hence, the metabolic rate is likely not a predictor of genome size in insects, regardless of flight capability.

Ancient TEs become unrecognizable

We found almost no ancestral TEs in species that diverged earlier than 100 Mya from their sister species (Figure 3.3). This is most likely a consequence of the TE nucleotide sequence similarity decaying over time and thus sequence homology becoming undetectable. Its effect is easily visualized when plotting the TE content distribution over the sequence divergence (or age, if conversion is available) and dividing the landscape in two parts, separated at the age of the species (Figure D.5). These findings are in line with other studies suggesting that inactive TEs become unrecognizable beyond 50 Mya due to high sequence divergence (*e.g.*, SINES (Shedlock & Okada, 2000)).

Genome contraction covaries with TE expansion

Insects have much larger effective population sizes than mammals or birds, which limits the effects of genetic drift and exacerbates the efficiency of natural and purifying selection (Szitenberg et al., 2016). As a result, we would expect TE activity to both be of limited detrimental effect to the host organism, and lead to widely distributed copies of active TEs among the individuals of a population. The latter can happen within a few generations, as has been shown in *Drosophila* fruit flies (Kofler et al., 2015); our analysis suggest a similar rate of intra-population TE proliferation in other insect species, however this remains to be tested experimentally.

TE activity has been shown to be a pivotal agent shaping genome size evolution in insects (Maumus et al., 2015), with DNA loss barely counteracting DNA gained by TE transposition to maintain a genome size equilibrium. For example, the large genome of the migratory locust *Lo-*

custa migratoria, which consists of over 60 % TEs, exhibits a moderate rate of DNA loss (DNA loss coefficient of $k = 0.003$), which did not prevent it from being inflated over time due to TE proliferation. On the other hand, there are examples to the contrary, documenting that a high rate of DNA loss can lead to small genomes despite high TE content; this is the case in *Mengenilla moldrzyki*. In these species, it appears that DNA loss is more efficient at keeping overly high TE activity in check. However, these traits appear lineage-specific and cannot be generalized to other representatives of the same orders.

Limitations of the methods in insect genomes

This analysis is of course heavily influenced by the node dating of the underlying phylogeny, and our approach using COI barcode sequences cannot rival the accuracy of phylogenomic studies (*e.g.*, Misof et al. (2014)). However, using calibration points from Misof et al. (2014) enabled us to estimate node ages with reasonable accuracy and therefore provide a robust dated phylogeny for the TE age classification. Unfortunately, for some species there were no closely related species in the dataset, which forced us to select an ancestral split that is older than the species would be. This was the case for all orders with only a single representative (Collembola, Diplura, Psocodea, Trichoptera, and Mecoptera). Here, the representative species were assigned an age that is even older than the age of the sister order, which likely led to an underestimation of the ancestral TE content. To solve this issue, genome size estimates for more representatives of these orders are required. This also highlights the importance of efforts like the genome size database (Gregory, 2018) in the age of whole-genome sequencing – not only because the estimates aid in establishing sequencing strategies, but also for comparative analyses like this one.

Kapusta et al. (2017) obtained a dataset that included a multiple whole genome alignment of 100 vertebrate species. Using this whole genome alignment, they were able to infer micro- and macrodeletions in the vertebrate lineage. These are lacking in our dataset simply because whole genome alignments are difficult in insects due to low conservation of synteny: while the human genome aligns with over 98 % to the chimpanzee genome and with around 70 % to the mouse genome (Mural et al., 2002) this is not the case in insects across larger evolutionary time scales. For example, the honey bee *A. mellifera* genome aligns to less than 20 % of the turnip sawfly *Athalia rosae* genome, also a representative of the order Hymenoptera (A. Donath, *pers. comm*). Thus, we omitted analysis of micro- and macrodeletions and segmental duplications in the insect genomes. However, since these events make up at most 10 % of the vertebrate DNA gain or loss (Kapusta et al., 2017), with the analysis on TEs we have covered the major source of DNA gain and loss in arthropod genomes. Our analysis is instead based on a wider dataset with twice as many species from all major insect and crustacean orders. This provided us with a broad comparative view on genome size dynamics in arthropods.

3.5 Conclusion

We show that genome size in insects is governed by complex dynamics that are not entirely explained by TE activity alone. There are large-scale differences even between (relatively) closely related taxa. We find that the "accordion" model that describes DNA gain and loss in birds and mammals (Kapusta et al., 2017) does not fit the DNA gain/loss dynamics in insect genomes. Instead, we find that DNA loss does not counteract TE proliferation: on average, the genome size remains more or less stable despite large amounts of DNA lost.

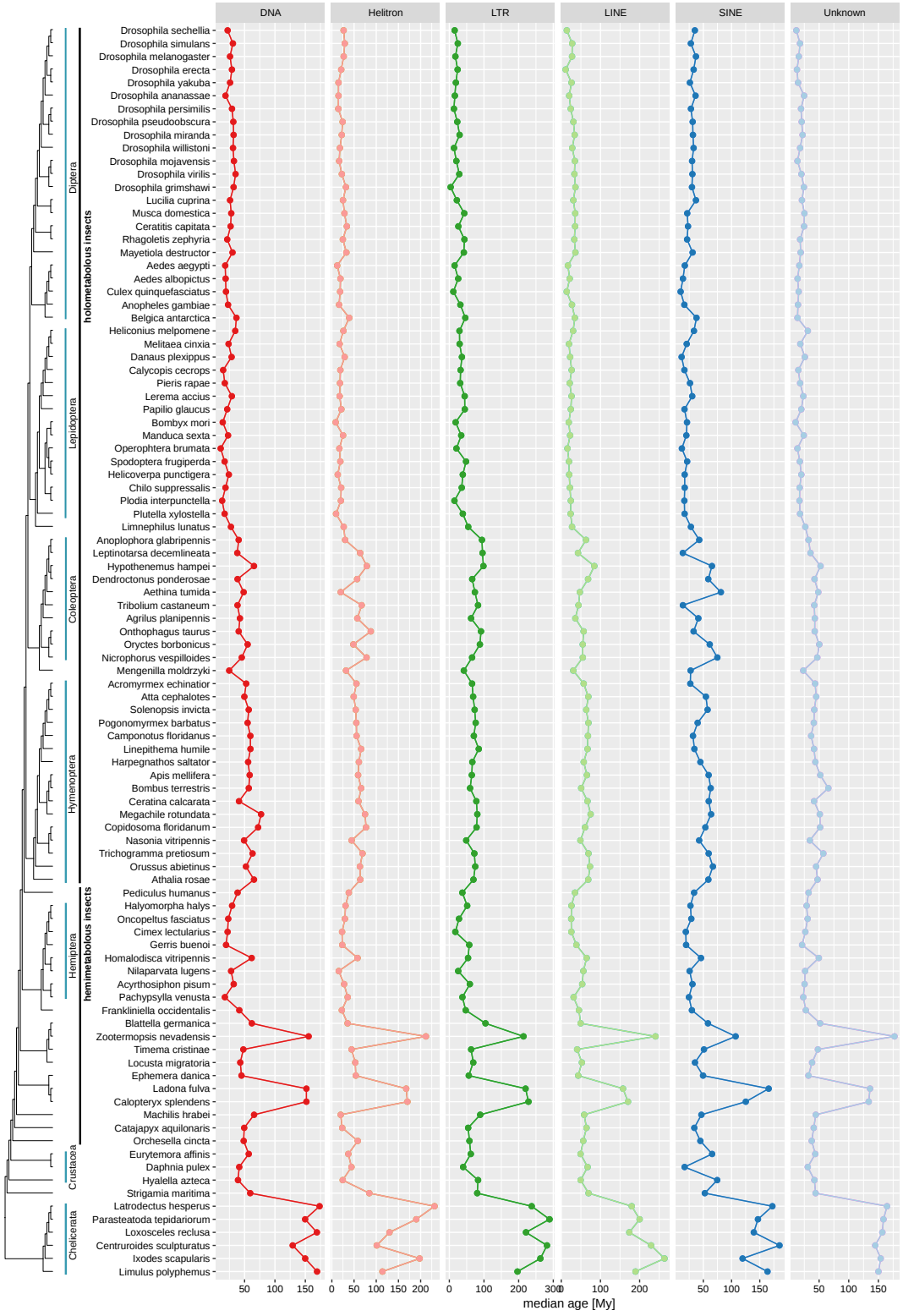

Figure 3.2: The median ages of six TE subclasses in all sampled species show variation between and within subclasses. Clade relationships after Misof et al. (2014). Species relationships within clades are based on the published phylogenies listed in Table D.8.

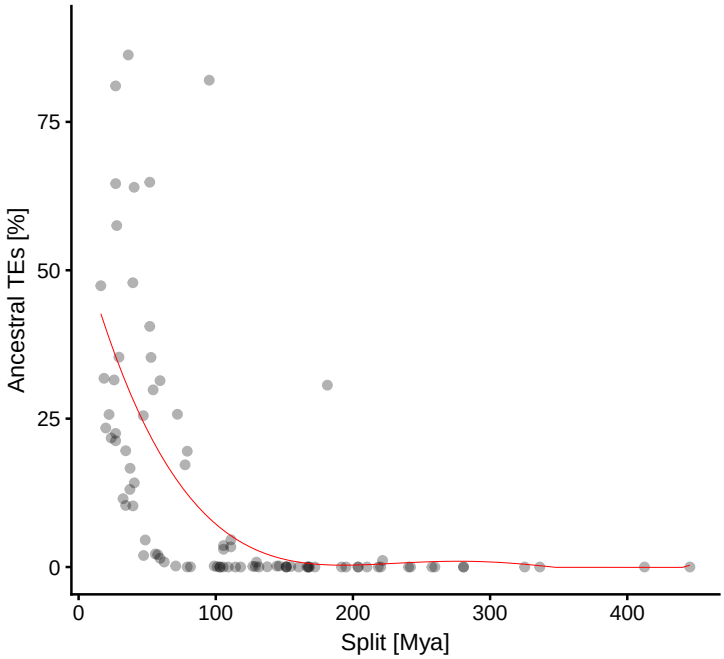

Figure 3.3: TEs are no longer recognized as “ancient” beyond a clade age of ~ 120 Mya. Dots: individual measurements; red line: polynomial regression. “Split” refers to the most recent common ancestor of the clade and its sister clade.

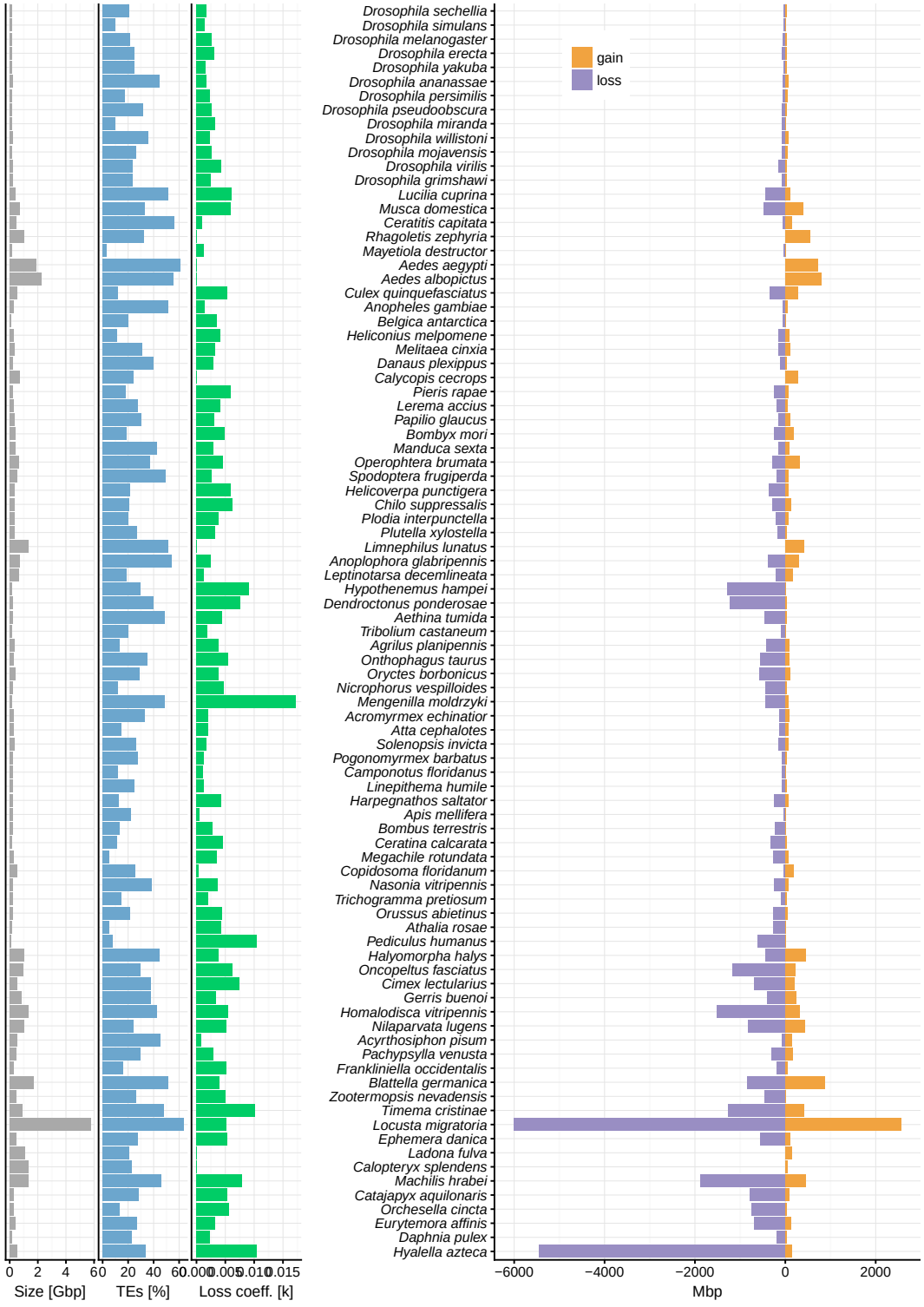

Figure 3.4: Increased DNA loss rate explains some of the observed genome size reductions in insects, but not all. The opposite is true, however: for species with negative loss coefficients we inferred increased genome sizes compared to other species in the same order, sometimes drastically so.

References

Alfsnes, K., Leinaas, H. P., & Hessen, D. O. (2017). Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecology and Evolution*, (pp. n/a–n/a).

Arkhipova, I. R. (2018). Neutral Theory, Transposable Elements, and Eukaryotic Genome Evolution. *Molecular Biology and Evolution*, 35(6), 1332–1337.

Bennett, M. D., Lewis, K. R., & Harberd, D. J. (1977). The Time and Duration of Meiosis [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 277(955), 201–226.

Bennetzen, J. L., Ma, J., & Devos, K. M. (2005). Mechanisms of Recent Genome Size Variation in Flowering Plants. *Annals of Botany*, 95(1), 127–132.

Blass, E., Bell, M., & Boissinot, S. (2012). Accumulation and Rapid Decay of Non-LTR Retrotransposons in the Genome of the Three-Spine Stickleback. *Genome Biology and Evolution*, 4(5), 687–702.

Charlesworth, B. & Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetical Research*, 42(01), 1.

Dufresne, F. & Jeffery, N. (2011). A guided tour of large genome size in animals: What we know and where we are heading. *Chromosome Research*, 19(7), 925–938.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1), 1–15.

Gregory, T. R., Ed. (2005). *The Evolution of the Genome*. Burlington, MA: Elsevier Academic. OCLC: ocm57727263.

Gregory, T. R. (2018). Animal Genome Size Database.

Gregory, T. R. & Johnston, J. S. (2008). Genome Size Diversity in the Family Drosophilidae. *Heredity*, 101(3), 228–238.

Hanrahan, S. J. & Johnston, J. S. (2011). New genome size estimates of 134 species of arthropods. *Chromosome Research*, 19(6), 809–823.

Hozza, M., Vinař, T., & Brejová, B. (2015). How Big is that Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra. In *String Processing and Information Retrieval*, Lecture Notes in Computer Science (pp. 199–209).: Springer, Cham.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6), 1635–1638.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, (pp. 201616702).

Kelly, L. J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., Lysak, M. A., Day, P. D., Berger, M., Fay, M. F., Nichols, R. A., Leitch, A. R., & Leitch, I. J. (2015). Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist*, 208(2), 596–607.

Kofler, R., Nolte, V., & Schlötterer, C. (2015). Tempo and Mode of Transposable Element Activity in Drosophila. *PLOS Genetics*, 11(7), e1005406.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., deJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Gra-

ham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.-P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., & Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069), 803–819.

Marburger, S., Alexandrou, M. A., Taggart, J. B., Creer, S., Carvalho, G., Oliveira, C., & Taylor, M. I. (2018). Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. *Proceedings of the Royal Society B: Biological Sciences*, 285(1872).

Maumus, F., Fiston-Lavier, A.-S., & Quesneville, H. (2015). Impact of Transposable Elements on Insect Genomes and Biology. *Current Opinion in Insect Science*, 7, 30–36.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von, R. B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K., & Zhou, X. (2014). Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science*, 346, 763–7.

Mural, R. J., Adams, M. D., Myers, E. W., Smith, H. O., Miklos, G. L. G., Wides, R., Halpern, A., Li, P. W., Sutton, G. G., Nadeau, J., Salzberg, S. L., Holt, R. A., Kodira, C. D., Lu, F., Chen, L., Deng, Z., Evangelista, C. C., Gan, W., Heiman, T. J., Li, J., Li, Z., Merkulov, G. V., Milshina, N. V., Naik, A. K., Qi, R., Shue, B. C., Wang, A., Wang, J., Wang, X., Yan, X., Ye, J., Yooseph, S., Zhao, Q., Zheng, L., Zhu, S. C., Biddick, K., Bolanos, R., Delcher, A. L., Dew,

I. M., Fasulo, D., Flanigan, M. J., Huson, D. H., Kravitz, S. A., Miller, J. R., Mobarry, C. M., Reinert, K., Remington, K. A., Zhang, Q., Zheng, X. H., Nusskern, D. R., Lai, Z., Lei, Y., Zhong, W., Yao, A., Guan, P., Ji, R. R., Gu, Z., Wang, Z. Y., Zhong, F., Xiao, C., Chiang, C. C., Yandell, M., Wortman, J. R., Amanatides, P. G., Hladun, S. L., Pratts, E. C., Johnson, J. E., Dodson, K. L., Woodford, K. J., Evans, C. A., Gropman, B., Rusch, D. B., Venter, E., Wang, M., Smith, T. J., Houck, J. T., Tompkins, D. E., Haynes, C., Jacob, D., Chin, S. H., Allen, D. R., Dahlke, C. E., Sanders, R., Li, K., Liu, X., Levitsky, A. A., Majoros, W. H., Chen, Q., Xia, A. C., Lopez, J. R., Donnelly, M. T., Newman, M. H., Glodek, A., Kraft, C. L., Nodell, M., Ali, F., An, H. J., Baldwin-Pitts, D., Beeson, K. Y., Cai, S., Carnes, M., Carver, A., Caulk, P. M., Center, A., Chen, Y. H., Cheng, M. L., Coyne, M. D., Crowder, M., Danaher, S., Davenport, L. B., Desilets, R., Dietz, S. M., Doup, L., Dullaghan, P., Ferriera, S., Fosler, C. R., Gire, H. C., Gluecksmann, A., Gocayne, J. D., Gray, J., Hart, B., Haynes, J., Hoover, J., Howland, T., Ibegwam, C., Jalali, M., Johns, D., Kline, L., Ma, D. S., MacCawley, S., Magoon, A., Mann, F., May, D., McIntosh, T. C., Mehta, S., Moy, L., Moy, M. C., Murphy, B. J., Murphy, S. D., Nelson, K. A., Nuri, Z., Parker, K. A., Prudhomme, A. C., Puri, V. N., Qureshi, H., Raley, J. C., Reardon, M. S., Regier, M. A., Rogers, Y. H. C., Romblad, D. L., Schutz, J., Scott, J. L., Scott, R., Sitter, C. D., Smallwood, M., Sprague, A. C., Stewart, E., Strong, R. V., Suh, E., Sylvester, K., Thomas, R., Tint, N. N., Tsonis, C., Wang, G., Wang, G., Williams, M. S., Williams, S. M., Windsor, S. M., Wolfe, K., Wu, M. M., Zaveri, J., Chaturvedi, K., Gabrielian, A. E., Ke, Z., Sun, J., Subramanian, G., & Venter, J. C. (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296(5573), 1661–1671.

Neafsey, D. E. & Palumbi, S. R. (2003). Genome Size Evolution in Pufferfish: A Comparative

Analysis of Diodontid and Tetraodontid Pufferfish Genomes. *Genome Research*, 13(5), 821–830.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, A., Rilakovic, N., Ritland, C., Rosselló, J. A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Lee Thompson, S., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P. K., Lundeberg, J., & Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451), 579–584.

Petersen, M., Armisén, D., Gibbs, R. A., Hering, L., Khila, A., Mayer, G., Richards, S., Niehuis, O., & Misof, B. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evolutionary Biology*, 19(1).

Petrov, D. A. (2001). Evolution of genome size: New approaches to an old problem. *Trends in Genetics*, 17(1), 23–28.

Petrov, D. A., Lozovskaya, E. R., & Hartl, D. L. (1996). High intrinsic rate of DNA loss in Drosophila. *Nature*, 384(6607), 346–349.

Petrov, D. A., Sangster, T. A., Johnston, J. S., Hartl, D. L., & Shaw, K. L. (2000). Evidence for DNA Loss as a Determinant of Genome Size. *Science*, 287(5455), 1060–1062.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A., & Panaud, O. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Research*, 16(10), 1262–1269.

Ratnasingham, S. & Hebert, P. D. N. (2007). Bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7(3), 355–364.

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things): Phytools: R package. *Methods in Ecology and Evolution*, 3(2), 217–223.

Sato, Y. & Nishida, M. (2010). Teleost fish with specific genome duplication as unique models of vertebrate evolution. *Environmental Biology of Fishes*, 88(2), 169–188.

Shedlock, A. M. & Okada, N. (2000). SINE insertions: Powerful tools for molecular systematics. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 22(2), 148–160.

Smit, A., Hubley, R., & Green, P. (2015a). RepeatMasker Open-4.0.

Smit, A., Hubley, R., & Green, P. (2015b). RepeatModeler Open-4.0.

Sotero-Caio, C. G., Platt, R. N., Suh, A., & Ray, D. A. (2017). Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, 9(1), 161–177.

Sun, C., Arriaza, L., R, J., & Mueller, R. L. (2012). Slow DNA Loss in the Gigantic Genomes of Salamanders. *Genome Biology and Evolution*, 4(12), 1340–1348.

Szitenberg, A., Cha, S., Opperman, C. H., Bird, D. M., Blaxter, M. L., & Lunt, D. H. (2016). Genetic drift, not life history or RNAi, determine long term evolution of transposable elements. *Genome Biology and Evolution*, (pp. evw208).

Vitte, C., Panaud, O., & Quesneville, H. (2007). LTR retrotransposons in rice (Oryza sativa, L.): Recent burst amplifications followed by rapid DNA loss. *BMC genomics*, 8, 218.

White, M. & McLaren, I. (2000). Copepod development rates in relation to genome size and 18S rDNA copy number. *Genome*, 43(5), 750–755.

Wright, N. A., Gregory, T. R., & Witt, C. C. (2014). Metabolic 'engines' of flight drive genome size reduction in birds. *Proceedings of the Royal Society B: Biological Sciences*, 281(1779), 20132780–20132780.

4

Orthograph: Mapping coding nucleotide sequences to clusters of orthologous genes

This chapter has been published in: Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., Niehuis, O. (2017). Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, 18, 111. doi:10.1186/s12859-017-1529-8

Author contributions to the original article:

MP, BM, and ON conceived the Orthograph algorithm. MP wrote the Orthograph software. ON, AV, AD, BM, and KM contributed with suggestions, code review, and helper scripts to the Orthograph package. MP, KM, AD, DD, SL, RSP, LP, XZ, and ON contributed to the apoid wasp transcriptomics. MP, BM, and ON wrote the manuscript.

4.1 Background

Inferring the evolution of gene families, the phylogeny of species, and tracing the biogeography of populations depend on reliable delineation of orthologous genes and paralogous copies of them. While delineation and identification of orthologous and paralogous genes has been firmly established for studying genomic data (reviewed by Kristensen et al. (2011) and benchmarked by Trachana et al. (2011)), few approaches are currently available for assessing transcripts in the same manner (proposed by, *e.g.*, Ebersberger et al. (2009) and Schreiber et al. (2009)). Each of these approaches exhibits, and suffers from, specific problems, potentially leading to erroneous species and gene tree inference (see below). We developed a novel software pipeline, called Orthograph, for convenient, fast, and reliable identification of orthologous (and paralogous) nucleotide or amino acid sequences, which resolves existing algorithmic and software-technical

issues. Orthograph builds on previously proposed graph-based clustering algorithms, but extends them without sacrificing accuracy or computational speed.

When comparing the gene repertoires of species, one of the first analytical steps is the delineation of orthologous genes (*orthologs*), *i.e.*, the identification of genes that originated from a single gene in the last common ancestor of the compared species. Each of the delineated orthologous groups (OGs) can also include species- or lineage-specific gene copies (*inparalogs*), that evolved by gene duplication after the evolutionary split of the ancestor into different species (Koonin, 2005). Finally, horizontal gene transfer can give rise to xenologous gene copies (*xenologs*) from a single ancestral gene (Koonin, 2005).

Two fundamentally different approaches to identify potential orthologs, paralogs, and xenologs have been established: tree-based and graph-based approaches. The benefit of graph-based approaches, which we will subsequently focus on, is their computational efficiency and scalability (for reviews and a comprehensive discussion of the benefits of the different approaches, see Dutilh et al. (2007) or Kristensen et al. (2011)). In general, graph-based approaches assessing gene orthology make use of the genome-wide best reciprocal hit (BRH) criterion. It rests on the assumption that orthologs in two genomes are more similar to each other than to any other gene in the compared genomes, since they are direct and exclusive descendants from a single ancestral gene (Altenhoff & Dessimoz, 2012).

Various graph-based approaches based on the BRH criterion have been developed that *de novo* infer orthology among genes and proteins in the gene or protein sets of sequenced and annotated organisms, such as OrthoMCL (Li et al., 2003), COCO-CL (Jothi et al., 2006), OrthoDB (Kriventseva et al., 2015), InParanoid (Sonnhammer & Östlund, 2015), OrthoFinder

(Emms & Kelly, 2015), and OMA (Altenhoff et al., 2015). The reliability of these methods critically depend on the fact that differential gene loss is the exception and that gene or protein repertoires are complete. This means that in order to apply a graph-based approach to infer gene orthology among genomes, the organisms' gene or protein repertoire must be reliably known. These methods are therefore not appropriate for assessing orthology among nucleotide sequences in sequenced transcriptomes, since transcript libraries contain only a subset of the organisms' actual gene repertoire. The nucleotide sequence of a gene may be missing in a given transcript library simply because the gene was not (sufficiently highly) expressed at the time of RNA preservation. Given that transcriptome sequencing represents an extremely valuable and cost-efficient strategy to sample coding nucleotide sequences of a large fraction of an organism's gene repertoire (Misof et al., 2014), several graph-based approaches have been developed that are dedicated to ortholog identification in transcript libraries.

A possible solution to the aforementioned problem in transcript orthology assessment is to assign transcripts to OGs whose genealogical relationships have already been reliably inferred, rather than to infer orthology of these genes *de novo* from the transcripts. Knowledge of the genealogical relationships of genes can be derived from comparative genomic analyses and may be retrievable from public databases such as OrthoDB (Kriventseva et al., 2015). This approach has been implemented in OrthoSelect (Schreiber et al., 2009) and HaMStR (Ebersberger et al., 2009). However, OrthoSelect does not implement the BRH criterion, but a unidirectional search. OrthoSelect is thus prone to false positives. HaMStR, on the other hand is more sophisticated since it applies a BRH orthology prediction strategy. Specifically, HaMStR uses profile hidden Markov models (pHMMs) that represent properties of the aligned amino acid

sequences of each known OG to search a transcript library on the amino acid level for matches. All retrieved hits are then searched against the entire set of proteins, *i.e.*, the proteome (also referred to as "official gene set") as reference gene set (RGS), of each of the species of which amino acid sequences were used to construct the pHMM. If this reciprocal search retrieves the same amino acid sequence(s) that was (were) used in the construction of the pHMM, a the respective transcript is mapped to the OG in question.

The algorithm of HaMStR is "memoryless", meaning that during evaluation of the BRH criterion for a given OG, it does not consider which transcripts have been assigned to other OGs. Since transcripts are assigned to OGs on a per-OG basis without considering results from evaluations for other OGs and keeping track of what transcripts have already been assigned, it is possible that a given transcript is mapped to more than one gene. This issue of redundant transcript assignments can result in a misled inference of phylogenetic relationships, as has been shown (Struck et al., 2011; Kvist & Siddall, 2013), and can potentially compromise downstream analyses. In HaMStR, it would be conceivable to prevent redundant transcript assignment by implementing a record of previously assigned transcripts. However, such a first-come-first-serve approach cannot be justified: transcripts must be assigned to the OG that they are most likely orthologous to, not to the OG that came first in the search order. Since this serious issue cannot be solved using the HaMStR algorithm, we developed Orthograph: a different algorithm that circumvents redundant transcript assignments and instead maps transcripts to the globally best matching OG.

To assess the sensitivity and accuracy of Orthograph, we tested whether or not Orthograph a) reliably identifies orthologs, b) detects known paralogs, and c) finds known isoforms or al-

ternative transcripts. We additionally searched 24 *de novo*-sequenced transcript libraries of apoid wasps for 5,561 orthologous genes to assess the computational performance of Orthograph. Finally, we verified that Orthograph does not map transcripts to more than one gene by re-analyzing a dataset that has been processed with HaMStR. Our results demonstrate that Orthograph's performance is on par with HaMStR's while not suffering from redundant transcript assignment. Further, we emphasize the flexibility of Orthograph and highlight features that are likely of particular interest for a wide array of analyses in molecular evolutionary biology and in comparative genomics in particular.

4.2 Implementation

The Orthograph software package is divided into three main tools that handle (i) database management (manager), (ii) forward and reverse searches (analyzer), and (iii) clustering of orthologous transcripts and output (reporter). The separation into three distinct tools is a deliberate design choice to address work environments where users do not have full administrative privileges. This facilitates implementation in a high-performance computing cluster setup where the administrator can use the appropriate tool to manage the database, while users only need to run the actual analysis tools. In addition, this design allows the user to evaluate the alignment search results using different settings (*e.g.*, different alignment bit score thresholds to fine-tune and optimize parameters) quickly without re-running the computationally expensive searches.

Orthograph builds on the transcript orthology assessment strategy via BRH suggested by Ebersberger et al. (2009). In contrast to the implementation of this strategy in HaMStR, Orthograph assigns a given transcript to the *globally* best matching OGs while making sure that

no transcript is assigned more than once. It additionally identifies all transcripts (splice variants and inparalogs) present in an assembled transcript library that are putatively homologous to a given OG. The specific transcript orthology assignment algorithm is as follows (Figure 4.1 on page 127); note that steps 1 through 3 are only required once since their output can be used for all subsequent analyses:

1. The proteomes ("reference gene sets", RGS) of reference species are used as input.
2. Orthologous genes from all reference proteomes are clustered to form orthologous groups (OGs). This information is provided from public databases or one's own orthology delineation in the RGS.
3. For each OG, the amino acid sequences are aligned and the multiple sequence alignment (MSA) is used to construct a profile HMM.
4. These pHMMs are used to search the transcript sequences on the amino acid level for candidate homologs.
5. Search results are stored in a relational database.
6. For each pHMM search hit, the target amino acid sequence section matching the pHMM is used as a query to search in a database that includes all genes from the RGS (including the genes that form OGs) on the amino acid level.
7. The results of the reverse search are also stored in the relational database.
8. After all forward and reverse searches have completed, the clustering of BRH pairs takes place: search results from all forward searches are sorted by descending alignment bit score. For each forward alignment search result, the corresponding reverse alignment search results are sorted by descending alignment bit score as well. They are evaluated in order of descending alignment bit score for the forward search results, starting with the highest alignment bit score.
9. If the best reverse search hit of a given transcript is part of the OG that the pHMM for the forward search is based on (*i.e.*, the BRH criterion is fulfilled), the target transcript is assigned to the OG. The target transcript section is marked so that it cannot be assigned again. Each entry in the database is evaluated in this manner.

Orthograph performs several post-processing steps on transcripts assigned to OGs. By aligning the transcript fulfilling the BRH criterion to the most similar orthologous amino acid sequence of a reference species using Exonerate (Slater & Birney, 2005), it infers a frameshift-corrected open reading frame (ORF). Orthograph allows to extend the ORF beyond the pHMM alignment sequence section for which the BRH criterion was fulfilled while making sure that the orthologous region is covered by a user-defined percentage of the ORF length. Subsequently, it provides both the amino acid sequence and the exactly corresponding frameshift-corrected nucleotide sequence of a given transcript. Additionally, Orthograph can concatenate transcripts of a given OG to simplify downstream analyses (*e.g.*, phylogenomic investigations). In all above analysis steps, the user can fine-tune all relevant search and evaluation parameters using configuration files for clarity, documentation, and reproducibility.

Orthograph has been developed with user friendliness in mind. As a result, it is easy to install and runs on any Unix/Linux system (including OS X) that provides its dependencies (see Materials and Methods). The generation of custom-tailored ortholog sets, *e.g.*, from public databases is facilitated by its ability to parse simple tab-delimited tables. Input from public databases such as OrthoDB is easily formatted accordingly using standard UNIX or spreadsheet tools. In addition, the Orthograph package contains helper scripts that simplify the preparation of RGS sequence files for custom-made ortholog sets as well as summarize results for multiple analyses, *e.g.*, different species or using different settings.

When designing a custom ortholog set, users should pay close attention to the taxon sampling. Genes that occur in at least two species in each OG are recommended so that the resulting pHMMs are more informative than when based on single sequences only. In terms of OG

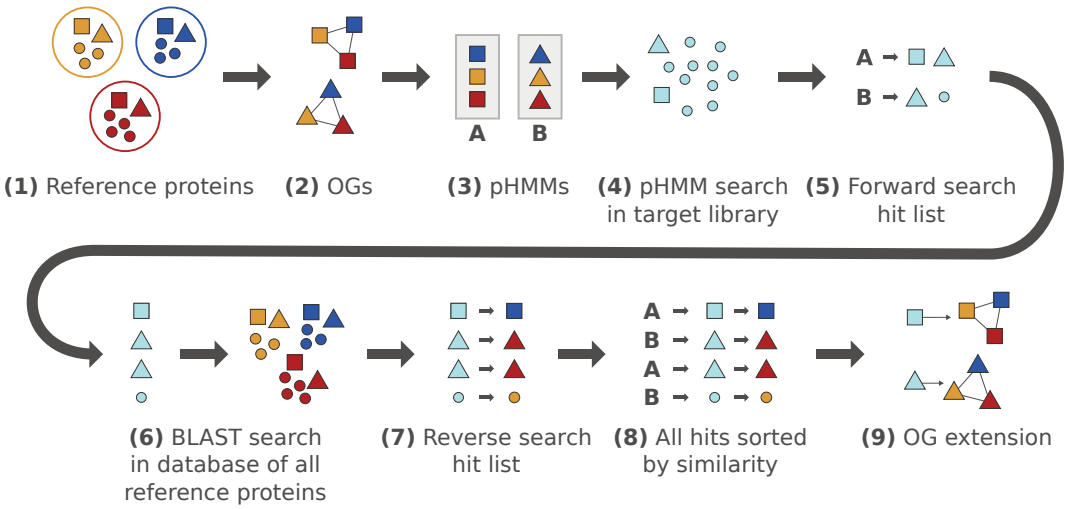

Figure 4.1: Orthograph workflow. From a set of reference proteins (1), the proteins are clustered to form orthologous groups (OGs) (2). These OGs are aligned to construct profile hidden Markov models (pHMMs) (3). The pHMMs are used to search for candidate orthologs in the target library (4). Each of the obtained hit amino acid sequences (5) is used as a query for a BLAST search in a database comprising all reference proteins (including the ones forming OGs) (6). Search results from both forward and reverse searches (7) are collated and sorted by bit score, with the reverse search result order being subordinated to the forward result order (8). This list is evaluated in descending order: if the reverse search hit a protein that is part of the OG used for the forward search, the candidate ortholog is mapped to the OG (9).

number, there is no lower or upper bound since the selection depends on the research question. Orthograph runtime increases linearly with each additional OG.

Detailed methods, data sources as well as system requirements are listed in the Supplemental Material (Figures E.1-E.5, Tables E.1-E.4).

4.3 Results and discussion

4.3.1 Sensitivity and accuracy when searching for single-copy orthologs

To assess the sensitivity and accuracy of Orthograph, we employed it to identify genes of known orthology in the RGS of the honeybee, *Apis mellifera* (15,314 genes, Honeybee Genome Se-

quencing Consortium (2006)), and Jerdon's jumping ant, *Harpegnathos saltator* (18,564 genes, Bonasio et al. (2010)). Specifically, we searched the RGS for 4,625 protein-coding genes provided by OrthoDB 5 (Waterhouse et al., 2011) as being single-copy across four species of Hymenoptera (*Apis mellifera* (Honeybee Genome Sequencing Consortium, 2006), *Camponotus floridanus* (Bonasio et al., 2010), *Harpegnathos saltator* (Bonasio et al., 2010), *Nasonia vitripennis* (Werren et al., 2010)) and the outgroup beetle *Tribolium castaneum* (Tribolium Genome Sequencing Consortium, 2008) (download URLs are listed in the Supplemental Material, Table E.3). Note that we removed all entries of the respective taxon whose RGS we analyzed for assessing the sensitivity and accuracy of Orthograph from this ortholog set (resulting in two sets: one without entries from *A. mellifera*, and one without entries from *H. saltator*). Of the 4,625 protein-coding genes that we searched for, Orthograph identified 4,582 (99.07 %) in the RGS of *A. mellifera* and 4,590 (99.24 %) in the RGS of *H. saltator* (Table 4.1 on page 138). In the case of *A. mellifera*, five proteins were assigned to other OGs than they were assigned by OrthoDB. We found a similar result for three proteins of the RGS of *H. saltator*. Visual inspection of these proteins suggested that the orthology assignment of these proteins in the OrthoDB database is not correct (for an in-depth assessment and discussion of an example see Supplemental Material, Figure E.5). The low fraction (less than 1 %) of non-recalled genes were caused by a comparable effect (Figure E.5). Thus, the sensitivity (true positive rate), defined as the ratio of true positives to true positives plus false negatives, was 0.9896 for the *A. mellifera* RGS and 0.9918 for the *H. saltator* RGS. The accuracy, defined as the ratio of true positives plus true negatives to the total number of genes in the RGS, was 0.9965 for the *A. mellifera* RGS and 0.9978 for the *H. saltator* RGS.

For comparison, HaMStR v13.2.3 was run on the same datasets with comparable parameters. HaMStR identified 4,589 genes (99.22 %) in the RGS of *A. mellifera* (1 false positive) and 4,573 genes (98.88 %) in the RGS of *H. saltator* (2 false positives). This results in a sensitivity of 0.992 in the *A. mellifera* RGS and of 0.9883 in the *H. saltator* RGS, and an accuracy of 0.9975 in the *A. mellifera* RGS and of 0.9969 in the *H. saltator* RGS.

The input data on ortholog relations were retrieved from OrthoDB which contains OG information inferred in a purely automated fashion (Waterhouse et al., 2011). OrthoDB has been attested low numbers of false positives and spurious assignments (Trachana et al., 2011); the proportion of less than 1 % of the genes that were recalled wrongly by Orthograph are in line with these benchmarks. Orthograph and HaMStR perform roughly equally in accuracy and sensitivity when it comes to identifying single-copy orthologs.

4.3.2 Identification of splice variants or isoforms

We used Orthograph to assess orthologous amino acid sequences including isoforms in the RGS of the Florida carpenter ant, *Camponotus floridanus*, a species whose genes and corresponding proteins are part of the ortholog set analyzed before (see above). In the *C. floridanus* RGS, eight genes that are part of the ortholog set each encode an alternative isoform. Orthograph readily assigned the alternative isoforms of seven of these genes to the correct OGs. In the remaining gene, however, the amino acid sequence of the isoform that Orthograph could not find was very short (46 amino acids) in length. Only 21 of the 46 amino acid sites can be well aligned to the OG and were identified as BRH. It is possible that amino acid sequences that are significantly shorter than the majority of the OG are scored poorly by the pHMM search and/or the subsequent re-

verse search so that they eventually do not fulfill the BRH criterion and are not recognized by Orthograph.

HaMStR, in comparison, also identified all isoforms of seven of the eight genes correctly. However, it reports them as co-orthologs. Strictly speaking, this term is only correct when, while searching for single-copy orthologs, one or more copies of the same gene are identified. Orthograph, in addition to reporting, provides tabular output with alignment coordinates, HMM alignment bit scores and e-values for further statistical analyses.

While it would be highly desirable for users to also obtain information on the occurrence of different isoforms (or alternative transcripts on the transcriptional level) in different species, alternative transcripts are difficult to distinguish from transcripts of inparalogs or from transcript assembly artifacts without additional information, for example on the genealogy of the species, whose transcript libraries have been investigated, and/or on the transcript's expression level. However, Orthograph provides tabular output files that can facilitate corresponding downstream analyses. Specifically, the Orthograph output files inform about a) what transcripts form BRHs with ortholog groups and b) what transcripts assigned by Orthograph to the same ortholog group overlap (*i.e.*, partially refer to the same coding sequence) and could thus represent alternative transcripts (or assembly artifacts).

Protein isoforms and splice variants in the reference ortholog set can lead to systematic errors and false positives

The presence of isoforms and splice variants in an RGS dataset can lead to wrong clustering to OGs and/or false negatives (discarded sequences that should have been mapped elsewhere). Be-

cause it is impossible to know in advance which isoform of a gene or transcribed gene is present in a given transcript library, it is likely that a BRH search will fail if more than one highly similar amino acid sequence are present in the reference RGSs. This occurs because the best reverse search hit of a candidate ortholog against the database comprising all proteins in an RGS may return an isoform of the protein that was not used in the pHMM, leading to a failure to fulfill the BRH criterion. Therefore, isoforms should either be removed from RGS databases prior to using them in Orthograph (or in any reference-based orthology prediction tool, for that matter), or the OGs should be extended to also include the isoforms.

4.3.3 Identification of inparalogs

In order to demonstrate Orthograph's capabilities to detect inparalogous gene copies, we used it to assess genes that are known to have inparalogous copies in the RGS of the leafcutter ant, *Atta cephalotes* (Suen et al., 2011). Specifically, we retrieved an ortholog set from OrthoDB 5 comprising 301 OGs that contain genes that are known to be single copy in the genomes of *A. mellifera*, *C. floridanus*, *H. saltator*, *N. vitripennis*, and *T. castaneum*, but are multi-copy genes in *A. cephalotes*. These 301 OGs include altogether 647 single-copy and multi-copy genes from *A. cephalotes*: 273 are duplicated, 18 are triplicated, seven have four copies, two have six copies and one has seven copies. Orthograph readily assigned 583 of the 647 multi-copy genes to the correct OG (90.1 %). Two of the 301 OGs were not assigned, one of which contained four, the other contained two gene copies. In both cases, the genes from *A. cephalotes* were much shorter than the remaining genes in the OG (18 % resp. 19 % of the average amino acid sequence length), possibly leading to the respective transcripts failing to fulfil the BRH criterion in the reverse

search step due to an insufficient alignment length. These edge cases again highlight the importance of high-quality genome sequencing and annotation efforts, as they provide the basis for many downstream analyses, including full-length gene sequences for reference-based orthology assessment.

4.3.4 Non-redundant mapping of transcripts

In order to test whether Orthograph indeed does not assign transcripts to more than one OG, we re-analyzed the dataset published by Struck et al. (2014), who used HaMStR version 8 (Ebersberger et al., 2009). Orthograph assigned transcripts to 1,253 OGs, the same number as obtained by Struck et al. (2014). However, Orthograph found transcripts of the analyzed genes in, on average, slightly more taxa (Orthograph: 28.079, Struck et al. (2014): 26.699). None of the transcripts was assigned to more than one OG. In the dataset published by Struck et al. (2014), 274 transcripts were assigned redundantly, however the orthologous regions were not overlapping. As Struck et al. (2014) removed a total of 1.3 % of their sequences from the dataset due to redundantly assigned transcripts, Orthograph yielded 1.4 % more taxa per gene, leading to a denser data matrix for downstream (phylogenetic) analyses.

4.3.5 Computational performance of Orthograph

To demonstrate the computational performance of Orthograph, we searched 24 apoid wasp transcriptome assemblies for 5,561 selected OGs (sequence data are deposited at NCBI GenBank; accession numbers are listed in Additional file 2). The analysis time when using a single thread increases linearly with total transcriptome assembly length (Spearman rank correlation, $S = 326$,

$p \ll 0.001$, Supplemental Material, Figure E.3). Single-threaded analysis time also increases with the number of assembled transcripts, showing a linear trend, but no significant correlation (Spearman rank correlation, $S = 1,430$, $p = 0.069$).

Given that next-generation RNAseq datasets tend to be large and current comparative genomic investigations analyze hundreds, if not thousands of genes (*e.g.*, Misof et al. (2014), Jarvis et al. (2014), the 1000 plants initiative (`https://sites.google.com/a/ualberta.ca/onekp/`)), with a linear runtime increase Orthograph does not pose a time bottleneck for current and future large-scale studies such as the numerous group-specific subprojects of the 1KITE consortium (`http://1kite.org/subprojects.html`). For employment in high-performance cluster computing environments, Orthograph supports multi-threading: it offers a linear speedup of about 1x until up to four threads (Fig. E.4). Orthograph scales well with a speedup of 15 to 80 % per additional thread up to 12 threads. Using 16 threads reduces Orthograph running time to around 11 % compared to a single-threaded analysis.

Because most of the data are stored in a relational database on the hard drive, Orthograph requires only little memory and allows to re-evaluate stored search results with different parameters, which takes only a fraction of the original analysis time. In a centralized server-client setup using the MySQL database backend, the database management overhead is solely handled by the server, freeing CPU resources for the alignment searches on the clients. For installation in a grid computing environment where adding a dedicated database server is not feasible, the SQLite database backend (Hipp et al., 2016) is provided. The file-based SQLite database system can be applied anywhere thanks to its portable and performant implementation (and is installed

by default in most Linux distributions and Mac OS X), thus it is the default database backend in Orthograph.

4.3.6 Advantages of graph-based orthology prediction strategies

Orthograph uses a graph-based approach, like HaMStR and OrthoSelect as well as orthology prediction tools that assess orthology among genes in completely sequenced and annotated genomes, such as OrthoMCL, OrthoDB, OMA, or InParanoid. In contrast, tree-based orthology prediction strategies such as TreeFam, Ensembl Compara, or the one implemented in Capella-Gutierrez et al. (2014), employ an algorithm that reconciles a phylogenetic tree topology of a gene or gene set with the topology of the respective species phylogenetic tree. This requires a) a multiple sequence alignment (MSA) of a gene's amino acid or nucleotide sequences, and b) a phylogenetic tree inference. Both steps are not only computationally expensive, but also introduce additional sources of bias at each step. The much reduced computational complexity of a bidirectional alignment search compared to a phylogenetic tree inference enables Orthograph to run on standard workstation computers without necessitating a high-performance computing environment. A number of graph-based and tree-based orthology assessment methods have been reviewed by Trachana et al. (2011).

4.3.7 Reference-based orthology search accuracy depends on reference database quality

Reference-based algorithms for assessing transcript orthology can only be as accurate as the content of the database providing reference OGs. The results from testing the performance of

Orthograph affirm that reference-based orthology prediction requires adequate orthology delineation in reference genomes. These findings further highlight the necessity for reliable identification of ortholog relations in completely sequenced genomes as well as continuously updated databases such as OrthoDB that lay the foundation for a plethora of downstream comparative analyses. In order to provide comprehensive information, these databases require high-quality genomic data as well as reliable structural and functional gene annotation; thus, the importance of continued genome sequencing and rigorous annotation efforts must not be underestimated. Likewise, many assembled (draft) genomes are far from complete in terms of having properly identified their *actual* gene content (Denton et al., 2014), which also hinders reliable inference of orthology among them.

4.3.8 Reciprocal search by using HMMER and BLAST

Orthograph makes use of both pHMM-based and BLAST search technology. By combining these two fundamentally different alignment search algorithms, it draws considerable sensitivity and accuracy. Profile HMM-based similarity searches have been shown to be more sensitive than BLAST when it comes to detecting remotely related sequences (Eddy, 2011). By restricting the reverse BLAST search to only the (sub)sequence that was found to be putatively homologous during the pHMM search, the BLAST query becomes more informative. Therefore, the practice of using BLAST for the reverse search in Orthograph improves confidence in the subsequent orthology hypothesis by applying a conservative search criterion. For an illustration of the interrelations between the search results and their respective subsequences, see Supplemental Material, Figures E.1 and E.2.

BLAST uses a heuristic algorithm and does not guarantee an optimal local alignment. To also support a non-heuristic Smith-Waterman algorithm, we have, in addition to BLAST, implemented SWIPE (Rognes, 2011), which is also used in OrthoDB. SWIPE uses a BLAST database, thus the BLAST package is required to generate the database; however the SWIPE search algorithm does not result in inconsistencies that are possible with BLAST's alignment heuristic. Users can opt to use the SWIPE algorithm with appropriate configuration settings.

4.3.9 Limits of the methods

Orthograph is intended to map transcripts of a single species to reference OGs. Orthology or paralogy relations between genes of more than one species cannot be established using transcriptomic datasets as they are inherently incomplete. For assessing orthology among genes in completely sequenced and annotated genomes, specialized tools exist, such as OrthoMCL (Li et al., 2003), InParanoid (Sonnhammer & Östlund, 2015), or the OrthoDB toolset, which is now public (Kriventseva et al., 2015). Additionally, alternative transcripts or splice variants are difficult to distinguish in a *de novo* transcriptome assembly without additional read coverage data, which is why Orthograph refrains from explicitly predicting them. Orthograph does, however, report transcripts that are potential alternative transcripts or splice variants in order to allow researchers to further investigate them.

4.4 Conclusion

With Orthograph, we provide a software solution to accurately assign transcripts (and other coding sequences) to known groups (clusters) of orthologous genes (OGs). Orthograph maps

transcripts to the globally best matching OG, circumventing the problem of redundantly assigning transcripts to more than one OG. With its specific algorithm, Orthograph solves this issue that earlier implementations of graph-based BRH mapping strategies suffered from, while maintaining the high sensitivity and accuracy of the BRH approach. We developed Orthograph to be an asset in many fields by offering additional functionality compared to earlier implementations of graph-based BRH mapping strategies. Orthograph is easy to install and use and thereby facilitates comparative analyses of transcriptomic and other coding sequence data. It was furthermore designed to point users to possibly existing alternative transcripts and paralogous genes, thereby significantly broadening the scope of the software. The wide applicability of Orthograph has been demonstrated by its application in a phylogenomic study on apoid wasps using target DNA sequencing baits (Mayer et al., 2016) and the numerous subprojects of the international 1KITE project, which investigate intraordinal phylogenetic relationships of insects. Orthograph provides researchers with a convenient, performant, general-purpose tool for analyses in a plethora of disciplines in evolutionary biology.

Table 4.1: Results from the tests that compare Orthograph performance to HaMStR (Ebersberger et al., 2009). Sensitivity is defined as the ratio of true positives (TP) to TP plus false negatives (FN). Accuracy is defined as the ratio of TP plus true negatives (TN) to the total number of genes in the official gene set (OGS). FP, false positives. Note that the results are meant to demonstrate equality in performance despite algorithmic differences.

Software	Test	Genes	Species	OGS	Found	TP	FP	FN	Sens.	Acc.
Orthograph	single-copy	4,625	*A. mellifera*	15,314	4,582	4,577	5	48	0.990	0.996
Orthograph	single-copy	4,625	*H. saltator*	18,564	4,590	4,587	3	38	0.992	0.997
HaMStR	single-copy	4,625	*A. mellifera*	15,314	4,589	4,588	3	39	0.992	0.997
HaMStR	single-copy	4,625	*H. saltator*	18,564	4,573	4,571	2	54	0.988	0.996
Orthograph	isoforms	8	*C. floridanus*	17,064	7	7	0	1	0.875	0.999
HaMStR	isoforms	8	*C. floridanus*	17,064	7	7	0	1	0.875	0.999
Orthograph	inparalogs	647	*A. cephalotes*	18,093	583	583	0	6	0.901	0.996

References

Altenhoff, A. M. & Dessimoz, C. (2012). Inferring orthology and paralogy. *Methods in molecular biology (Clifton, N.J.)*, 855, 259–279.

Altenhoff, A. M., Škunca, N., Glover, N., Train, C.-M., Sueki, A., Piližota, I., Gori, K., Tomiczek, B., Müller, S., Redestig, H., Gonnet, G. H., & Dessimoz, C. (2015). The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*, 43(D1), D240–D249.

Bonasio, R., Zhang, G., Ye, C., Mutti, N. S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., Zhang, P., Huang, Z., Berger, S. L., Reinberg, D., Wang, J., & Liebig, J. (2010). Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. *Science (New York, N.Y.)*, 329(5995), 1068–1071.

Capella-Gutierrez, S., Kauff, F., & Gabaldón, T. (2014). A phylogenomics approach for selecting robust sets of phylogenetic markers. *Nucleic Acids Research*, (pp. gku071).

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLoS computational biology*, 10(12), e1003998.

Dutilh, B. E., van Noort, V., van der Heijden, R. T. J. M., Boekhout, T., Snel, B., & Huynen, M. A. (2007). Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics (Oxford, England)*, 23(7), 815–824.

Ebersberger, I., Strauss, S., & Von Haeseler, A. (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, 9(1), 157.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10), e1002195.

Emms, D. M. & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157.

Hipp, R. D., Kennedy, D., & Mistachkin, J. (2016). SQLite.

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114), 931–949.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldon, T., Capella-Gutierrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Nunez, A., Campos, P. F., Petersen, B., Sicheritz-

Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jonsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., & Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215), 1320–1331.

Jothi, R., Zotenko, E., Tasneem, A., & Przytycka, T. M. (2006). COCO-CL: Hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, 22(7), 779–788.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, 39, 309–338.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., & Koonin, E. V. (2011). Computational methods for gene orthology inference. *Briefings in Bioinformatics*, 12(5), 379–391.

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simao, F. A., Pozdnyakov, I. A., Ioannidis, P., & Zdobnov, E. M. (2015). OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, 43(D1), D250–D256.

Kvist, S. & Siddall, M. E. (2013). Phylogenomics of Annelida revisited: A cladistic approach using genome-wide expressed sequence tag data mining and examining the effects of missing data. *Cladistics*, 29(4), 435–448.

Li, L., Stoeckert, C., & Roos, D. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9), 2178–2189.

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design. *Molecular Biology and Evolution*, 33(7), 1875–1886.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von, R. B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang,

H., Wang, J., Wang, J., Kjer, K., & Zhou, X. (2014). Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science*, 346, 763–7.

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R. S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., & Niehuis, O. (2017). Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics*, 18, 111.

Rognes, T. (2011). Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics*, 12, 221.

Schreiber, F., Pick, K., Erpenbeck, D., Wörheide, G., & Morgenstern, B. (2009). OrthoSelect: A protocol for selecting orthologous groups in phylogenomics. *BMC Bioinformatics*, 10(1), 219.

Slater, G. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *Bmc Bioinformatics*, 6(1), 31.

Sonnhammer, E. L. L. & Östlund, G. (2015). InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43(Database issue), D234–239.

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., & Bleidorn, C. (2011). Phylogenomic analyses unravel annelid evolution. *Nature*, 471(7336), 95–98.

Struck, T. H., Wey-Fabrizius, A. R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., Klebow, S., Iakovenko, N., Hausdorf, B., Petersen, M., Kuck, P., Herlyn, H., & Hankeln,

T. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. *Molecular Biology and Evolution*, 31(7), 1833–1849.

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., Bouffard, P., Caldera, E. J., Cash, E., Cavanaugh, A., Denas, O., Elhaik, E., Favé, M.-J., Gadau, J., Gibson, J. D., Graur, D., Grubbs, K. J., Hagen, D. E., Harkins, T. T., Helmkampf, M., Hu, H., Johnson, B. R., Kim, J., Marsh, S. E., Moeller, J. A., Muñoz-Torres, M. C., Murphy, M. C., Naughton, M. C., Nigam, S., Overson, R., Rajakumar, R., Reese, J. T., Scott, J. J., Smith, C. R., Tao, S., Tsutsui, N. D., Viljakainen, L., Wissler, L., Yandell, M. D., Zimmer, F., Taylor, J., Slater, S. C., Clifton, S. W., Warren, W. C., Elsik, C. G., Smith, C. D., Weinstock, G. M., Gerardo, N. M., & Currie, C. R. (2011). The genome gequence of the leaf-cutter ant Atta cephalotes reveals insights into Its obligate symbiotic lifestyle. *PLoS Genet*, 7(2), e1002007.

Trachana, K., Larsson, T. A., Powell, S., Chen, W.-H., Doerks, T., Muller, J., & Bork, P. (2011). Orthology prediction methods: A quality assessment using curated protein families. *BioEssays*, 33(10), 769–780.

Tribolium Genome Sequencing Consortium (2008). The genome of the model beetle and pest Tribolium castaneum. *Nature*, 452(7190), 949–955.

Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J., & Kriventseva, E. V. (2011). OrthoDB: The hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Research*, 39(Database issue), D283–288.

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., The Nasonia Genome Working Group, Beukeboom, L. W., Desplan, C., Elsik, C. G., Grimme-

likhuijzen, C. J. P., Kitts, P., Lynch, J. A., Murphy, T., Oliveira, D. C. S. G., Smith, C. D., v. d. Zande, L., Worley, K. C., Zdobnov, E. M., Aerts, M., Albert, S., Anaya, V. H., Anzola, J. M., Barchuk, A. R., Behura, S. K., Bera, A. N., Berenbaum, M. R., Bertossa, R. C., Bitondi, M. M. G., Bordenstein, S. R., Bork, P., Bornberg-Bauer, E., Brunain, M., Cazzamali, G., Chaboub, L., Chacko, J., Chavez, D., Childers, C. P., Choi, J.-H., Clark, M. E., Claudianos, C., Clinton, R. A., Cree, A. G., Cristino, A. S., Dang, P. M., Darby, A. C., de Graaf, D. C., Devreese, B., Dinh, H. H., Edwards, R., Elango, N., Elhaik, E., Ermolaeva, O., Evans, J. D., Foret, S., Fowler, G. R., Gerlach, D., Gibson, J. D., Gilbert, D. G., Graur, D., Grunder, S., Hagen, D. E., Han, Y., Hauser, F., Hultmark, D., Hunter, H. C., Hurst, G. D. D., Jhangian, S. N., Jiang, H., Johnson, R. M., Jones, A. K., Junier, T., Kadowaki, T., Kamping, A., Kapustin, Y., Kechavarzi, B., Kim, J., Kim, J., Kiryutin, B., Koevoets, T., Kovar, C. L., Kriventseva, E. V., Kucharski, R., Lee, H., Lee, S. L., Lees, K., Lewis, L. R., Loehlin, D. W., Logsdon, J. M., Lopez, J. A., Lozado, R. J., Maglott, D., Maleszka, R., Mayampurath, A., Mazur, D. J., McClure, M. A., Moore, A. D., Morgan, M. B., Muller, J., Munoz-Torres, M. C., Muzny, D. M., Nazareth, L. V., Neupert, S., Nguyen, N. B., Nunes, F. M. F., Oakeshott, J. G., Okwuonu, G. O., Pannebakker, B. A., Pejaver, V. R., Peng, Z., Pratt, S. C., Predel, R., Pu, L.-L., Ranson, H., Raychoudhury, R., Rechtsteiner, A., Reid, J. G., Riddle, M., Romero-Severson, J., Rosenberg, M., Sackton, T. B., Sattelle, D. B., Schluns, H., Schmitt, T., Schneider, M., Schuler, A., Schurko, A. M., Shuker, D. M., Simoes, Z. L. P., Sinha, S., Smith, Z., Souvorov, A., Springauf, A., Stafflinger, E., Stage, D. E., Stanke, M., Tanaka, Y., Telschow, A., Trent, C., Vattathil, S., Viljakainen, L., Wanner, K. W., Waterhouse, R. M., Whitfield, J. B., Wilkes, T. E., Williamson, M., Willis, J. H., Wolschin, F., Wyder, S., Yamada, T., Yi, S. V., Zecher, C. N., Zhang, L., & Gibbs, R. A.

(2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, 327(5963), 343–348.

5

General Conclusion

The present thesis comprises research using a multitude of approaches from comparative genomics, ancestral reconstruction, phylogenetics, as well as algorithm design and implementation. By using a standardized TE annotation and a large taxon sampling that encompasses representatives from all major insect orders, it is able to draw conclusions that surpass inferences from intra-ordinal comparisons. Additionally, the development of a new software tool enables researchers to easily assess orthology within coding nucleotide data.

The software Orthograph (chapter 4) has proven to be a valuable asset which was used in 26 publications to date (Google Scholar, `https://scholar.google.com`, accessed 2019-03-01). Many of these aim to resolve order-level or family-level phylogenies from transcriptomic data (*e.g.*, Hymenoptera: Vespidae (Bank et al., 2017), Hymenoptera (Peters et al., 2017), Diptera: Acroceridae (Gillung et al., 2018), Hemiptera (Johnson et al., 2018), Diptera (Kutty et al., 2018), Palaeoptera (Simon et al., 2018)). In other publications, it was used to map target enrichment data to the correct genes (Mayer et al., 2016; Sann et al., 2018; Shin et al., 2018), often also for phylogenetic analyses. However, Orthograph was designed to be versatile, and this shows in its application in studies that investigate the evolution of specific gene families (Pauli et al., 2016; Dowling et al., 2017) or the distribution of DNA methylation in insects Provataris et al. (2018). Orthograph was reviewed in Nichio et al. (2017) and has recieved critical acclaim.

By providing, with Orthograph, a powerful and easy to use tool to identify orthologs in coding nucleotide data, the efforts described in chapter 4 facilitate future phylogenetic analyses. The growing amount of available transcriptomic data for more and more species enables researchers to further resolve phylogenies based on molecular datasets with increasing resolution and accuracy. Although for many species, molecular data will never be obtained (see page 10), our understanding of the remaining species' relationships will develop. This endeavor continues to be important since all studies dealing with aspects of evolution — also the ones in this thesis — necessitate a concept of species interrelationships. Thus, Orthograph contributes to furthering the field of biological (molecular) systematics, which in turn enables other fields such as evolutionary biology or comparative genomics to make meaningful inferences.

The focus of the empirical comparative studies in this thesis is on insect genomes. Insects are very different from vertebrates, plants, and fungi in both phenotype and genotype. This is also reflected in the processes that define their genome size with respect to the TE content. As in vertebrates, TE content is a predictor for genome size, however, DNA gain and loss rates do not affect genome size in insects as much as they do in vertebrates (Kapusta et al., 2017; Lindblad-Toh et al., 2005). Nevertheless, insect genome size does exhibit large fluctuations (Alfsnes et al., 2017), but these cannot be explained by differential TE activity alone. Similarly, the patterns of DNA methylation, which has been hypothesized to be involved in TE defense mechanisms, in insect genomes are drastically different from what has been observed in vertebrates or plants (Provataris et al., 2018; Suzuki & Bird, 2008). Apparently, insects do not rely on DNA methylation to inhibit TE proliferation and thereby genome size expansion. What else, then, could explain the large spread in insect genome size?

The answer to that question is likely not straigthforward. Instead of definitive answers, the present thesis provides a broad array of pointers for future research. For example, the RNA interference pathway genes were implicated in TE inhibition (Aravin et al., 2001; Czech et al., 2008) and are absent in some butterfly species that exhibited high TE content (Dowling et al., 2017). The large number of publicly available lepidopteran genomes provides ample opportunity to closely investigate these genes and shed more light on TE defense mechanisms. Incorporating some of these genomes, the study in chapter 2 characterizes the TE repertoire of 73 arthropod species, the largest taxon sampling for a comparative study on TE diversity in arthropods to date. The TE annotation data that were generated for the study are valuable as a resource to investigate the interaction of TEs with other genome components such as protein-coding

genes. The annotation results from six insect species were used in investigations on the protein-coding gene repertoire in insects (Wilbrandt et al. (in prep.). Another part of the TE annotation data was used by (Provataris et al., 2018), and the annotation pipeline was used to identify TEs in additional insect genomes. The annotation procedure is largely identical to the one used by Reinar (2016) who benchmarked the approach and showed it to be accurate and efficient. By combining several well-established algorithm implementations into an easy to use and fully automated pipeline, the method provides a tool to reliably annotate TEs in assembled genomes of non-model species.

The construction of a dated phylogeny for over 600 insect and arthropod species from the literature and publicly available DNA barcode data is unprecendented (chapter 3). This phylogeny will be valuable to researchers in many disciplines because it allows to set insights from other studies into context with the evolution of genome size in insects. In fact, the phylogeny also allows to map other phenotypic characters and to infer ancestral states for them, which is often a means to study and understand their evolution.

Also in chapter 3, I inferred ancestral genome sizes for 613 arthropod taxa including 520 insect species using that dated phylogeny and likewise publicly available genome size data (Gregory, 2018). While other studies have also exploited this database to set extant insect genome size into context with other phenotypic traits (Alfsnes et al., 2017; Gregory, 2011), none had information on the ancestral states of these traits due to lack of a phylogeny with branch lengths. Using the obtained ancestral genome size estimates, it was possible to classify the annotated TE content into ancestral and lineage-specific TEs. The study shows that there are practically no ancestral TEs in arthropod species that diverged from the common ancestor of their sister species ear-

lier than about 100 Mya. This result is consistent with prior findings that inactive TEs become unrecognizable after more than 50 Mya due to random mutations (Shedlock & Okada, 2000) and leads to the hypothesis that the majority of TEs in extant genomes might be dormant and possibly suppressed by the host genome defenses such as the RNAi or piRNA pathways or DNA methylation. Save from sustaining a high gene deletion rate (along with its drawbacks), no mechanism for targeted removal of TEs has been identified in eukaryotes. Thus, the most efficient defense appears to be to reduce TE activity and let random mutation degrade the TEs. In fact, in a study on nematode genomes, Szitenberg et al. (2016) argue that long-term TE dynamics are largely independent of host genome defenses, and that TE evolution in the host genome is determined by genetic drift. Only during a period of inefficient silencing, for example due to relaxed epigenetic modification, would the TEs be able to successfully proliferate (Slotkin & Martienssen, 2007; Zeh et al., 2009; Rebollo et al., 2010), leading to a burst in TE activity as often observed in the study in chapter 2. These periods of epigenetic silencing could be caused by environmental stress (Horvath & Slotte, 2017), an opportunity for adaptive evolution to work.

In general, the role of TEs in adaptive evolution cannot be disregarded. After decades of being viewed as mainly deleterious or neutral in effect on the host genome, the reputation of TEs changed when evidence for beneficial functions conferred by TEs was discovered (reviewed in Oliver & Greene (2012); Fedoroff (2013)). Especially in times of stress, when the organism is in need of genomic innovation to survive and adapt to new environmental conditions, TEs are thought to play an important role. For instance, TEs have been implicated in the rewiring of regulatory networks conferring dosage compensation (Ellison & Bachtrog, 2013; Chuong et al., 2016) or in adaptation to a different climate (González et al., 2010). Additionally, there

is no difference in the ratio of beneficial to deleterious TE-derived mutations when compared to mutations caused by single nucleotide polymorphisms (SNPs) (Akagi et al., 2013; Barrón et al., 2014). Therefore, the rate of beneficial or destructive effects due to TE activity is no different than that of random nucleotide substitutions, however, the effects are more profound when they are caused by TEs because they affect a larger region of the genome and can cause chromosomal rearrangements due to ectopic recombination (Gray, 2000; Fiston-Lavier et al., 2007). In *Drosophila melanogaster* and *D. miranda*, which exhibit similar rates of adaptation (Bachtrog, 2008) (*Drosophila melanogaster* also shows a high rate of TE-induced adaptation (González et al., 2008)), the study in chapter 2 inferred a two-fold difference in TE coverage. This is only an apparent contradiction, however: *D. miranda* exhibits a smaller current population size (Bachtrog, 2008), where the impact of genetic drift is amplified. The rate of fixation of a mutation is also higher in small populations (Kimura & Ohta, 1969), thus it is not surprising that a lower TE content in *D. miranda*, as found in chapter 2, is not reflected in a lower rate of adaptation.

Genotypic adaptation determines phenotypic adaptation, and thus defines the evolution of the species. The information encoded in the genome as well as the mechanisms and processes on the genomic and epigenetic level shape the interface between the organism and the outside world in ways both subtle and profound. While many of the building blocks and pathways that comprise the genome of complex organisms, such as *Drosophila*, mouse, or human, have been characterized, a large fraction of the genome and its components remains of unknown function. In particular, the function and purpose — if there is any — of TEs beside their role in adaptive evolution is still unclear. With the thorough characterization of the insect TE repertoire (chap-

ter 2) and the assessment of their influence on genome size dynamics (chapter 3), the research comprised in this thesis has added to the foundation for illuminating the many mysteries that remain in the genomes of modern metazoa.

References

Akagi, K., Li, J., & Symer, D. E. (2013). How do mammalian transposons induce genetic variation? A conceptual framework: The age, structure, allele frequency, and genome context of transposable elements may define their wide-ranging biological impacts. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 35(4), 397–407.

Alfsnes, K., Leinaas, H. P., & Hessen, D. O. (2017). Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecology and Evolution*, (pp. n/a–n/a).

Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., & Gvozdev, V. A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Current Biology*, 11(13), 1017–1027.

Bachtrog, D. (2008). Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes. *BMC Evolutionary Biology*, 8(1), 334.

Bank, S., Sann, M., Mayer, C., Meusemann, K., Donath, A., Podsiadlowski, L., Kozlov, A., Petersen, M., Krogmann, L., Meier, R., Rosa, P., Schmitt, T., Wurdack, M., Liu, S., Zhou, X.,

Misof, B., Peters, R. S., & Niehuis, O. (2017). Transcriptome and target DNA enrichment sequence data provide new insights into the phylogeny of vespid wasps (Hymenoptera: Aculeata: Vespidae). *Molecular Phylogenetics and Evolution*, 116, 213–226.

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population Genomics of Transposable Elements in Drosophila. *Annual Review of Genetics*, 48(1), 561–581.

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, 18(2), 71–86.

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J. A., Sachidanandam, R., Hannon, G. J., & Brennecke, J. (2008). An endogenous small interfering RNA pathway in Drosophila. *Nature*, 453(7196), 798–802.

Dowling, D., Pauli, T., Donath, A., Meusemann, K., Podsiadlowski, L., Petersen, M., Peters, R. S., Mayer, C., Liu, S., Zhou, X., Misof, B., & Niehuis, O. (2017). Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects. *Genome Biology and Evolution*, (pp. evw281).

Ellison, C. E. & Bachtrog, D. (2013). Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. *Science*, 342(6160), 846–850.

Fedoroff, N. V., Ed. (2013). *Plant Transposons and Genome Dynamics in Evolution: Fedoroff/Plant Transposons and Genome Dynamics in Evolution*. Oxford, UK: Wiley-Blackwell.

Fiston-Lavier, A.-S., Anxolabehere, D., & Quesneville, H. (2007). A model of segmental duplication formation in Drosophila melanogaster. *Genome Research*, 17(10), 1458–1470.

Gillung, J. P., Winterton, S. L., Bayless, K. M., Khouri, Z., Borowiec, M. L., Yeates, D. K., Kimsey, L. S., Meusemann, K., Misof, B., Shin, S., Zhou, X., Mayer, C., Petersen, M., & Wiegmann, B. M. (2018). Bias in big-data phylogenetics: Anchored phylogenomics unravels the evolution of spider flies (Acroceridae) and reveals discordance between nucleotides and amino acids. *Molecular Biology and Evolution*.

González, J., Karasov, T. L., Messer, P. W., & Petrov, D. A. (2010). Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila. *PLoS Genetics*, 6(4), e1000905.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., & Petrov, D. A. (2008). High Rate of Recent Transposable Element–Induced Adaptation in Drosophila melanogaster. *PLoS Biology*, 6(10), e251.

Gray, Y. H. (2000). It takes two transposons to tango: Transposable-element-mediated chromosomal rearrangements. *Trends in genetics: TIG*, 16(10), 461–468.

Gregory, T. R. (2011). *The Evolution of the Genome*. Academic Press.

Gregory, T. R. (2018). Animal Genome Size Database.

Horvath, R. & Slotte, T. (2017). The Role of Small RNA-Based Epigenetic Silencing for Purifying Selection on Transposable Elements in Capsella grandiflora. *Genome Biology and Evolution*, 9(10), 2911–2920.

Johnson, K. P., Dietrich, C. H., Friedrich, F., Beutel, R. G., Wipfler, B., Peters, R. S., Allen, J. M., Petersen, M., Donath, A., Walden, K. K. O., Kozlov, A. M., Podsiadlowski, L., Mayer,

C., Meusemann, K., Vasilikopoulos, A., Waterhouse, R. M., Cameron, S. L., Weirauch, C., Swanson, D. R., Percy, D. M., Hardy, N. B., Terry, I., Liu, S., Zhou, X., Misof, B., Robertson, H. M., & Yoshizawa, K. (2018). Phylogenomics and the evolution of hemipteroid insects. *Proceedings of the National Academy of Sciences*, 115(50), 12775–12780.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, (pp. 201616702).

Kimura, M. & Ohta, T. (1969). The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics*, 61(3), 763–771.

Kutty, S. N., Wong, W. H., Meusemann, K., Meier, R., & Cranston, P. S. (2018). A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families. *Systematic Entomology*, 43(3), 434–446.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., deJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill,

P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.-P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., & Lander, E. S. (2005). Genome sequence, comparative

analysis and haplotype structure of the domestic dog. *Nature*, 438(7069), 803–819.

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design. *Molecular Biology and Evolution*, 33(7), 1875–1886.

Nichio, B. T. L., Marchaukoski, J. N., & Raittz, R. T. (2017). New Tools in Orthology Analysis: A Brief Review of Promising Perspectives. *Frontiers in Genetics*, 8.

Oliver, K. R. & Greene, W. K. (2012). Transposable elements and viruses as factors in adaptation and evolution: An expansion and strengthening of the TE-Thrust hypothesis. *Ecology and Evolution*, 2(11), 2912–2933.

Pauli, T., Vedder, L., Dowling, D., Petersen, M., Meusemann, K., Donath, A., Peters, R. S., Podsiadlowski, L., Mayer, C., Liu, S., Zhou, X., Heger, P., Wiehe, T., Hering, L., Mayer, G., Misof, B., & Niehuis, O. (2016). Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects. *BMC Genomics*, 17, 861.

Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P. A., Heraty, J., Kjer, K. M., Klopfstein, S., Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B., & Niehuis, O. (2017). Evolutionary History of the Hymenoptera. *Current Biology*.

Provataris, P., Meusemann, K., Niehuis, O., Grath, S., Misof, B., & Wagner, G. (2018). Signatures of DNA Methylation across Insects Suggest Reduced DNA Methylation Levels in

Holometabola. *Genome Biology and Evolution*, 10(4), 1185–1197.

Rebollo, R., Horard, B., Hubert, B., & Vieira, C. (2010). Jumping genes and epigenetics: Towards new species. *Gene*, 454(1-2), 1–7.

Reinar, W. B. (2016). *In Silico Explorations of TE Activity, Diversity and Abundance across 74 Teleost Fish Genomes*. PhD thesis, University of Oslo, Oslo.

Sann, M., Niehuis, O., Peters, R. S., Mayer, C., Kozlov, A., Podsiadlowski, L., Bank, S., Meusemann, K., Misof, B., Bleidorn, C., & Ohl, M. (2018). Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. *BMC Evolutionary Biology*, 18(1), 71.

Shedlock, A. M. & Okada, N. (2000). SINE insertions: Powerful tools for molecular systematics. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 22(2), 148–160.

Shin, S., Clarke, D. J., Lemmon, A. R., Moriarty Lemmon, E., Aitken, A. L., Haddad, S., Farrell, B. D., Marvaldi, A. E., Oberprieler, R. G., & McKenna, D. D. (2018). Phylogenomic Data Yield New and Robust Insights into the Phylogeny and Evolution of Weevils. *Molecular Biology and Evolution*, 35(4), 823–836.

Simon, S., Blanke, A., & Meusemann, K. (2018). Reanalyzing the Palaeoptera problem – The origin of insect flight remains obscure. *Arthropod Structure & Development*, 47(4), 328–338.

Slotkin, R. K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4), 272–285.

Suzuki, M. M. & Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6), 465–476.

Szitenberg, A., Cha, S., Opperman, C. H., Bird, D. M., Blaxter, M. L., & Lunt, D. H. (2016). Genetic drift, not life history or RNAi, determine long term evolution of transposable elements. *Genome Biology and Evolution*, (pp. evw208).

Zeh, D. W., Zeh, J. A., & Ishida, Y. (2009). Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays*, 31(7), 715–726.

A

Co-authored publications using Orthograph

Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects

Benjamin Wipfler[a,b,1,2], Harald Letsch[c,1], Paul B. Frandsen[d,e,1], Paschalia Kapli[f,g,1], Christoph Mayer[h,1], Daniela Bartel[i], Thomas R. Buckley[j,k], Alexander Donath[h], Janice S. Edgerly-Rooks[l], Mari Fujita[m], Shanlin Liu[n,o], Ryuichiro Machida[m], Yuta Mashimo[m], Bernhard Misof[h], Oliver Niehuis[p], Ralph S. Peters[b], Malte Petersen[h], Lars Podsiadlowski[h], Kai Schütte[q], Shota Shimizu[m], Toshiki Uchifune[m,r], Jeanne Wilbrandt[h], Evgeny Yan[a,s], Xin Zhou[t], and Sabrina Simon[u,1,2]

[a]Institut für Spezielle Zoologie und Evolutionsbiologie, Friedrich-Schiller-University Jena, 07743 Jena, Germany; [b]Center of Taxonomy and Evolutionary Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; [c]Department für Botanik und Biodiversitätsforschung, Universität Wien, 1030 Vienna, Austria; [d]Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT 84604; [e]Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, Washington, DC 20002; [f]The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany; [g]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom; [h]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; [i]Department of Integrative Zoology, Universität Wien, 1090 Vienna, Austria; [j]New Zealand Arthropod Collection, Manaaki Whenua – Landcare Research, Auckland 1142, New Zealand; [k]School of Biological Sciences, The University of Auckland, Auckland 1142, New Zealand; [l]Department of Biology, College of Arts and Sciences, Santa Clara University, Santa Clara, CA 95053; [m]Sugadaira Research Station, Mountain Science Center, University of Tsukuba, Sugadaira Kogen, Ueda, Nagano 386-2204, Japan; [n]BGI-Shenzhen, Shenzhen 518083, China; [o]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen K, Denmark; [p]Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert Ludwig University, 79104 Freiburg, Germany; [q]Tierökologie und Naturschutz, Universität Hamburg, 20146 Hamburg, Germany; [r]Yokosuka City Museum, Fukadadai, Kanagawa 238-0016, Japan; [s]Borissiak Palaeontological Institute, Russian Academy of Sciences, 123 Moscow, Russia; [t]Department of Entomology, College of Plant Protection, China Agricultural University, Beijing 100083, China; and [u]Biosystematics Group, Wageningen University and Research, 6708 PB Wageningen, The Netherlands

Edited by Scott V. Edwards, Harvard University, Cambridge, MA, and approved December 11, 2018 (received for review November 8, 2018)

Polyneoptera represents one of the major lineages of winged insects, comprising around 40,000 extant species in 10 traditional orders, including grasshoppers, roaches, and stoneflies. Many important aspects of polyneopteran evolution, such as their phylogenetic relationships, changes in their external appearance, their habitat preferences, and social behavior, are unresolved and are a major enigma in entomology. These ambiguities also have direct consequences for our understanding of the evolution of winged insects in general; for example, with respect to the ancestral habitats of adults and juveniles. We addressed these issues with a large-scale phylogenomic analysis and used the reconstructed phylogenetic relationships to trace the evolution of 112 characters associated with the external appearance and the lifestyle of winged insects. Our inferences suggest that the last common ancestors of Polyneoptera and of the winged insects were terrestrial throughout their lives, implying that wings did not evolve in an aquatic environment. The appearance of the first polyneopteran insect was mainly characterized by ancestral traits such as long segmented abdominal appendages and biting mouthparts held below the head capsule. This ancestor lived in association with the ground, which led to various specializations including hardened forewings and unique tarsal attachment structures. However, within Polyneoptera, several groups switched separately to a life on plants. In contrast to a previous hypothesis, we found that social behavior was not part of the polyneopteran ground plan. In other traits, such as the biting mouthparts, Polyneoptera shows a high degree of evolutionary conservatism unique among the major lineages of winged insects.

lower winged insects | Polyneoptera | Pterygota | Neoptera | phylogenomics

The evolution of insect wings, which happened ∼400 Mya, led to a unique radiation and gave rise to the most species-rich group of organisms relative to their phylogenetic age (1, 2). One of the major lineages of winged insects is Polyneoptera, which comprises ∼40,000 described species in a total of 10 taxonomic orders. These include the well-known grasshoppers, crickets and allies (Orthoptera), stoneflies (Plecoptera), earwigs (Dermaptera), roaches and termites (Blattodea), mantids (Mantodea), stick and leaf insects (Phasmatodea), and also some of the least known and species-poor insect groups, including heelwalkers (Mantophasmatodea), ice crawlers (Grylloblattodea), webspinners (Embioptera), and ground lice (Zoraptera). Polyneoptera feature a wide spectrum of different lifestyles and body shapes. Some groups (e.g., roaches) exhibit extreme adaptations toward a ground-dwelling lifestyle, with hardened forewings and a dorsoventrally flattened body. Other groups, such

Significance

Polyneoptera is the only major lineage of winged insects (Pterygota) with an unresolved evolutionary history concerning important phenotypic traits like external shape, social behavior, and lifestyle. These ambiguities have far-reaching consequences for our understanding of the early evolution of winged insects. We closed this knowledge gap through large-scale phylogenomic analyses tracing traits concerning lifestyle and habitus within Polyneoptera and Pterygota. Both groups were ancestrally terrestrial in all developmental stages, implying that wings did not evolve in species living in water. All polyneopteran insects derive from a ground-dwelling insect with a largely unmodified body relative to the last common ancestor of winged insects. Intriguingly, different forms of social behavior, changes in lifestyle, and associated morphological specializations evolved multiple times within Polyneoptera.

Author contributions: B.W., H.L., P.B.F., and S. Simon designed research; B.W., H.L., P.B.F., C.M., and S. Simon performed research; D.B., A.D., B.M., O.N., R.S.P., M.P., L.P., and J.W. contributed new reagents/analytic tools; B.W., T.R.B., J.S.E.-R., R.M., R.S.P., and K.S. provided samples; B.W., H.L., P.B.F., P.K., C.M., D.B., T.R.B., A.D., J.S.E.-R., M.F., S.L., R.M., Y.M., K.S., S. Shimizu, T.U., E.Y., X.Z., and S. Simon analyzed data; and B.W., H.L., and S. Simon wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the National Center for Biotechnology Information database (accession nos. PRJNA273018–PRJNA273027 and PRJNA273029–PRJNA273062).

[1]B.W., H.L., P.B.F., P.K., C.M., and S. Simon contributed equally to this work.

[2]To whom correspondence may be addressed. Email: benjamin.wipfler@leibniz-zfmk.de or sabrina.simon@wur.nl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1817794116/-/DCSupplemental.

as stick and leaf insects and some mantids, live in the foliage and mimic leaves or twigs. Polyneoptera also feature a wide range of diets: some species, including most roaches and earwigs, are omnivorous, while others, such as stick and leaf insects and some grasshoppers, are strictly herbivorous. Additionally, the group includes carnivorous taxa (e.g., ambush predators like mantids and heelwalkers). Polyneoptera have also evolved a wide spectrum of insect social behaviors, ranging from maternal and biparental brood care to eusociality with a complex caste system.

The evolution of the above-mentioned traits is poorly understood, largely due to the lack of studies on character evolution and unresolved phylogenetic relationships among Polyneoptera. Previously published phylogenetic hypotheses were incongruent, including disagreement concerning the common ancestry of the group (*SI Appendix*, Fig. S9). As a result, Polyneoptera is the only major lineage of winged insects with a largely unresolved evolutionary history and thus strongly differs in this respect from Holometabola (insects with a complete metamorphosis) and from Acercaria (mostly sucking insects such as lice or true bugs). For the latter two lineages, detailed evolutionary scenarios, including ground-plan reconstructions for various character systems such as habitus (3), the holometabolan larvae (4), or social patterns (5), are available, although the common ancestry of Acercaria was recently challenged by a transcriptomic study (6). The evolutionary history of Polyneoptera is considered one of the major unresolved subjects in insect evolution (7) and not only affects our knowledge on Polyneoptera itself, but also has broad implications for our understanding of the early evolution of the winged insects. A major unresolved question related to this is whether winged insects evolved in an aquatic or terrestrial environment. The immatures of the two early diverging groups of winged insects—mayflies (Ephemeroptera) and damsel- and dragonflies (Odonata)—have an aquatic lifestyle; the same is observed in the polyneopteran stoneflies (Plecoptera), which have been hypothesized to represent the sister group of all remaining Polyneoptera or of all insects that can flex their wings above the abdomen (Neoptera) (8, 9). Thus, various authors assumed that early winged insect evolution occurred in an aquatic environment (e.g., refs. 9–12). Another unanswered question in the early evolution of winged insects concerns the ancestral lifestyle of adults. Many Polyneoptera inhabit narrow spaces such as litter, soil, and small cracks, and a similar lifestyle is found in the closest relatives of the winged insects: wingless bristletails (Archaeognatha) and silverfish (Zygentoma). However, this lifestyle is not found in mayflies, in damselflies or dragonflies, or in most groups of the other major lineages of winged insects (Acercaria and Holometabola). The question thus remains whether a ground-dwelling lifestyle represents an ancestral condition of the winged insects or whether some Polyneoptera returned secondarily to a life on ground. The current fossil record does not provide answers to these questions due to the lack of transitional fossils (13). Thus, a sound understanding of the phylogenetic relationships among the extant lineages of Polyneoptera is essential to trace the currently unresolved evolutionary trends within the in group, with possibly major implications for our knowledge of the evolution of winged insects.

We aim to close the above-outlined knowledge gaps in insect evolution by combining phylogenetic analyses of the largest transcriptomic dataset ever used for this purpose, comprising 3,014 protein-coding genes sampled from a total of 106 extant insect species, with a critical reevaluation of morphological and embryological arguments for all recovered interordinal nodes. We use the obtained phylogeny to reconstruct the evolution of 112 characters associated with habitus, habitat of larvae and adults, diet, and social behavior. Our study provides a formal reconstruction of the evolutionary history of both Polyneoptera and early winged insects (Pterygota).

Results

Phylogenomic Analyses. Our dataset comprised, in total, 106 insect species, representing all currently recognized polyneopteran orders and a representative sampling of outgroup taxa (Dataset S1). Phylogenetic analyses are based on five different datasets derived from 3,014 protein-coding genes: (*i*) $D_{AA,all}$, the complete dataset comprising 1,246,506 aligned amino acid sites; (*ii*) $D_{AA,decisive}$, a protein domain-based decisive dataset (i.e., a dataset which included only data blocks with representatives of selected taxonomic groups, see *Materials and Methods*) comprising 909,873 aligned amino acid sites; (*iii*) $D_{nuc,decisive}$, a corresponding decisive dataset comprising 909,873 aligned sites of second codon positions only; (*iv*) $D_{AA,genes}$, a gene-based decisive amino acid dataset of 2,061 genes comprising 832,237 amino acid sites; and (*v*) $D_{nuc,genes}$, a corresponding dataset comprising 832,237 aligned sites of second codon positions only. In addition to a maximum likelihood (ML) tree reconstruction based on a supermatrix approach and a multispecies coalescent (MSC) tree reconstruction, we applied Four-cluster Likelihood Mapping (FcLM) (14) to evaluate alternative signal for the major phylogenetic splits within Polyneoptera and to assess potential incongruent signal in our datasets that might not be revealed by a multispecies tree. To assess plausibility of our phylogenomic results, we compiled and assessed arguments from morphological and embryological data that support the inferred phylogenetic relationships (*SI Appendix*).

Our various phylogenomic analyses consistently revealed the monophyly of Pterygota, Neoptera, Eumetabola (Holometabola + Acercaria), and Holometabola. The only notable difference we observed was that the phylogenetic inferences from the analysis of amino acids (Fig. 1 and *SI Appendix*, Figs. S3 and S9) support monophyletic Acercaria, while the phylogenetic inferences from the analysis of the second codon positions (*SI Appendix*, Figs. S4 and S10) support lice (Psocodea) as the sister group to Holometabola. All our phylogenomic analyses found strong support for a monophyletic origin of Polyneoptera (Fig. 1 and *SI Appendix*, Figs. S3–S10), a result that is also corroborated by morphological and embryological evidence (*SI Appendix*, section 5) and that is not contradicted by the FcLM analyses (*SI Appendix*, section 4.4). Within Polyneoptera, all five supermatrix phylogenomic analyses place earwigs (Dermaptera) and ground lice (Zoraptera) as a sister group to the remaining Polyneoptera (Fig. 1 and *SI Appendix*, Figs. S3, S4, S9, and S10). Although this phylogenetic relationship is challenged by the MSC analyses, additional in-depth analyses of confounding signal and heterogeneity revealed further support for a sister group relationship of Dermaptera and Zoraptera (see discussion in *SI Appendix*, section 4.5). Consistent with earlier studies (6, 15), a polyneopteran clade ("core Polyneoptera") comprising grasshoppers, crickets and allies (Orthoptera), roaches and termites (Blattodea), mantids (Mantodea), stick and leaf insects (Phasmatodea), webspinners (Embioptera), heelwalkers (Mantophasmatodea), and ice crawlers (Grylloblattodea) is well supported by all our analyses. Stoneflies (Plecoptera) are placed as sister group to these core Polyneoptera. Furthermore, our analyses provide strong support for Dictyoptera, which is a close relationship of mantids, termites, and roaches. Stick and leaf insects (Phasmatodea) are inferred as the sister group of the webspinners (Embioptera) [Phasmatodea + Embioptera = Eukinolabia *sensu* (15)]. Eukinolabia form the sister group of Xenonomia *sensu* (15) [i.e., ice crawlers (Grylloblattodea) and heelwalkers (Mantophasmatodea)] (Fig. 1).

Character Evolution. To understand major evolutionary transitions, we applied the maximum parsimony and ML optimality criteria to trace a total of 112 behavioral, ecological, and morphological characters that played a major role during the evolution of Pterygota and of Polyneoptera (*SI Appendix*, section 6). Specifically, we studied characters associated with (*i*) the habitus of adults, (*ii*) social behavior, (*iii*) the habitat of larvae and of

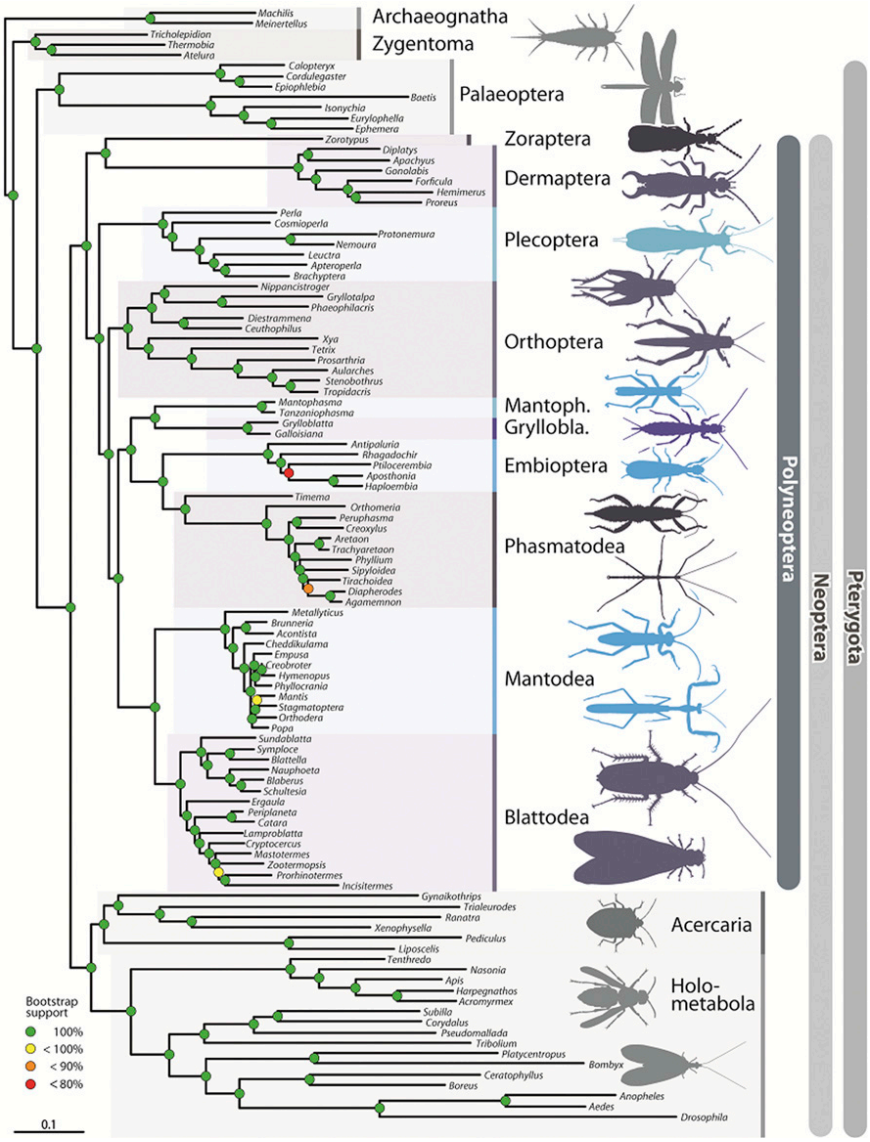

Fig. 1. Phylogenetic relationships among the major lineages of Polyneoptera, inferred from analyzing a decisive dataset comprising 909,873 amino acids sites and applying protein domain-based partitioning scheme ($D_{AA,decisive}$). Circles on nodes indicate bootstrap support values. Outgroup taxa are drawn in gray.

adults, and (*iv*) diet and lifestyle (Fig. 2). *SI Appendix*, section 6.1 provides a detailed list of these characters, their coding, and the results of their evolutionary reconstruction. In cases in which our results were ambiguous or contradictory to the paleontological record, we critically discuss the selected character state (*SI Appendix*, section 6.3). Based on our results, we created a model that illustrates the reconstructed characters of the last common ancestor of Polyneoptera (Fig. 3). *SI Appendix*, Fig. S17 illustrates which characters of the model are based on the analyses.

Discussion

The phylogenetic relationships among the polyneopteran groups were one of the most controversially discussed issues in systematic entomology (7). This ambiguity was caused by the fact that virtually every published phylogeny differed strongly from previous hypotheses (*SI Appendix*, Fig. S15). As a result, major evolutionary transitions and changes among the group were rarely addressed and thus remained poorly or not at all understood. The combination of our phylogenomic inferences (Fig. 1) together with our detailed evaluation of morphological and embryological arguments (*SI Appendix*, section 5) breaks this circle of phylogenetic ambiguity and we thus can reliably trace and interpret evolutionary transitions such as changes in lifestyle and phenotypic features within Polyneoptera and Pterygota.

To the best of our knowledge, no previous study addressed the possible outer appearance of the last common ancestor of Polyneoptera. Our results suggest that it had unspecialized biting mouthparts comparable to those of primarily wingless insects and of early branching winged insects such as damsel- and dragonflies or mayflies. Interestingly, there is not a single known polyneopteran species that secondarily modified these unspecialized biting mouthparts. This stands in contrast to the other major groups of winged insects, in which far-reaching transformations of the mouthparts have occurred several times independently, such as the suction feeding apparatus of lice (Psocodea) or mosquitos (Diptera) (e.g., ref. 16), or the "beak" of the extinct Paleodictyopterida (13). However, our data strongly suggest that there were multiple changes in the positioning of mouthparts within Polyneoptera: according to our analyses, the last common ancestor of the group held its mouthparts below the head capsule (orthognathy), a character state that represents the ancestral condition of Neoptera and Pterygota (Fig. 2*A*). Frontally oriented mouthparts (prognathy) might have

evolved at least four times separately within Polyneoptera (Fig. 2*A*). The actual number might even be higher, because our morphological reconstruction for a clade containing Eukinolabia (= Embioptera + Phasmatodea) and Xenonomia (= Grylloblattodea + Mantophasmatodea) remained ambiguous. Prognathy is usually associated with a raptorial lifestyle (17). Intriguingly, however, the

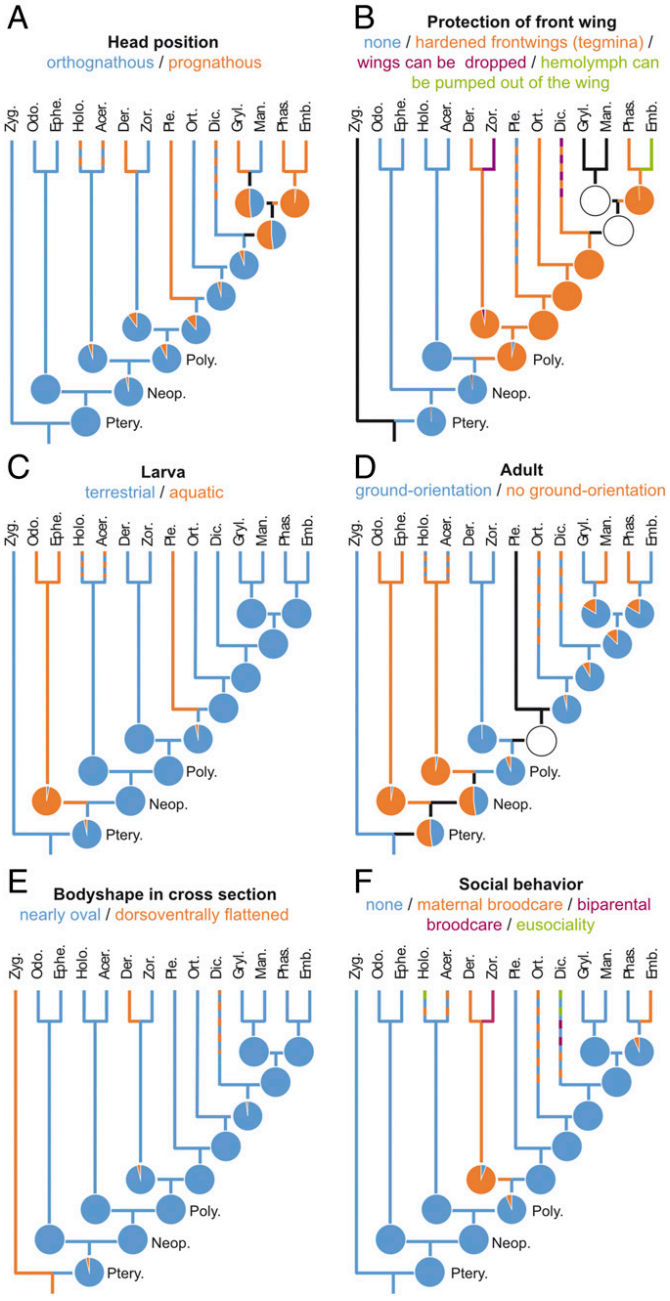

Fig. 2. (*A–F*) Reconstructed evolution of selected characters in Polyneoptera and related lineages (full list in Dataset S11). Pie charts indicate the ML results for the respective hypothesis at that node. Black lines and white pie charts imply ambiguous results or characters that are not applicable. Dotted lines indicate changes within the terminal taxa. Acer., Acercaria; Der., Dermaptera (earwigs); Dic., Dictyoptera (mantids, roaches, and termites); Emb., Embioptera (webspinners); Ephe., Ephemeroptera (mayflies); Gryl., Grylloblattodea (ice crawlers); Holo., Holometabola; Man., Mantophasmatodea (heelwalkers); Neop., Neoptera; Odo., Odonata (damsel- and dragonflies); Orth., Orthoptera (crickets, katydids, and grasshoppers); Phas., Phasmatodea (stick and leaf insects); Ple., Plecoptera (stoneflies); Poly., Polyneoptera; Ptery., Pterygota (winged insects); Zor., Zoraptera (ground lice); Zyg., Zygentoma (silverfish).

Fig. 3. Virtual model of the last common ancestor of Polyneoptera, inferred from analyzing 112 morphological characters. *SI Appendix*, Fig. S17 illustrates which parts of the model are based on results of the present analyses.

only exclusively predatory polyneopteran groups—the mantids (Mantodea) and the heelwalkers (Mantophasmatodea)—are not prognathous but orthognathous. A possible explanation for this evolutionary conservation in these two predatory groups is the fact that species of these two lineages use their forelegs rather than their mouthparts to catch prey (18). Prognathous polyneopteran insects show a wide spectrum of different diets, which include pure or partial herbivory (stick and leaf insects, stoneflies, and webspinners), omnivory (most earwigs), wood feeding (termites), or feeding on dead animals (ice crawlers). Prognathy also evolved in various subgroups of Holometabola and Acercaria (19), the other two major clades of Neoptera. Our data thus imply that a change of mouthpart orientation, typically resulting in far-reaching modifications of the head capsule, is a comparatively frequent evolutionary transition. Based on our reconstruction, the first polyneopteran insects also exhibited other ancestral pterygote traits such as long-segmented abdominal appendages (cerci), thoracic segments with approximately equal dimensions, and pentamerous tarsi. Although extant representatives of many orders of Polyneoptera, like stoneflies (Plecoptera) and earwigs (Dermaptera), reduced at least some of these features, the fossil record shows that their stem group representatives still had these ancestral characteristics (13). This implies that such reductions occurred several times separately (13) (*SI Appendix*, section 6.3).

Given the wide distribution of social behaviors among extant polyneopterans, it has been hypothesized that the last common polyneopteran ancestor exhibited social behavior in the form of maternal care (20). Although different forms of social behavior, such as maternal or biparental care, are indeed found in almost all polyneopteran insect groups, including ground lice (Zoraptera), earwigs (Dermaptera), crickets and grasshoppers (Orthoptera), roaches and termites (Blattodea), mantids (Mantodea), and webspinners (Embioptera) (20), our study contradicts this hypothesis. Instead, our results strongly suggest that maternal care evolved at least five times independently within Polyneoptera (Fig. 2*F*). The actual number might even be higher, since the social behavior of several lineages of mantids (Mantodea) is not documented and the evolution of maternal care in the roaches (Blattodea) (21) and crickets and grasshoppers (Orthoptera) (22) is only poorly understood. In accordance with Gilbert and Manica (20), we find that biparental care likely evolved separately in ground lice (Zoraptera) (23) and multiple lineages of roaches (21). Additionally, we confirm that the eusocial termites are the sister group of one of the subsocial groups of roaches, the Cryptocercidae (Fig. 1) (24).

It has been assumed that the first winged insects had aquatic larvae and that wings also evolved as an adaptation to an aquatic environment (e.g., refs. 9–12). This hypothesis is based on the presence of aquatic nymphs in mayflies (Ephemeroptera), in damsel- and dragonflies (Odonata), as well as in stoneflies (Plecoptera). In this respect, the morphology of stoneflies was considered to

reflect the ancestral condition of both Polyneoptera and Neoptera (8, 9). However, our results confirm the contradicting hypothesis, which states that the last common ancestors of Polyneoptera, Neoptera, and Pterygota were terrestrial throughout their entire lives (Fig. 2*C*) (13, 25). Specifically, this hypothesis is based on the derived position of the stoneflies within the polyneopteran insects and the increasing evidence for monophyletic Paleoptera (26) that is also supported by our analyses (Fig. 1). In addition, there is a third group of paleopteran insects, the extinct Paleodictyopterida: these large insects with beaks and, in many of its species, strongly widened prothoracic plates that contained a venation similar to those of wings ("six-winged insects") were also terrestrial throughout their life stages (13). The question concerning the habitat of the first winged insects also has direct implications for the ancestral function of wings. Although extant winged insects use their wings primarily for flight, this was likely not their original purpose, since early winglets were likely much shorter, immoveable, and therefore incapable of supporting powered flight. Three competing hypotheses have been proposed concerning the early function of these winglets: (*i*) winglets were used to control the descent while falling or jumping from a raised stand (e.g., a plant) (27), (*ii*) winglets evolved as organs of steering and propulsion in aquatic nymphs (10), and (*iii*) winglets were used as sails to achieve a quick distribution after the adults hatched from aquatic larvae (11, 12). Recent developmental studies have shown that the anatomical origin of wings does not provide any insight into this question because wings most likely evolved from a combination of the dorsal plates of the thoracic segments and branches of the legs (2, 28). However, because our results suggest that an aquatic larva was not part of the ground plan of winged insects, they consequently also reject the latter two hypotheses, which postulate the most recent common ancestor of winged insects having lived in an aquatic habitat. Our results thus favor the theory that wings originally developed as organs used for directed aerial descent when gliding from a raised stand (27). This concept is further supported by the fact that this behavior is also observed in some primarily wingless bristletails (Archaeognatha) (29).

Our comprehensive datasets allowed us to shed light on the ancestral habitat of adult early winged insects. Many adult extant Polyneoptera live on the ground or inhabit narrow spaces, such as leaf litter, cracks, crevices, or the spaces under bark. Most representatives of silverfish and bristletails, the closest relatives of winged insects, prefer a similar habitat. However, Paleoptera and most representatives of the other two major groups of neopteran insects, Holometabola and Acercaria, do not live in this kind of habitat. It thus remains unclear whether a preference for the ground represents an ancestral condition in the winged insects or whether Polyneoptera returned secondarily to this habitat. Our analyses suggest that the last common ancestor of Polyneoptera had a ground-dwelling lifestyle (Fig. 2*D*), although it remains ambiguous whether this is an ancestral or a derived feature (Fig. 2*D*). The shape of the body provides some hints on the evolutionary origin of the ground preference: specialized ground-dwelling insects, such as silverfish (Zygentoma), earwigs (Dermaptera), and roaches (Blattodea), usually have dorsoventrally flattened bodies. Based on our data, this was not the case for the ancestral polyneopteran insects (Fig. 2*E*). Instead, dorsoventrally flattened bodies evolved secondarily and separately in several polyneopteran lineages. Additionally, Paleodictyopterida apparently exhibited a non–ground-dwelling lifestyle (13). Thus, it appears likely to us that the ancestral polyneopteran returned secondarily to the ground and that the ancestral terrestrial Pterygota and Neoptera lived on plants or trees, which is congruent with the idea of early wings being used for directed aerial descent. According to our analyses, this change of habitat led to many adaptations in the body of the first polyneopteran insects (Fig. 3): Its antennae were comparatively long, it evolved unique attachment structures on the tarsi (euplantulae) that provided additional grip, and it had hardened forewings with a complete wing venation (tegmina) that protected the delicate hind wings when entering the substrate (Fig. 2*B*). The sclerotization of the forewings resulted in a reduced lift during flight that had to be counterbalanced by the hind wings that became triangular in shape (30). In contrast to other insects with sclerotized forewings, such as beetles and true bugs, Polyneoptera achieved this counterbalancing by a distinctly enlarged hind wing vannus with additional anal veins, which is also found in the last common ancestor (Fig. 2*B*). However, our data show that some polyneopteran insects secondarily and separately from each other reduced the sclerotized forewings—that is, most stoneflies (Plecoptera), ground lice (Zoraptera), webspinners (Embioptera), and termites (subgroup of Blattodea). Interestingly, species of the latter three groups live in habitats that provide little space to move and are closely associated to the substrate. In these groups, completely different mechanisms for wing protection evolved: Ground lice and termites evolved separately from each other the ability to drop their wings when entering a life in the ground. Male webspinners (females are wingless) can pump the hemolymph out of the wings, which then become extremely flexible and can be folded or crumbled over the thorax without damage (31). As effective as sclerotized forewings are as a protective measure, there is a tendency within exclusively ground-dwelling polyneopteran insects to either replace them with another mechanism (webspinners) or to use the flight capability only as a one-time method of dispersal and then to dispose the wings (ground lice and termites). Although the first polyneopteran lived most likely on the ground or was associated with the substrate, we can show that, within Polyneoptera, several groups adapted secondarily to a life in bushes and trees. Our analyses suggest at least four independent transitions to this habitat (Fig. 2*D*), either related to a plant diet (stick and leaf insects and some crickets, katydids, and grasshoppers) or to a predatory life on plants (most mantids and heelwalkers). However, different groups of grasshoppers and katydids (Orthoptera) might have colonized an arboreal habitat separately from each other (22), which would increase the number of transitions within Polyneoptera. To disguise themselves, all these lineages developed camouflage patterns, including extreme forms such as morphological and behavioral leaf and twig mimicry (e.g., ref. 32).

In summary, our study reveals that the highly specialized polyneopteran groups we observe today, such as the herbivorous stick and leaf insects, the eusocial termites, and the predatory mantids, are derived from an insect with many ancestral traits in both morphology and behavior. We demonstrate that some transitions, such as the evolution of social behavior or a life in the foliage, occurred several times separately. However, some polyneopteran features, such as the retention of the biting mouthparts, are uniquely conserved compared with the other major groups of winged insects. With our approach to combine a robust phylogeny with a formal reconstruction of character evaluation, we provide a comprehensive evolutionary picture of the Polyneoptera, thus closing a major gap in our understanding of insect evolution.

Materials and Methods

Phylogenomic Analyses. Our taxon sampling comprises a total of 106 extant insect species, including 72 polyneopteran representatives. Dataset S1 provides a detailed list of all species, including their collection data and National Center for Biotechnology Information (NCBI) accession numbers. Detailed information on the orthology prediction, the matrix assembly, the phylogenetic analyses, the FcLM strategy, and analyses to detect confounding signal are available in *SI Appendix*, sections 2–4. RNA extraction, cDNA library preparation, transcriptome sequencing, de novo assembly, and transcriptome quality assessment, as well as the submission procedure to the NCBI Transcriptome Shotgun Assembly database were performed as described by Peters et al. (5). Final assemblies were searched for transcripts of 3,014 protein-coding single-copy genes. Orthologous amino acid sequences were aligned, and resulting multiple sequence alignments (MSA) were assessed for quality and, if necessary, improved (or removed). We performed

ML phylogenetic analyses using a supermatrix approach with a partitioning scheme based on protein domains. According to ref. 6, we assume that protein domains are better evolutionary units to model sequence evolution than genes. In addition, we performed phylogenetic analyses with an MSC method and gene-based partitioned supermatrices applying the ML optimality criterion. All analyses have been carried out (*i*) on amino acid datasets and (*ii*) on corresponding nucleotide datasets using second codon positions only. For the gene-based ML and MSC analyses as well as for the protein domain-based ML analyses, MSA segments with putative alignment ambiguities or randomized data, identified with Aliscore (v.1.2) (33), were removed from the genes and the protein domain-based data blocks. The information content of each gene and protein domain-based data block was characterized with Mare (34), and uninformative genes and data blocks were removed. The resulting datasets were further optimized by including only those data blocks or genes that contained sequence information of selected taxa [decisive dataset, *sensu* (35)]. For the selection of optimal partitions and appropriate substitution models for the protein domain-based partitioning scheme, we applied PartitionFinder 2.0.0 (prerelease 10) using the rcluster algorithm (36). For the nucleotide supermatrix of the decisive dataset, we used PartitionFinder v.2.0.0 (prerelease 5) to select the partitions and the best-fitting substitution model using the iterative k-means search (37). Fifty independent phylogenetic tree inferences were performed using ExaML (v.3.0.16) (38), starting from different starting trees (40 random starting trees and 10 random stepwise addition parsimony starting trees). Phylogenetic analyses of the supermatrices with a gene-based partitioning scheme were inferred under the ML optimality criterion as implemented in IQ-TREE (v.1.5.5) (39) using the best-scoring substitution matrix for each gene partition as selected with ModelFinder implemented in IQ-TREE (40). In addition to the nonparametric bootstrap analysis, support for specific phylogenetic relationships was further assessed by FcLM (14) implemented in the software IQ-TREE v.1.4.1 (39). Gene trees used in the two MSC analyses (decisive amino acid datasets and corresponding nucleotide datasets using second codon positions only) were computed with IQ-TREE (v.1.6.3) with a model selection described in *SI Appendix*, section 4.1 and performing 1,000 nonparametric bootstrap replicates. The MSC analyses were carried out in ASTRAL III (v.5.5.6 and v.5.6.1) (41) on each dataset separately, with and without bootstrapping.

Character Evolution. To reconstruct the major transitions in morphology and lifestyle, we coded a total of 112 behavioral, ecological, and morphological characters for 106 selected species. We traced evolutionary transformations by maximum parsimony (all characters unordered) and ML-based mapping using Mesquite (42). Dataset S10 contains the character matrix. Dataset S11 contains the reconstructed character states for Polyneoptera. We discuss our results with respect to plausibility and paleontological findings in detail in *SI Appendix*, section 6.3. Based on these results, we created a virtual model that illustrates the retrieved characters of the last common ancestor of Polyneoptera (Fig. 3).

ACKNOWLEDGMENTS. We thank the entire team at 1K Insect Transcriptome Evolution (1KITE) (www.1kite.org); Karen Meusemann for the tremendous help with the phylogenomic analyses; Siavash Mirarab and Tandy Warnow for advice on ASTRAL analyses; Timothy Stephens for help with the R Package PhyloSortR; colleagues from the China National GeneBank and BGI-Shenzhen for their contributions, especially Guanliang Meng, Chengran Zhou, and Xu Su for careful maintenance of the 1KITE database and RNA vouchers; and Akino Akintola, Dieter Schulten, Evil Avian, Hans Pohl, Makiko Fukui, Reinhard Predel, Romolo Fochetti, Rüdiger Plarre, Sven Bradler, and Yoshie Jintsu-Uchifun for help with providing specimens. The study is part of the 1KITE project, which was supported by China National GeneBank and BGI-Shenzhen. We acknowledge financial support from the Deutsche Forschungsgemeinschaft Grant WI 4324/1-1 (to B.W.), the Klaus Tschira Foundation, and the European Research Council Grant ERC-2012- AdG 322790 (to P.K.).

1. Stork NE, McBroom J, Gely C, Hamilton AJ (2015) New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc Natl Acad Sci USA* 112:7519–7523.
2. Medved V, et al. (2015) Origin and diversification of wings: Insights from a neopteran insect. *Proc Natl Acad Sci USA* 112:15946–15951.
3. Beutel RG, et al. (2011) Morphological and molecular evidence converge upon a robust phylogeny of the megadiverse Holometabola. *Cladistics* 27:341–355.
4. Truman JW, Riddiford LM (1999) The origins of insect metamorphosis. *Nature* 401: 447–452.
5. Peters RS, et al. (2017) Evolutionary history of the hymenoptera. *Curr Biol* 27: 1013–1018.
6. Misof B, et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.
7. Whitfield JB, Kjer KM (2008) Ancient rapid radiations of insects: Challenges for phylogenetic analysis. *Annu Rev Entomol* 53:449–472.
8. Beutel RG, Gorb SN (2006) A revised interpretation of the evolution of attachment structures in Hexapoda with special emphasis on Mantophasmatodea. *Arthropod Syst Phylogeny* 64:3–25.
9. Zwick P (2009) The plecoptera–who are they? The problematic placement of stoneflies in the phylogenetic system of insects. *Aquat Insects* 31:181–194.
10. Kukalova-Peck J (1978) Origin and evolution of insect wings and their relation to metamorphosis, as documented by the fossil record. *J Morphol* 156:53–125.
11. Marden JH, Kramer MG (1994) Surface-skimming stoneflies: A possible intermediate stage in insect flight evolution. *Science* 266:427–430.
12. Thomas MA, Walsh KA, Wolf MR, McPheron BA, Marden JH (2000) Molecular phylogenetic analysis of evolutionary trends in stonefly wing structure and locomotor behavior. *Proc Natl Acad Sci USA* 97:13178–13183.
13. Grimaldi D, Engel MS (2005) *Evolution of the Insects* (Cambridge Univ Press, New York), p 772.
14. Strimmer K, von Haeseler A (1997) Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci USA* 94:6815–6819.
15. Terry MD, Whiting MF (2005) Mantophasmatodea and phylogeny of the lower neopterous insects. *Cladistics* 21:240–257.
16. Huang DY, et al. (2016) New fossil insect order Permopsocida elucidates major radiation and evolution of suction feeding in hemimetabolous insects (Hexapoda: Acercaria). *Sci Rep* 6:23004.
17. Beutel RG, Friedrich F, Yang X-K, Ge S-Q (2014) *Insect Morphology and Phylogeny: A Textbook for Students of Entomology* (Walter de Gruyter, Berlin).
18. Prete F, Wells H, Wells P, Hurd L (1999) *The Praying Mantids* (The Johns Hopkins Univ Press, Baltimore).
19. Matsuda R (1965) Morphology and evolution of the insect head. *Mem Am Entomol Inst* 4:1–334.
20. Gilbert JD, Manica A (2015) The evolution of parental care in insects: A test of current hypotheses. *Evolution* 69:1255–1270.
21. Bell WJ, Roth LM, Nalepa CA (2007) *Cockroaches: Ecology, Behavior, and Natural History* (JHU Press, Baltimore).
22. Gwynne DT (1995) Phylogeny of the Ensifera (Orthoptera): A hypothesis supporting multiple origins of acoustical signalling, complex spermatophores and maternal care in crickets, katydids, and weta. *J Orthoptera Res* 4:203–218.
23. Choe JC (1994) Sexual selection and mating system in *Zorotypus gurneyi* Choe (Insecta: Zoraptera): I. Dominance hierarchy and mating success. *Behav Ecol Sociobiol* 34:87–93.
24. Inward D, Beccaloni G, Eggleton P (2007) Death of an order: A comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. *Biol Lett* 3:331–335.
25. Will KW (1995) Plecopteran surface-skimming and insect flight evolution. *Science* 270: 1684–1686.
26. Thomas JA, Trueman JW, Rambaut A, Welch JJ (2013) Relaxed phylogenetics and the palaeoptera problem: Resolving deep ancestral splits in the insect phylogeny. *Syst Biol* 62:285–297.
27. Wootton RJ, Ellington CP (1991) Biomechanics and the origin of insect flight. *Biomechanics and Evolution*, eds Rayner JMV, Wootton RJ (Cambridge Univ Press, Cambridge, UK), pp 99–112.
28. Linz DM, Tomoyasu Y (2018) Dual evolutionary origin of insect wings supported by an investigation of the abdominal wing serial homologs in *Tribolium*. *Proc Natl Acad Sci USA* 115:E658–E667.
29. Yanoviak SP, Kaspari M, Dudley R (2009) Gliding hexapods and the origins of insect aerial behaviour. *Biol Lett* 5:510–512.
30. Brodsky AK (1994) *The Evolution of Insect Flight* (Oxford Univ Press, Oxford).
31. Ross ES (1970) Biosystematics of the Embioptera. *Annu Rev Entomol* 15:157–172.
32. Wedmann S, Bradler S, Rust J (2007) The first fossil leaf insect: 47 million years of specialized cryptic morphology and behavior. *Proc Natl Acad Sci USA* 104:565–569.
33. Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst Biol* 58: 21–34.
34. Misof B, et al. (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics* 14:348.
35. Dell'Ampio E, et al. (2014) Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects. *Mol Biol Evol* 31:239–249.
36. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol* 14:82.
37. Frandsen PB, Calcott B, Mayer C, Lanfear R (2015) Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evol Biol* 15:13.
38. Kozlov AM, Aberer AJ, Stamatakis A (2015) ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31:2577–2579.
39. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274.
40. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589.
41. Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.
42. Maddison W, Maddison D (2015) Mesquite: A modular system for evolutionary analysis, version 3.02. Available at www.mesquiteproject.org. Accessed December 5, 2017.

Phylogenomics and the evolution of hemipteroid insects

Kevin P. Johnson[a,1], Christopher H. Dietrich[a], Frank Friedrich[b], Rolf G. Beutel[c], Benjamin Wipfler[c,d], Ralph S. Peters[d], Julie M. Allen[a,e], Malte Petersen[f], Alexander Donath[f], Kimberly K. O. Walden[g], Alexey M. Kozlov[h], Lars Podsiadlowski[f,i], Christoph Mayer[f], Karen Meusemann[f,j,k], Alexandros Vasilikopoulos[f], Robert M. Waterhouse[l], Stephen L. Cameron[m], Christiane Weirauch[n], Daniel R. Swanson[a], Diana M. Percy[o,p], Nate B. Hardy[q], Irene Terry[r], Shanlin Liu[s], Xin Zhou[t], Bernhard Misof[f], Hugh M. Robertson[g], and Kazunori Yoshizawa[u]

[a]Illinois Natural History Survey, Prairie Research Institute, University of Illinois at Urbana–Champaign, Champaign, IL 61820; [b]Institut für Zoologie, Universität Hamburg, 20146 Hamburg, Germany; [c]Institut für Zoologie und Evolutionsforschung, Friedrich-Schiller-Universität Jena, 07743 Jena, Germany; [d]Center of Taxonomy and Evolutionary Research, Arthropoda Department, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; [e]Department of Biology, University of Nevada, Reno, NV 89557; [f]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; [g]Department of Entomology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [h]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany; [i]Institute of Evolutionary Biology and Ecology, University of Bonn, 53121 Bonn, Germany; [j]Evolutionary Biology and Ecology, Institute for Biology I (Zoology), University of Freiburg, 79104 Freiburg, Germany; [k]Australian National Insect Collection, Commonwealth Scientific and Industrial Research Organisation National Research Collections Australia, Acton, ACT 2601 Canberra, Australia; [l]Department of Ecology and Evolution, University of Lausanne and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [m]Department of Entomology, Purdue University, West Lafayette, IN 47907; [n]Department of Entomology, University of California, Riverside, CA 92521; [o]Department of Life Sciences, Natural History Museum, London, SW7 5BD United Kingdom; [p]Department of Botany, University of British Columbia, Vancouver V6T 1Z4, Canada; [q]Department of Entomology and Plant Pathology, Auburn University, Auburn, AL 36849; [r]School of Biological Sciences, University of Utah, Salt Lake City, UT 84112; [s]BGI-Shenzhen, Shenzhen, 518083 Guangdong Province, People's Republic of China; [t]Department of Entomology, China Agricultural University, 100193 Beijing, People's Republic of China; and [u]Systematic Entomology, Hokkaido University, Sapporo, 060-8589 Japan

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved October 25, 2018 (received for review September 13, 2018)

Hemipteroid insects (Paraneoptera), with over 10% of all known insect diversity, are a major component of terrestrial and aquatic ecosystems. Previous phylogenetic analyses have not consistently resolved the relationships among major hemipteroid lineages. We provide maximum likelihood-based phylogenomic analyses of a taxonomically comprehensive dataset comprising sequences of 2,395 single-copy, protein-coding genes for 193 samples of hemipteroid insects and outgroups. These analyses yield a well-supported phylogeny for hemipteroid insects. Monophyly of each of the three hemipteroid orders (Psocodea, Thysanoptera, and Hemiptera) is strongly supported, as are most relationships among suborders and families. Thysanoptera (thrips) is strongly supported as sister to Hemiptera. However, as in a recent large-scale analysis sampling all insect orders, trees from our data matrices support Psocodea (bark lice and parasitic lice) as the sister group to the holometabolous insects (those with complete metamorphosis). In contrast, four-cluster likelihood mapping of these data does not support this result. A molecular dating analysis using 23 fossil calibration points suggests hemipteroid insects began diversifying before the Carboniferous, over 365 million years ago. We also explore implications for understanding the timing of diversification, the evolution of morphological traits, and the evolution of mitochondrial genome organization. These results provide a phylogenetic framework for future studies of the group.

phylogeny | systematics | transcriptomes | Hemiptera | Psocodea

The hemipteroid insect orders, Psocodea (bark lice and parasitic lice), Thysanoptera (thrips), and Hemiptera (true bugs and allies; i.e., hemipterans), with over 120,000 described species, comprise well over 10% of known insect diversity. However, the evolutionary relationships among the major lineages of these insects are not yet resolved. Recent phylogenomic analyses questioned the monophyly of this group (1) demanding a reconsideration of the evolution of hemipteroid and holometabolous insects. We assess these prior results, which placed Psocodea as the sister taxon to Holometabola (insects with complete metamorphosis; e.g., wasps, flies, beetles, butterflies), and uncover relationships within and among hemipteroid insect orders by analyzing a large phylogenomic dataset covering all major lineages of hemipteroid insects.

Knowledge of the phylogeny of these insects is important for several reasons. First, major transitions between the mandibulate (chewing) mouthpart insect groundplan and "piercing–sucking" mouthparts occurred in this group. In particular, thrips and hemipterans, and some ectoparasite lice in Psocodea, have highly modified mouthparts adapted for feeding on fluids and, hence, differ markedly from their mandibulate ancestors. Through a series of remarkable modifications, hemipteroids acquired a piercing–sucking mode of feeding in both immature and adult stages that enabled them to feed not only on plant vascular fluids, but also on blood and other liquid diets. Resolution of the evolutionary tree of hemipteroid insects is needed to provide a framework for

Significance

Hemipteroid insects constitute a major fraction of insect diversity, comprising three orders and over 120,000 described species. We used a comprehensive sample of the diversity of this group involving 193 genome-scale datasets and sequences from 2,395 genes to uncover the evolutionary tree for these insects and provide a timescale for their diversification. Our results indicated that thrips (Thysanoptera) are the closest living relatives of true bugs and allies (Hemiptera) and that these insects started diversifying before the Carboniferous period, over 365 million years ago. The evolutionary tree from this research provides a backbone framework for future studies of this important group of insects.

Author contributions: K.P.J., C.H.D., F.F., R.G.B., B.W., R.S.P., K.M., X.Z., B.M., H.M.R., and K.Y. designed research; K.P.J., C.H.D., R.G.B., B.W., R.S.P., J.M.A., M.P., A.D., K.K.O.W., A.M.K., L.P., C.M., K.M., A.V., R.M.W., S.L., X.Z., and K.Y. performed research; K.P.J., C.H.D., F.F., B.W., R.S.P., K.M., C.W., D.R.S., D.M.P., N.B.H., I.T., and K.Y. contributed new reagents/analytic tools; K.P.J., C.H.D., R.G.B., B.W., R.S.P., J.M.A., M.P., A.D., K.K.O.W., A.M.K., L.P., C.M., K.M., A.V., R.M.W., and K.Y. analyzed data; and K.P.J., C.H.D., F.F., R.G.B., B.W., R.S.P., J.M.A., M.P., A.D., K.K.O.W., A.M.K., L.P., C.M., K.M., A.V., R.M.W., S.L.C., C.W., D.R.S., D.M.P., N.B.H., I.T., S.L., X.Z., B.M., H.M.R., and K.Y. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in NCBI (accession nos. SRA SRR1821891–SRR1821980, SRR2051465–SRR2051515, and SRR921611–SRR921660). Gene sets, alignments, trees, quartet likelihood mapping results, morphological data matrices, and dating analyses results were deposited in Dryad repository, 10.5061/dryad.t4f4g85.

[1]To whom correspondence should be addressed. Email: kpjohnso@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1815820115/-/DCSupplemental.

understanding morphological transitions that occurred in this group, as well as to provide a timeframe over which these changes occurred.

In addition, several lineages of hemipteroid insects (particularly thrips and Psocodea) underwent major reorganizations of their mitochondrial genomes, including the emergence of minicircles (2). Understanding how these changes in mitochondrial genome organization occurred requires knowledge of evolutionary relationships to document in which lineages these changes first arose. Finally, hemipteroids are among the most abundant insects (3) and are therefore key components of terrestrial and aquatic food webs (4). Thus, a robust backbone phylogenetic framework is needed to place ecological studies in their evolutionary context and for use in comparative genomic and macroevolutionary analyses.

Despite their importance, relatively few studies have addressed the relationships among the major groups of hemipteroid insects [Paraneoptera, sensu stricto (excluding Zoraptera), also termed Acercaria]. While a recent large transcriptome-based phylogenomic analysis of insects (1) provided a well-resolved and strongly supported phylogenetic framework for the insect orders in general, it did not sample intensively within individual orders and recovered some unexpected relationships. Among the most puzzling was the nonmonophyly of the hemipteroid insects, with Psocodea as the sister taxon of holometabolous insects rather than as sister to thrips plus hemipterans (Condylognatha). Although this result was congruent with one earlier analysis based on three nuclear protein-coding genes (5), it had not been proposed in other molecular phylogenetic or morphological studies. Previous morphological studies indicated monophyly of hemipteroid insects with Psocodea sister to thrips plus hemipterans (6–9), or sometimes a group comprising thrips plus Psocodea (10, 11).

Another unexpected relationship recovered by Misof et al. (1) was the placement of moss bugs (Coleorrhyncha) as sister to a group comprising leafhoppers, cicadas, and relatives (Auchenorrhyncha) instead of sister to true bugs (Heteroptera). A recent morphological study also found some support for moss bugs sister to Auchenorrhyncha (12). In contrast, prior analyses based on morphology (e.g., ref. 9) and DNA sequence data (e.g., ref. 13) consistently placed moss bugs as sister to true bugs. An analysis of a reduced gene set from transcriptome data (14) also recovered moss bugs as sister to true bugs, while the full gene set placed moss bugs as sister to Auchenorrhyncha. Analysis of mitochondrial genomes (15) produced an even more unconventional result, with moss bugs placed as the sister taxon of planthoppers (Fulgoroidea), making Auchenorrhyncha paraphyletic. Thus, it is important to investigate the placement of moss bugs in more detail with both expanded taxon and gene sampling.

We evaluated these possible conflicts among analyses by analyzing a more comprehensive dataset comprising an increased number of clusters of orthologous sequence groups (2,395 protein-coding, single-copy genes) as well as an increased taxon sample within hemipteroid insects: 160 samples vs. 22 sampled by Misof et al. (1). We included representatives of all major hemipteroid lineages (sub- and infraorders). Outgroups comprised 33 species of holometabolous and nonholometabolous insect orders. This dataset enabled us to test the hypothesis of nonmonophyly of hemipteroid insects and also provides a more detailed backbone framework for the hemipteroid phylogeny. We evaluate the implications of this phylogeny for understanding the evolution of feeding strategy, morphology, and mitochondrial genome organization of this major group of insects.

Results

Phylogeny of Hemipteroid Insect Orders. Separate amino acid sequence alignments of the 2,395 single-copy genes across 193 terminal taxa (*SI Appendix*, Tables S1–S4) yielded a concatenated supermatrix of 859,518 aligned amino acid positions, which was used in subsequent phylogenetic analyses. A concatenated nucleotide sequence supermatrix of only first and second codon positions resulted in ∼1.72 million aligned nucleotide sequence sites. Tree reconstructions based on the nucleotide sequence data supported a phylogenetic tree (Fig. 1 and *SI Appendix*, Figs. S1 and S2) with 172/190 (∼90%) of all nodes supported in 100% of bootstrap replicates. The tree based on amino acid sequence data (*SI Appendix*, Fig. S3) was highly concordant with that based on nucleotide data. Analysis of an optimized amino acid dataset (*SI Appendix*, *Supplemental Materials and Methods*) produced a tree (*SI Appendix*, Fig. S4) that was identical to that based on all amino acids with respect to relationships among orders, suborders, infraorders, and superfamilies, but had some minor rearrangements within these groups.

Considering relationships within and among orders in more detail, the thrips (Thysanoptera) were recovered with 100% bootstrap support as the sister taxon of Hemiptera (i.e., monophyletic Condylognatha), although only 68% of quartets supported this result in four-cluster likelihood mapping (FcLM) (*SI Appendix*, Tables S5 and S6). As in the study of Misof et al. (1), Psocodea was placed as the sister taxon of Holometabola in 100% of bootstrap replicates, rendering hemipteroid insects paraphyletic. However, only 25% of quartets supported Psocodea as sister to Holometabola, compared with 67% of the quartets supporting hemipteroid insect monophyly. Results from the FcLM imply that the placement of Psocodea as sister to Holometabola is unstable and may be due to confounding phylogenetic signal (e.g., from heterogeneous composition of amino acid sequences, nonstationarity of substitution processes, or nonrandom distribution of missing data) and is also dependent on the taxon sample. However, permutation tests of these results suggested the impact of these potential confounding signals on the topology was minor (*SI Appendix*, Table S6). To evaluate whether the parasitic lice in particular (Phthiraptera), which have elevated substitution rates compared with other hemipteroids (16), were a possible source of conflicting signal, we compared quartets with and without these ectoparasitic insects as the representative of Psocodea. However, the support from FcLM for monophyly of hemipteroid insects was highly similar whether parasitic lice were included (66%) or not (67%).

Morphological character mapping over three possible alternative topologies (*SI Appendix*, Fig. S5) revealed no apomorphies supporting Psocodea + Holometabola. In contrast, there are 14 potential apomorphies for the monophyly of Paraneoptera. These results indicate that there is more agreement between morphology and the FcLM results, compared with the supermatrix analyses with all taxa. For Coleorrhyncha (moss bugs), three characters are apomorphies for a sister relationship to Auchenorrhyncha (leafhoppers and relatives) but two other characters appear to support a sister relationship to Heteroptera (true bugs).

In general, the phylogenetic results from transcriptomes are congruent with the generally accepted classification schemes within these insect orders. Bark lice and parasitic lice (Psocodea) together are monophyletic. As has been suggested based on both morphological (17) and molecular (16, 18) analyses, the parasitic lice are embedded within free-living bark lice, being the sister taxon of book lice (Liposcelididae), which makes the bark lice ("Psocoptera") paraphyletic. In contrast to results based on 18S rDNA sequences (18), parasitic lice (Phthiraptera) were supported as a monophyletic group in our analyses, which included representatives of all four suborders of parasitic lice.

The thrips (Thysanoptera) were found to be monophyletic. The thrips family Phlaeothripidae was recovered as the sister taxon to the remaining thrips (Aeolothripidae + Thripidae), congruent with previous molecular analyses and the current classification of Thysanoptera into the suborders Tubulifera (i.e., Phlaeothripidae) and Terebrantia (all other thrips) (19).

The order Hemiptera was also monophyletic. Within Hemiptera, Sternorrhyncha (whiteflies, psyllids, scales, and aphids) was recovered as the sister taxon of the remaining hemipterans. Recent classification schemes (20) and prior molecular studies (13, 21) have placed the enigmatic moss bugs as the sister taxon of true bugs. However, our results recovered moss bugs as the sister taxon of Auchenorrhyncha (leafhoppers, planthoppers, and relatives), which was also found by Misof et al. (1). In FcLM analyses, 96% of quartets placed moss bugs with Auchenorrhyncha, suggesting little underlying conflict in the data for this result (*SI Appendix*, Table S6).

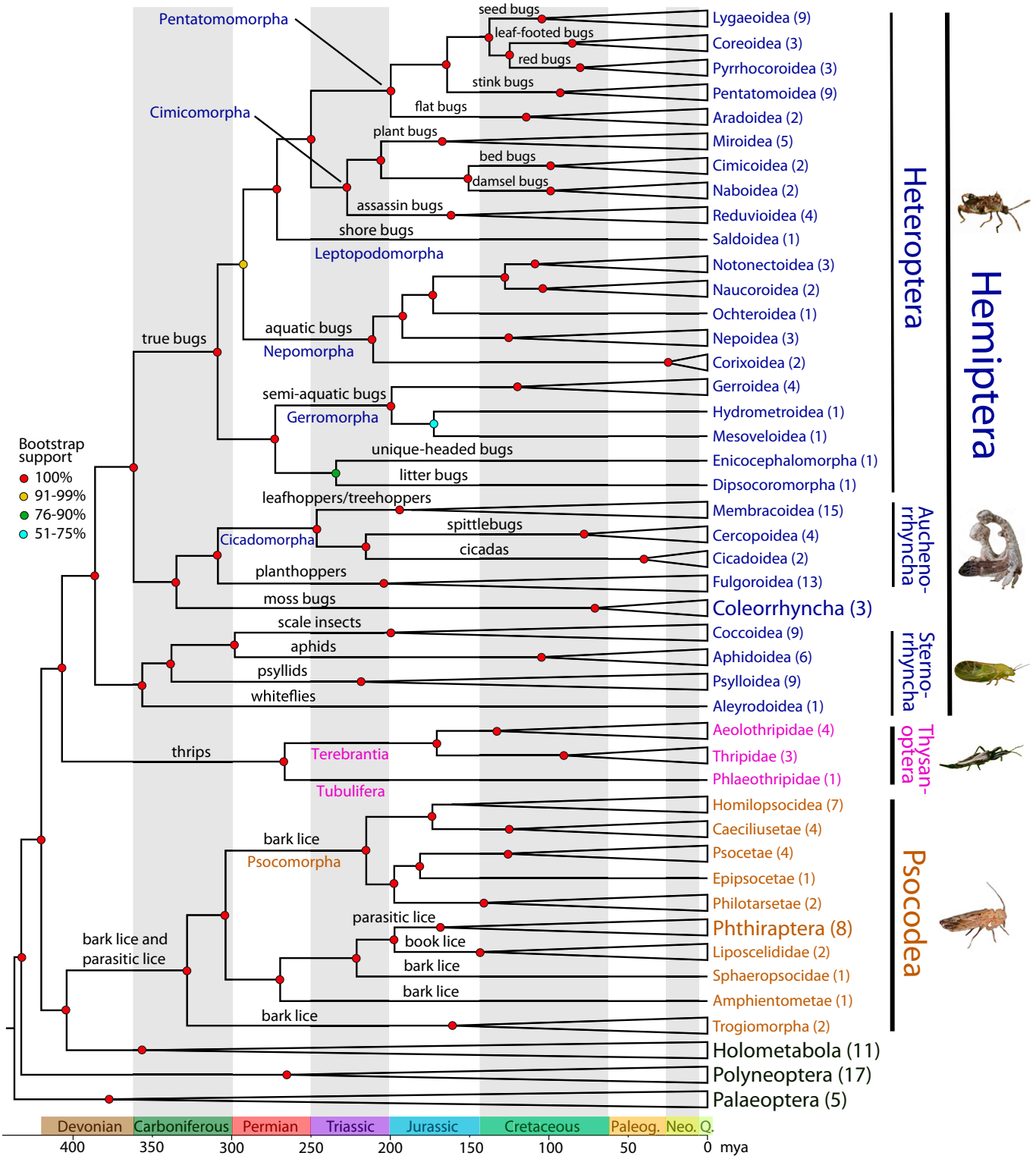

Fig. 1. Dated phylogeny of hemipteroid insects (Hemiptera, Thysanoptera, and Psocodea) based on maximum likelihood analysis of a supermatrix of first and second codon position nucleotides corresponding to 859,518 aligned amino acid positions from transcriptome or genome sequences of 193 samples. Colored circles indicate bootstrap support. Timescale in millions of years (*Bottom*) estimated from MCMCTree Bayesian divergence time analyses using 23 fossil calibration points and a reduced dataset. Number of species sampled from each group indicated in parentheses. Higher taxa are indicated as taxon labels and below branches; most convenient generalized common names are above branches. Images represent five major groups: Heteroptera, Auchenorrhyncha, Sternorrhyncha, Thysanoptera, and Psocodea.

Within Sternorrhyncha, whiteflies (Aleyrodoidea) were sister to the remainder of the suborder, and psyllids (Psylloidea) were sister to a clade composed of aphids (Aphidoidea) + scale insects (Coccoidea), also supported by 91% of quartets in FcLM analyses. Previous phylogenetic analyses of Sternorrhyncha have tended to focus within particular superfamilies or families (e.g., refs. 22–24)

rather than addressing relationships among major lineages (superfamilies).

The earliest molecular phylogenetic analyses of Hemiptera (e.g., refs. 25 and 26) failed to recover Auchenorrhyncha as a monophyletic group, as has a more recent analysis of mitochondrial genomes (15). However, our analyses provided strong support for monophyly of this group, corroborating results of other studies based on multiple loci (13, 14). Within Auchenorrhyncha, our results strongly support the taxonomic status of the two recognized infraorders Fulgoromorpha (i.e., Fulgoroidea, planthoppers) and Cicadomorpha (leafhoppers/treehoppers, spittlebugs, and cicadas) as monophyletic, as found previously (13). However, relationships among the three superfamilies of Cicadomorpha were inconsistently resolved. Cicadas (Cicadoidea) plus spittlebugs (Cercopoidea) were sister to leafhoppers/treehoppers (Membracoidea) in the analysis of nucleotide sequences (Fig. 1, FcLM 52% of quartets), but cicadas were sister to spittlebugs plus leafhoppers/treehoppers in the analysis of amino acid sequence data (*SI Appendix*, Fig. S1), which was also found in 48% of quartets of nucleotide data in FcLM analyses.

Relationships among the earlier diverging lineages of true bugs (Heteroptera) have not been resolved consistently across previous analyses (14, 27–29), in which the deepest divergences received low statistical branch support and recovered different relationships among infraorders. In our analysis, which included representatives of all seven currently recognized infraorders, the four infraorders for which more than one species was included were found to be monophyletic. Like two recent studies based on combined molecular and morphological data (29) and transcriptome data (14), we found 100% bootstrap support for (*i*) a clade comprising litter bugs (Dipsocoromorpha), unique-headed bugs (Enicocephalomorpha), and semiaquatic bugs (Gerromorpha) (also found in 100% of quartets in FcLM analyses) and (*ii*) shore bugs (Leptopodomorpha) as the sister to Cimicomorpha + Pentatomomorpha (also found in 100% of quartets in FcLM analyses).

Divergence Time Analysis. The estimate of the root age for our tree, the split between Paleoptera (dragonflies, damselflies, and mayflies) and Neoptera (all other insects) at 437 million years ago (mya) (95% CI 401–486) was only slightly older than that estimated for this node by Misof et al. (1), at 406 mya. Divergence dates for more interior nodes tended to be older than those estimated by Misof et al. (1) and more similar to those of Tong et al. (30), possibly due either to much denser sampling of minimum age fossil calibration points throughout this part of the insect tree or to different methodology (e.g., MCMCtree versus BEAST or different prior distributions of expected ages for Bayesian analyses). Analyses of divergence times postulated a common ancestor of thrips and hemipterans as early as the Devonian (~407 mya, 95% CI 373–451). Radiation within Hemiptera is also inferred to have begun in this period (~386 mya, 95% CI 354–427), with radiations within Sternorrhyncha, Auchenorrhyncha, and Heteroptera having commenced by the late Carboniferous (all before 300 mya). Radiation within modern Psocodea dates to the Carboniferous (328 mya, 95% CI 292–376), with divergence of this lineage from other insects as early as 404 mya (95% CI 367–451).

Discussion

Analysis of 2,395 protein-coding, single-copy genes derived from transcriptomes of hemipteroid insects and outgroups provided strong support for a backbone tree of hemipteroid insects largely congruent with previous analyses and classification schemes. In particular, we recovered with strong support monophyly of the three orders of hemipteroid insects: Psocodea, Thysanoptera, and Hemiptera. We also recovered monophyly of most currently recognized suborders, infraorders, and superfamilies within these groups as well as resolving relationships among these major groups. Although the unconventional result of a sister relationship between Psocodea and Holometabola of Misof et al. (1) appeared to be robust to our substantially increased taxon sampling based on maximum likelihood bootstrapping, it was not supported by four-cluster likelihood mapping analyses. FcLM, which can detect potentially confounding signal, suggests extensive underlying conflict for this result, with the majority of quartets placing Psocodea with thrips and hemipterans, which would imply monophyly of Paraneoptera in rooted trees. However, permutations appear to rule out several possible types of confounding signal (e.g., among-lineage heterogeneity or nonrandom distribution of missing data) in our dataset. Recent work has suggested that bootstrap support from very large datasets may provide an overestimate of confidence for phylogenetic results (31–33). Thus, the position of Psocodea in the insect tree is still an open question. Monophyly of hemipteroid insects is supported by several morphological autapomorphies (34); therefore, nonmonophyly of the group would imply homoplasy in these traits. In addition, there is no known morphological apomorphy supporting Psocodea + Holometabola (*SI Appendix*, Fig. S5). In contrast, the other less conventional relationship, a clade comprising Coleorrhyncha and Auchenorrhyncha uncovered by Misof et al. (1), was recovered by our trees with increased taxon sampling and is supported by 96% of quartets in the FcLM analyses and three morphological apomorphies, suggesting that this result is robust.

Divergence time estimates using a dense sampling of 23 fossil calibration points suggest that the radiation of the hemipteroid insect orders is relatively ancient, beginning before the early Carboniferous, considerably older than initial expectations based on available fossils. However, the insect fossil record of this period is extremely fragmentary, and relatively old fossils of modern lineages that are used as calibration points imply that branches uniting these lineages must be older still, given that fossil ages represent minimum ages.

Implications for Evolution of Feeding Strategy. Our phylogenetic results generally agree with evidence from the fossil record that the earliest hemipteroids fed on detritus, pollen, fungi, or spores (as in most modern bark lice and thrips). Plant-fluid feeding probably coincided with the origin of Hemiptera and was independently derived in thrips. Today, Hemiptera is the fifth largest insect order, surpassed only by the four major holometabolous orders (Hymenoptera, Coleoptera, Lepidoptera, and Diptera). It remains one of the most abundant and diverse groups of plant-feeding insects. Within Hemiptera, the origin of true bugs apparently coincided with a shift from herbivory to predation, with subsequent shifts back to herbivory (29, 35) in the more derived lineages (Pentatomomorpha and Cimicomorpha). The two other large suborders of Hemiptera (Auchenorrhyncha and Sternorrhyncha) feed almost exclusively on vascular plant fluids.

Our results also suggest that the earliest hemipterans fed preferentially on phloem. Phloem feeding remains predominant in extant plant-feeding hemipterans, including nearly all Sternorrhyncha and most Auchenorrhyncha (36), while modern moss bugs feed on phloem-like tissues in mosses (37). A shift to xylem feeding appears to have coincided with the origin of Cicadomorpha (at least the crown group of this lineage), in which all cicadas and spittlebugs retain this preference. This is also supported by the fossil record in which the earliest leafhoppers had inflated faces (38), indicating a preference for xylem feeding, despite the predominance of phloem feeding among modern leafhoppers and treehoppers (Membracoidea). A shift to phloem feeding apparently occurred early in the evolution of Membracoidea but at least one reversal to xylem feeding [in Cicadellinae (sharpshooters)] has been inferred previously (39), consistent with our results.

Implications for Morphological Evolution. Based on the conflicting statistical support between the supermatrix analysis and four-cluster likelihood mapping, the position of lice (Psocodea) appears to be unstable. Morphological evidence, in contrast, supports the monophyly of hemipteroid insects (Paraneoptera). Our parsimony mapping of 142 morphological characters (*SI Appendix*, Fig. S5) found no apomorphies supporting Psocodea + Holometabola but 14 apomorphies supporting hemipteroid insect monophyly.

Some of these are reductions or losses, including the reduced number of tarsomeres (three in modern hemipteroids), reduced number of Malpighian tubules (four), and presence of only one abdominal ganglionic complex. Nevertheless, these characters, together with characters of the forewing base, still appear to support the sister group relationship between Psocodea and thrips plus hemipterans (11, 34, 40). Thus, the phylogenetic position of Psocodea requires further study of morphological and molecular data.

In contrast to the equivocal support for Paraneoptera, Condylognatha is strongly supported not only in the phylogenomic analyses, but also with six morphological apomorphies. The origin of this group apparently coincided with a distinct shift in mouthpart morphology and feeding habits toward piercing and sucking. These changes include anterior shifting of tentorial pits, elongated and slender mandibles, stylet-like laciniae, and a narrowed labium (*SI Appendix*, Fig. S5). Subsequent evolutionary transformations led to the very distinct and unique piercing–sucking mouthparts of hemipterans that facilitate ingestion of liquid from plant or animal tissues.

The sister-group relationship that we found between moss bugs (Coleorrhyncha) and Auchenorrhyncha has not, to our knowledge, been proposed previously in any explicit phylogenetic analysis other than in recent phylogenomic analyses of transcriptomes (1, 14). Traditionally, moss bugs were treated as one of three suborders of "Homoptera" (along with Sternorrhyncha and Auchenorrhyncha), largely based on the structure of the head. The mouthparts of moss bugs arise posteroventrally (41), as in leafhoppers and relatives, rather than anteriorly as in true bugs (42). Nevertheless, morphological evidence from fossil and living moss bugs, primarily from wing structure and musculature, suggested a closer relationship to true bugs (9, 41, 43). However, a recent comparative morphological study (12) revealed that moss bugs share a unique derived feature of the wing base with Auchenorrhyncha; a membranous proximal median plate. The same study also showed that some previously suggested morphological synapomorphies of moss bugs and true bugs (*SI Appendix*, Fig. S5*C*) are either ambiguous or have been misinterpreted (12). Prior molecular evidence supporting moss bugs plus true bugs was also somewhat equivocal [ref. 13: maximum likelihood (ML) bootstrap 83% and maximum parsimony (MP) bootstrap 63%]. Our results support those of other transcriptome studies (1, 14) in placing Coleorrhyncha sister to Auchenorrhyncha.

Implications for Evolution of Mitochondrial Genome Organization. Several groups of hemipteroid insects have been shown to have highly rearranged mitochondrial genomes (2). The sister relationship between thrips and hemipterans indicates that the heightened rates of mitochondrial (mt) genome rearrangements observed in the lice (44) and thrips (45) are the result of convergence between these two clades. Even if Psocodea is sister to thrips plus hemipterans, and not to holometabolous insects, recent analyses indicating that the ancestor of all Psocodea had a generally standard insect mitochondrial gene order still result in an interpretation involving convergence (46). This phylogenetic evidence is also consistent with the absence of any shared, derived gene arrangements between Psocodea and thrips, as both have independently diverged from the inferred ancestral insect mt genome arrangement (2, 45).

An interpretation involving convergence is also consistent with the varying degrees of rearrangement observed within each order. Within Psocodea, mt genomes vary wildly across different taxonomic scales, from a single derived arrangement found in all Psocomorpha (46), to wide variation within a single genus (*Liposcelis*, ref. 47), and between closely related species of parasitic lice. In contrast, for the thrips, mitochondrial genome arrangements are relatively consistent at the family level (with only tRNA rearrangements observed), albeit still highly rearranged relative to the ancestral insect mt genome (48). Very few rearrangements of any type are observed in the Hemiptera, with the vast majority of families possessing the inferred ancestral arrangement (2).

In summary, although the exact phylogenetic position of Psocodea remains to be resolved convincingly, our results based on transcriptomes for hemipteroid insects provide a strong phylogenetic framework for future studies of genomic, morphological, ecological, and behavioral characteristics of this important group of insects.

Materials and Methods

Our general approach closely followed methods described previously by Misof et al. (1) and Peters et al. (49) for phylogenomic analyses of insect transcriptomes (*SI Appendix*, Dryad repository, 10.5061/dryad.t4f4g85). Transcriptomes of 140 samples of Paraneoptera were newly sequenced with 100 bp paired-end reads for this study using Illumina HiSeq2000 or HiSeq2500 machines to achieve at least 2.5 Gbp per taxon. The final taxon sample of 193 includes representatives of 97 hemipteroid families with several larger families represented by multiple subfamilies.

All paired-end reads were assembled with SOAPdenovo-Trans (version 1.01; ref. 50) and the assembled transcripts were filtered for possible contaminants (*SI Appendix*, Table S2) as described in Peters et al. (49). The raw reads and filtered assemblies were submitted to the NCBI SRA and TSA archives (*SI Appendix*, Table S1). We searched the assemblies for transcripts of 2,395 protein-coding genes that the OrthoDB v7 database (51) suggested to be single copy across the genomes of six species (*SI Appendix*, Table S3) using the software Orthograph (version beta4, ref. 52; for results of the orthology search see *SI Appendix*, Table S4). Orthologous transcripts were aligned with MAFFT (version 7.123; ref. 53) at the translational (amino acid) level. Corresponding nucleotide multiple sequence alignments were generated with a modified version of the software Pal2Nal (54) (version 14).

Alignment sections that could not be discriminated from randomly aligned regions at the amino acid level of each gene were identified with Aliscore version 1.2 (55, 56). To maximize the fit of our substitution models, we identified for each gene the protein domains (clans, families) and unannotated regions using the Pfam database (refs. 1 and 57 and *SI Appendix, Supplemental Materials and Methods*). The phylogenetic information content of each data block was assessed with MARE (version 0.1.2-rc) (58), and all uninformative data blocks (IC = 0) were removed. We subsequently used PartitionFinder (developer version 2.0.0-pre14, ref. 59) to simultaneously infer the best partitioning scheme and amino acid or nucleotide (removing third positions because of heterogeneity, *SI Appendix*, Fig. S6) substitution models, using the rclusterf algorithm.

Phylogenetic trees were inferred using a maximum likelihood approach with ExaML version 3.0.17 (60) for both the nucleotide and amino acid datasets. We performed 50 nonparametric bootstrap replicates mapping the support on the best ML tree after checking for bootstrap convergence with the default bootstopping criteria (61). An optimized dataset, which requires the presence of at least one species from a given taxonomic group (*SI Appendix*, Table S5) in each data block of the supermatrix (62), was used for testing the possible impact of missing data at the partition level. Four-cluster likelihood mapping (63) was used for assessing the phylogenetic signal for alternative phylogenetic relationships (*SI Appendix*, Tables S5 and S6). Permutation tests in these analyses assessed the impact of heterogeneous amino acid sequence composition among lineages, nonstationarity of substitution processes, and nonrandom distribution of missing data on the inferred phylogenetic tree (1).

To understand the morphological transformations underlying the evolution of the hemipteroid groups and to identify potential shared derived characters (synapomorphies), we used the morphological data matrix of Friedemann et al. (9) with 118 characters of the entire body (with modifications from ref. 12) and additionally 25 characters associated with the wing base (8). By tracing characters over the tree using maximum parsimony using Winclada (64), we evaluated three possible phylogenetic alternatives: (*i*) paraphyletic Paraneoptera and Coleorrhyncha sister to Auchenorrhyncha (result from ML analysis of transcriptomes); (*ii*) monophyletic Paraneoptera (as suggested by FcLM analyses); and (*iii*) paraphyletic Paraneoptera, but with Coleorrhyncha sister to Heteroptera (as suggested in previous literature).

To estimate divergence dates, we used the topology resulting from ML analysis of first and second position nucleotides as the input tree and assigned 23 ingroup fossil calibration points (65) throughout the tree (*SI Appendix*, Table S7). These calibrations were used as minimum ages in soft bound uniform priors with a root age of 406 mya (1) as a soft bound maximum. These priors were used in a Bayesian MCMCTree (66) molecular dating analysis of a first and second position nucleotide dataset for which sites were present in at least 95% of taxa.

ACKNOWLEDGMENTS. We thank E. Anton, M. Bowser, C. Bramer, T. Catanach, D. H. Clayton, J. R. Cooley, G. Gibbs, A. Hansen, E. Hdez, A. Katz, K. Kjer, J. Light, A. Melber, B. Morris, D. Papura, H. Pohl, R. Rakitov, C. Ray, S. Schneider, K. Schütte, W. Smith, K.-Q. Song, T. Sota, N. Szucsich, G. Taylor, S. Taylor, S. Tiwari, and X. Tong for assistance with obtaining specimens; G. Meng and Beijing Genomics Institute staff for their efforts in data curation; and O. Niehuis

for assistance preparing the ortholog gene set. R.M.W. was supported by Swiss National Science Foundation Grant PP00P3_1706642. K.M. was supported by David Yeates, the Schlinger Endowment, CSIRO National Research Council Australia, J. Korb, and the University of Freiburg. This work was also supported by National Science Foundation DEB-1239788 (to K.P.J., C.H.D., and H.M.R.).

1. Misof B, et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.
2. Cameron SL (2014) Insect mitochondrial genomics: Implications for evolution and phylogeny. *Annu Rev Entomol* 59:95–117.
3. Adis J, Lubin YD, Montgomery GG (1984) Arthropods from the canopy of inundated and terra firme forests near Manaus, Brazil, with critical considerations on the pyrethrum-fogging technique. *Stud Neotrop Fauna Environ* 19:223–236.
4. Schaefer CW, Panizzi AR (2000) *Heteroptera of Economic Importance* (CRC Press, Boca Raton, FL).
5. Ishiwata K, Sasaki G, Ogawa J, Miyata T, Su Z-H (2011) Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. *Mol Phylogenet Evol* 58:169–180.
6. Beutel RG, Gorb SN (2001) Ultrastructure of attachment specializations of hexapods (Arthropoda): Evolutionary patterns inferred from a revised ordinal phylogeny. *J Zool Syst Evol Res* 39:177–207.
7. Wheeler WC, Whiting M, Wheeler QD, Carpenter JM (2001) The phylogeny of the extant hexapod orders. *Cladistics* 17:113–169.
8. Yoshizawa K, Saigusa T (2001) Phylogenetic analysis of paraneopteran orders (Insecta: Neoptera) based on forewing base structure, with comments on monophyly of Auchenorrhyncha (Hemiptera). *Syst Ent* 26:1–13.
9. Friedemann K, Spangenberg R, Yoshizawa K, Beutel RG (2014) Evolution of attachment structure in the highly diverse Acercaria (Hexapoda). *Cladistics* 30:170–201.
10. Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC (1997) The Strepsiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Syst Biol* 46:1–68.
11. Kristensen NP (1991) Phylogeny of extant hexapods. *The Insects of Australia*, ed CSIRO (Melbourne Univ Press, Melbourne), pp 125–140.
12. Yoshizawa K, Ogawa N, Dietrich CH (2017) Wing base structure supports Coleorrhyncha + Auchenorrhyncha (Insecta: Hemiptera). *J Zool Syst Evol Res* 55:199–207.
13. Cryan JR, Urban JM (2012) Higher-level phylogeny of the insect order Hemiptera: Is Auchenorrhyncha really paraphyletic? *Syst Ent* 37:7–21.
14. Wang Y-H, et al. (2017) When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera-Heteroptera. *Cladistics*, in press.
15. Li H, et al. (2017) Mitochondrial phylogenomics of Hemiptera reveals adaptive innovations driving the diversification of true bugs. *Proc Biol Sci* 284:20171223.
16. Yoshizawa K, Johnson KP (2013) Changes in base composition bias of nuclear and mitochondrial genes in lice (Insecta: Psocodea). *Genetica* 141:491–499.
17. Lyal CHC (1985) Phylogeny and classification of the Psocodea, with particular reference to the lice (Psocodea: Phthiraptera). *Syst Ent* 10:145–165.
18. Johnson KP, Yoshizawa K, Smith VS (2004) Multiple origins of parasitism in lice. *Proc Biol Sci* 271:1771–1776.
19. Buckman RS, Mound LA, Whiting MF (2013) Phylogeny of thrips (Insecta: Thysanoptera) based on five molecular loci. *Syst Ent* 38:123–133.
20. Bourgoin T, Campbell BC (2002) Inferring a phylogeny for Hemiptera: Falling into the 'autapomorphic trap'. *Denisia* 4:67–82.
21. Ouvrard D, Campbell BC, Bourgoin T, Chan KL (2000) 18S rRNA secondary structure and phylogenetic position of Peloridiidae (Insecta, Hemiptera). *Mol Phylogenet Evol* 16:403–417.
22. Von Dohlen CD, Moran NA (2000) Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. *Biol J Linn Soc Lond* 71: 689–717.
23. Gullan PJ, Cook LG (2007) Phylogeny and higher classification of the scale insects (Hemiptera: Sternorrhyncha: Coccoidea). *Zootaxa* 1668:413–425.
24. Percy DM, et al. (2018) Resolving the psyllid tree of life: Phylogenomic analyses of the superfamily Psylloidea (Hemiptera). *Syst Ent* 43:762–776.
25. Campbell BC, Steffen-Campbell JD, Sorensen HT, Gill RJ (1995) Paraphyly of Homoptera and Auchenorrhyncha inferred from 18S rDNA nucleotide sequences. *Syst Ent* 20:175–194.
26. von Dohlen CD, Moran NA (1995) Molecular phylogeny of the Homoptera: A paraphyletic taxon. *J Mol Evol* 41:211–223.
27. Li H, et al. (2012) The complete mitochondrial genome and novel gene arrangement of the unique-headed bug *Stenopirates* sp. (Hemiptera: Enicocephalidae). *PLoS One* 7: e29419.
28. Weirauch C, Štys P (2014) Litter bugs exposed: Phylogenetic relationships of Dipsocoromorpha (Hemiptera: Heteroptera) based on molecular data. *Insect Syst Evol* 45: 351–370.
29. Weirauch C, Schuh RT, Cassis G, Wheeler WC (2018) Revisiting habitat and lifestyle transitions in Heteroptera (Insecta: Hemiptera): Insights from combined morphological and molecular phylogeny. *Cladistics*, in press.
30. Tong KJ, Duchêne S, Ho SYW, Lo N (2015) INSECT PHYLOGENOMICS. Comment on "Phylogenomics resolves the timing and pattern of insect evolution". *Science* 349: 487.
31. Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.
32. Kobert K, Salichos L, Rokas A, Stamatakis A (2016) Computing the internode certainty and related measures from partial gene trees. *Mol Biol Evol* 33:1606–1617.
33. Shen X-X, Hittinger CT, Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecol Evol* 1:126.
34. Yoshizawa K, Lienhard C (2016) Bridging the gap between chewing and sucking in the hemipteroid insects: New insights from Cretaceous amber. *Zootaxa* 4079:229–245.
35. Cobben RH (1979) On the original feeding habits of the Hemiptera (Insecta): A reply to Merrill sweet. *Ann Entomol Soc Am* 72:711–715.
36. Backus EA (1988) Sensory systems and behaviors which mediate hemipteran plant-feeding: A taxonomic overview. *J Insect Physiol* 34:151–165.
37. Cronk QCB, Forest F (2017) The evolution of Angiosperm trees: From Palaeobotany to genomics. *Comparative and Evolutionary Genomics of Angiosperm Trees*, eds Groover A, Cronk Q (Springer, New York), pp 1–17.
38. Shcherbakov D (1996) Origin and evolution of the Auchenorrhyncha as shown by the fossil record. *Studies on Hemipteran Phylogeny*, ed Schaeffer CW (Entomol Soc Am, Lanham, MD).
39. Dietrich CH, et al. (2017) Anchored hybrid enrichment-based phylogenomics of leafhoppers and treehoppers (Hemiptera: Cicadomorpha: Membracoidea). *Insect Syst Diver* 1:57–72.
40. Beutel RG, Friedrich F, Ge S-Q, Yang X-K (2014) *Insect Morphology and Phylogeny: A Textbook for Students of Entomology* (Walter de Gruyter, Berlin).
41. Spangenberg R, et al. (2013) The cephalic morphology of the Gondwanan key taxon *Hackeriella* (Coleorrhyncha, Hemiptera). *Arthropod Struct Dev* 42:315–337.
42. Spangenberg R, Friedemann K, Weirauch C, Beutel RG (2013) The head morphology of the potentially basal heteropteran lineages Enicocephalomorpha and Dipsocoromorpha (Insecta: Hemiptera: Heteroptera). *Arthropod Syst Phyl* 71:103–136.
43. Shcherbakov D, Popov YA (2002) Superorder Cimicidea Laicharting, 1781, Order Hemiptera Linné, 1758. The bugs, cicadas, plantlice, scale insects, etc. *History of Insects*, eds Rasnitsyn AP, Quicke DLJ (Kluwer, Dordrecht, The Netherlands), pp 143–157.
44. Cameron SL, Yoshizawa K, Mizukoshi A, Whiting MF, Johnson KP (2011) Mitochondrial genome deletions and minicircles are common in lice (Insecta: Phthiraptera). *BMC Genomics* 12:394.
45. Dickey AM, et al. (2015) A novel mitochondrial genome architecture in thrips (Insecta: Thysanoptera): Extreme size asymmetry among chromosomes and possible recent control region duplication. *BMC Genomics* 16:439.
46. Yoshizawa K, et al. (2018) Mitochondrial phylogenomics and genome rearrangements in the barklice (Insecta: Psocodea). *Mol Phylogenet Evol* 119:118–127.
47. Shi Y, et al. (2016) The mitochondrial genome of booklouse, *Liposcelis sculptilis* (Psocoptera: Liposcelididae) and the evolutionary timescale of *Liposcelis*. *Sci Rep* 6: 30660.
48. Yan D, et al. (2014) The mitochondrial genome of *Frankliniella intonsa*: Insights into the evolution of mitochondrial genomes at lower taxonomic levels in Thysanoptera. *Genomics* 104:306–312.
49. Peters RS, et al. (2017) Evolutionary history of the Hymenoptera. *Curr Biol* 27: 1013–1018.
50. Xie Y, et al. (2014) SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-seq reads. *Bioinformatics* 30:1660–1666.
51. Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM (2008) OrthoDB: The hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* 36:D271–D275.
52. Petersen M, et al. (2017) Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* 18:111.
53. Katoh K, Standley DM (2016) A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 32:1933–1942.
54. Suyama M, Torrents D, Bork P (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34: W609–W612.
55. Misof B, Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst Biol* 58: 21–34.
56. Kück P, et al. (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool* 7:10.
57. Finn RD, et al. (2014) Pfam: The protein families database. *Nucleic Acids Res* 42: D222–D230.
58. Misof B, et al. (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics* 14:348.
59. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol* 14:82.
60. Kozlov AM, Aberer AJ, Stamatakis A (2015) ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31:2577–2579.
61. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A (2010) How many bootstrap replicates are necessary? *J Comput Biol* 17:337–354.
62. Dell'Ampio E, et al. (2014) Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects. *Mol Biol Evol* 31: 239–249.
63. Strimmer K, von Haeseler A (1997) Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci USA* 94:6815–6819.
64. Nixon K (2002) Winclada (Ithaca, NY), version 1.00. 08. Available at http://www.diversityoflife.org/winclada/. Accessed January 24, 2018.
65. Parham JF, et al. (2012) Best practices for justifying fossil calibrations. *Syst Biol* 61: 346–359.
66. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

Molecular Phylogenetics and Evolution xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids

Jessica P. Gillung[a,b,*], Shaun L. Winterton[b], Keith M. Bayless[c], Ziad Khouri[a], Marek L. Borowiec[d], David Yeates[e], Lynn S. Kimsey[a], Bernhard Misof[f], Seunggwan Shin[g], Xin Zhou[h], Christoph Mayer[f], Malte Petersen[f], Brian M. Wiegmann[i]

[a] Bohart Museum of Entomology, University of California, One Shields Ave, Davis, CA 95616, USA
[b] California State Collection of Arthropods, 3294 Meadowview Rd, Sacramento, CA 95832, USA
[c] California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118, USA
[d] School of Life Sciences, Social Insect Research Group, Arizona State University, Tempe, AZ, 85287, USA
[e] National Research Collections Australia, Clunies Ross Street, Acton, ACT 2601, GPO Box 1700, Canberra, ACT 2601, Australia
[f] Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany
[g] Department of Biological Sciences, University of Memphis, 3700 Walker Avenue, Memphis, TN 38152, USA
[h] Department of Entomology, China Agricultural University, Beijing 100193, China
[i] Department of Entomology & Plant Pathology, North Carolina State University, 3114 Gardner Hall, Raleigh, NC 27695-7613, USA

ARTICLE INFO

Keywords:
Bayesian inference
Bioinformatics
Conflict
Diptera
Fossilized birth-death process
Systematic error

ABSTRACT

The onset of phylogenomics has contributed to the resolution of numerous challenging evolutionary questions while offering new perspectives regarding biodiversity. However, in some instances, analyses of large genomic datasets can also result in conflicting estimates of phylogeny. Here, we present the first phylogenomic scale study of a dipteran parasitoid family, built upon anchored hybrid enrichment and transcriptomic data of 240 loci of 43 ingroup acrocerid taxa. A new hypothesis for the timing of spider fly evolution is proposed, wielding recent advances in divergence time dating, including the fossilized birth-death process to show that the origin of Acroceridae is younger than previously proposed. To test the robustness of our phylogenetic inferences, we analyzed our datasets using different phylogenetic estimation criteria, including supermatrix and coalescent-based approaches, maximum-likelihood and Bayesian methods, combined with other approaches such as permutations of the data, homogeneous versus heterogeneous models, and alternative data and taxon sets. Resulting topologies based on amino acids and nucleotides are both strongly supported but critically discordant, primarily in terms of the monophyly of Panopinae. Conflict was not resolved by controlling for compositional heterogeneity and saturation in third codon positions, which highlights the need for a better understanding of how different biases affect different data sources. In our study, results based on nucleotides were both more robust to alterations of the data and different analytical methods and more compatible with our current understanding of acrocerid morphology and patterns of host usage.

1. Introduction

The size of molecular datasets in phylogenetics has been growing greatly since the introduction of high-throughput sequencing. The combination of the advances in genomic data acquisition with new bioinformatics tools has resulted in a novel field of evolutionary biology, phylogenomics. The onset of phylogenomics has resolved some of the most challenging evolutionary questions while giving us a new perspective on biodiversity (e.g., Misof et al., 2014; Prum et al., 2015; Garrison et al., 2016; Hamilton et al., 2016; Kocot et al., 2016; Branstetter et al., 2017; Shin et al., 2017; Espeland et al., 2018; Winterton et al., 2018).

Increasing the quantity of phylogenomic data successfully alleviates stochastic error caused by limited data sampling, but the impact of systematic error is potentially augmented (Yeates et al., 2016). Several sources of systematic error have been identified, including compositional heterogeneity, missing data, heterogeneity in evolutionary rates among lineages, among others (Felsenstein, 1978; Jermiin et al., 2004; Bininda-Emonds 2007; Lartillot et al., 2007; Edwards, 2009; Nabholz et al., 2011; Roure et al., 2013; Mirarab et al., 2014; Goremykin et al.,

* Corresponding author at: Bohart Museum of Entomology, University of California, One Shields Ave, Davis, CA 95616, USA.
E-mail address: jpgillung@ucdavis.edu (J.P. Gillung).

https://doi.org/10.1016/j.ympev.2018.08.007
Received 5 April 2018; Received in revised form 3 August 2018; Accepted 7 August 2018

2015; Streicher et al., 2016). Thus, merely increasing the number of gene sequences in datasets does not necessarily resolve all phylogenetic incongruence. Instead, a number of cases have been observed in which alternative phylogenomic datasets strongly support conflicting conclusions, each with highly resolved phylogenetic estimates and maximal nodal support values (e.g., Crawford et al., 2012; Shaffer et al., 2013; Wang et al., 2013; Jarvis et al., 2014; Chang et al., 2015; Pisani et al., 2015; Prum et al., 2015). Phylogenetic conflict, however, can originate not only from different datasets, but also from alternative coding of the same data (Fučíková et al., 2016). Protein-coding genes can be analyzed as amino acids, nucleotides or codons, and choosing which data type to analyze in phylogenomics is a challenge that could significantly affect reliability and confidence of the results.

In the case of phylogenetic studies that focus on recent divergences, nucleotides are probably more informative than amino acids. This is because substitutions are more likely to have occurred at synonymous sites. For deep divergences, however, the choice is not as straightforward. Even though analyses of amino acid datasets are suggested to be less prone to systematic error due to compositional heterogeneity across sites and taxa (Jeffroy et al., 2006; Rodríguez-Ezpeleta et al., 2007; Rota-Stabelli et al., 2013), the statistical phylogenetic analysis of amino acid data presents challenges beyond those often faced with the analysis of DNA sequences. Most approaches to the analysis of amino acid datasets make use of empirical amino acid models, in which all of the potentially free parameters are fixed to specific values estimated from a large number of sequences (Dayhoff et al., 1978; Henikoff and Henikoff, 1992; Jones et al., 1992; Adachi and Hasegawa, 1996; Cao et al., 1998; Adachi et al., 2000; Whelan and Goldman, 2001; Le et al., 2012). Although the fixed amino acid models succeed in reducing the number of free parameters to be estimated, it is possible that even the best-fitting fixed amino acid model is not particularly appropriate for the data at hand. Consequently, if the model is misspecified, the phylogeny estimate might be inaccurate, potentially resulting in conflicting estimates of phylogeny under either nucleotides or amino acids. Conflict among topologies due to alternative data coding as nucleotides or amino acids is relatively common in phylogenomics (Zwick et al., 2012; Rota-Stabelli et al., 2013; Cox et al., 2014; Reddy et al., 2017; Shin et al., 2017; Haddad et al., 2018), but our knowledge of systematic error in big-data phylogenetics is still incipient.

Here, we attempt to understand the basis for the incongruence among phylogenomic trees originating from alternative data types (nucleotides versus amino acids) by investigating the evolution of spider flies (Acroceridae), the only family of flies that exclusively parasitize spiders. Acroceridae is a relatively ancient and morphologically derived lineage of lower Brachycera, consisting of a charismatic and remarkably diverse assemblage of insects. Spider fly origins have been estimated in the Early Mesozoic (173–221 MYA) (Winterton et al., 2007), but their fossil record extends only to the Upper Jurassic (~150 MYA) (Gillung and Winterton, 2017). Species of Acroceridae attack spiders in 26 families (Cady et al., 1993; Gillung and Borkent, 2017) and are currently distributed in 55 genera and approximately 530 species (Winterton et al., 2007; Schlinger et al., 2013). Three subfamilies are recognized, Acrocerinae, Panopinae and Philopotinae. Monophyly of Philopotinae is based on a series of morphological characters, while Panopinae is defined based on their unique mygalomorph attacking behaviour. The monophyly of Acrocerinae is contentious, and its internal relationships are poorly known (Winterton et al., 2007). Thus, additional data and analyses are needed to test the monophyly of the subfamilies and to establish a robust higher-level classification.

In this study, we address the two-fold problem of data type choice and Acroceridae relationships, bringing the greatly expanded gene sampling of anchored phylogenomics to bear on spider fly phylogeny. We recovered 240 unique orthologous loci of 43 species representing all major lineages of spider flies, plus seven representatives of outgroup families. Through the integration of high-throughput sequencing and comparative methods, we provide a robust hypothesis for the pattern and timing of spider fly evolution. Using Acroceridae as a system, we explore the potential of genomic data to resolve relationships in relatively ancient radiations and explore the effects of potential confounding factors in phylogenomic reconstruction.

2. Material and methods

2.1. Taxon sampling and DNA acquisition

Taxa were carefully selected to represent the greatest diversity within Acroceridae and to ensure as close to proportional sampling as possible, based on ongoing taxonomic studies (Gillung and Winterton, 2011; Winterton and Gillung, 2012; Schlinger et al., 2013; Borkent et al., 2016; Gillung and Nihei, 2016). Newly generated Anchored Hybrid Enrichment (AHE) data for 42 species of Acroceridae plus the transcriptome of one additional spider fly species were included as the ingroup. Transcriptomes of six species and AHE data of one species in the lower Brachycera were used as outgroup taxa, representing the families Asilidae, Bombyliidae, Hilarimorphidae, Nemestrinidae, Pantophthalmidae, Tabanidae and Xylophagidae (Supplementary Table 4). Genetic material was extracted from the legs and thorax, with genitalia, remaining legs, head and wings preserved in 95% ethanol as vouchers (Supplementary Table 4). DNA was extracted from frozen specimens preserved in 95% ethanol using a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA). RNA was extracted from specimens preserved in RNA-later following methods described in Misof et al. (2014) and Peters et al. (2017). AHE capture was carried out following the general methods of Lemmon et al. (2012) for sonication, library preparation, indexing and enrichment. Probes were developed specifically for Diptera at the Center for Anchored Phylogenomics at Florida State University, as described in Young et al. (2016). The AHE Diptera Probe Set targets 559 loci, with sequences publicly available as supplementary information from Young et al. (2016). AHE data was sequenced as single reads, with up to 48 multiplexed samples per lane on an Illumina MiSeq platform at the NCSU Genomic Sciences Laboratory (Raleigh, NC). Transcriptome libraries were prepared following methods described by Misof et al. (2014) and Peters et al. (2017). Reads were inspected for quality with Fastqc (Andrews, 2010) and trimmed with Trimmomatic (Bolger et al., 2014), with minimum per base sequence quality set to 20, and minimum read length set to 25 bp.

2.2. Sequence assembly and orthology prediction

De novo assemblies were carried out using Trinity v. 2.2 (Grabherr et al., 2011). For data provided by 1KITE, raw reads were quality checked, assembled with SOAPdenovo-Trans-31kmer (version 1.01) (Xie et al., 2014) and cleaned from potential contaminants as described by Peters et al. (2017). We used Orthograph v.0.5.8 (Petersen et al., 2017) to infer orthology of sequence contigs, with single copy genes extracted and assembled from OrthoDB5 (Waterhouse et al., 2013) and reciprocal search set to relaxed. Orthologous genes were identified based on an ortholog reference set of 3288 orthologous clusters of sequences groups (single copy protein-coding genes) from five reference species: *Anopheles gambiae* Giles, *Tribolium castaneum* (Herbst), *Drosophila melanogaster* Meigen, *Mayetiola destructor* (Say) and *Bombyx mori* (Linnaeus) (Kutty et al., 2018). Following orthology prediction, contaminant viral, bacterial and fungal sequences were identified using NCBI BLAST; loci not matching Diptera or other insects were removed.

2.3. Dataset construction

Internal stop codons and “U” (Selenocysteine) were replaced with an “X” in the amino acid dataset and with “NNN” on the nucleotide dataset, respectively. Amino acid sequences were aligned using MAFFT v.7.123b (Katoh and Standley, 2013) with the *L-INS-i* algorithm.

Ambiguously or randomly aligned sections identified by Aliscore v2.2 (Misof and Misof, 2009; Kück et al., 2010) were removed from the amino acid alignment, and the corresponding codons from the nucleotide loci were removed using Alicut and custom Perl scripts (Misof and Misof, 2009; Kück et al., 2010). Nucleotide sequences were then aligned using the amino acid alignment as blueprint in Pal2Nal (Suyama et al., 2006), using a slightly modified version (see Misof et al., 2014). Individual loci were concatenated using AMAS (Borowiec, 2016). We combined transcriptomic and AHE data for all 50 species included in this study, which resulted in a dataset containing 3234 genes. Because many of these genes were present only in the taxa represented by transcriptomes, which leads to non-random distribution of missing data as these were mainly outgroups, we filtered loci based on taxon occupancy, keeping only the loci present in at least 24 taxa (out of 50). The final nucleotide and amino acid datasets contained 240 loci, with 172,905 base pairs and 57,635 amino acid sites, respectively (Supplementary Files 1–4).

2.4. Dataset exploration

Pairwise sequence comparisons using Bowker's matched-pairs tests of symmetry (Bowker, 1948) were performed in SymTest version 2.0.47 (https://github.com/ottmi/symtest) (Jermiin et al., 2004; Ababneh et al., 2006). The software was also used to generate heat maps based on the inferred p-values, using default window and step sizes. We applied Bowker's test as implemented in SymTest on the amino acid dataset, and on the nucleotide dataset with and without 3rd codon positions.

2.5. Phylogenetic analyses

Both supermatrix and species tree approaches were used for tree estimation on both amino acid and nucleotide datasets. We performed multiple alternative rooting strategies to account for uncertainty in the placement of Acroceridae within the lower Brachycera (e.g., Wiegmann et al., 2011; Shin et al., 2018). Alternative rooting along branches of the outgroup did not affect the relationships within Acroceridae (results not shown), thus we arbitrarily constrained Pantophthalmidae as the root in the topologies presented here. Multispecies coalescent analyses (MSC) were performed using ASTRAL v4.9.7 (Mirarab and Warnow, 2015), with gene trees estimated using RAxML v.8.2.10 (Stamatakis, 2014), and branch support values calculated using 500 bootstrap replicates from RAxML. For the concatenated analyses, alignments were initially partitioned by genes, which were then grouped into meta-partitions using PartitionFinder 2 (Lanfear et al., 2016), with the *rcluster* search algorithm (Lanfear et al., 2014) and BIC for model selection. For the nucleotide model selection analyses, we did not include the GTR + I + G mixture model because this approach has been demonstrated to result in undesirable interactions among parameters (Yang, 1993, 1996, 2006; Sullivan et al., 1999; Mayrose et al., 2005; Jia et al., 2014). Model selection for the amino acid dataset was performed including all models available in PartitionFinder 2, using the *–raxml* option. The best fitting model was selected using BIC. Basic alignment statistics, including percentage of missing data, A/T and G/C content, alignment length and proportion of variable sites were obtained using AMAS (Borowiec, 2016). ExaML (Kozlov et al., 2015) was used to estimate phylogenies under Maximum Likelihood (ML), with parsimony starting trees inferred with RaxML v8.2.10. Node support was estimated via slow non-parametric bootstrapping, with 500 bootstrap replicates per dataset generated with RaxML. Ten different ExaML tree searches were performed and compared with each other to ensure that the analyses were not trapped in a local optimum (i.e., the same topology was recovered). Bayesian tree inference (BI) was carried out by running four independent replicates, with four chains each, using either ExaBayes v1.4 (Aberer et al., 2014) or MrBayes (Ronquist and Huelsenbeck, 2003) through the Cipres Science Gateway v3.3 (Miller et al., 2010). Runs were carried on for at least 50,000,000 generations and were sampled every 1000 generations. Branch lengths were linked among partitions and a relative burn-in of 25% was used. Convergence was evaluated by ensuring effective sample size values (ESS) greater than 200 for each parameter in Tracer v1.6 (Rambaut et al., 2014), as well as potential scale reduction factors (PSRF) ranging close to one and average standard deviations of split frequencies (ASDSF) smaller than 0.01%.

A site-heterogeneous CAT-GTR-G mixture model (Lartillot and Philippe, 2004) was implemented in PhyloBayes (Lartillot et al., 2009). Two independent Markov chains with a total length of 10,000 cycles were run for each analysis, with the first 4000 trees being discarded as burn-in and the posterior consensus determined using the remaining 6000 trees. Convergence between the two chains was assured, with the largest discrepancy observed across all bipartitions (maxdiff) being less than 0.1.

We implemented the *degen1* v1.4 approach (Regier et al., 2010; Zwick et al., 2012) to mask synonymous signal and keep only non-synonymous changes at all coding positions. The degenerated alignment was initially partitioned by locus, and the best fitting substitution model and partition scheme were selected using PartitionFinder 2 as described above. Analysis of the degenerated nucleotide dataset was executed in MrBayes via the Cipres Science Gateway, with four coupled chains and settings as described previously.

2.6. Substitution rate heterogeneity

We used a simplified binning approach as proposed by Mirarab et al. (2014), but grouping genes based on rate of evolution as opposed to bootstrap values on branches as originally proposed by the authors. Gene trees were estimated under ML in RaxML v8.2.10, using the best-fitting model identified by PartitionFinder 2 for each locus both as amino acids and nucleotides. Utilizing the gene_stats R script used in Borowiec et al. (2015), we inferred the average branch lengths, used here as a proxy for rate of evolution, with short branch lengths indicating relatively slowly evolving loci, and long branch lengths indicating relatively faster evolving loci. After sorting genes based on average branch lengths (lowest to highest), we divided the entire set of 240 loci – both as amino acids and nucleotides – into three subsets of 80 loci, so that each subset consisted of a set of loci evolving at roughly under the same rate – namely 'slow', 'intermediate' and 'fast'. We discarded the intermediate population of average branch lengths to ensure two discrete loci populations separated by a large buffer population and concatenated the genes in each of the fast and slow subsets. We then estimated phylogenetic trees separately for each subset using BI in MrBayes 3.2 via the Cipres Science Gateway as described above and assessed their topological congruence with the tree generated from all loci.

2.7. Four-cluster likelihood mapping

We performed four-cluster likelihood mapping (FcLM; Strimmer and von Haeseler, 1997) to quantify the support for the monophyly of Panopinae in the amino acid and nucleotide datasets as implemented in IQTree (Nguyen et al., 2015). We defined four taxon clusters: Panopinae1 (7 species), Panopinae2 (7 species), *Turbopsebius* Schlinger + *Cyrtus* Latreille (2 species), and *Psilodera* Gray + *Pterodontia* Gray (4 species). All remaining species were ignored during analyses. IQTree analyses were conducted using the -m TEST option, which implements ModelFinder (Kalyaanamoorthy et al., 2017) to automatically select the best fitting model, with alignment partitioned by locus.

2.8. Divergence times estimation

The chronogram for Acroceridae was estimated using BEAST v2.4.7

(Bouckaert et al., 2014). We used the nucleotide alignment for the dating analyses and removed six outgroups to enforce proportional sampling of terminal taxa, keeping only *Hilarimorpha* Schiner (Hilarimorphidae) as outgroup. Because the whole nucleotide alignment is too large for a computationally feasible BEAST analysis, we used a method of matrix reduction as implemented in MARE (Misof et al., 2013) using phylogenetic information content as the criterion for keeping or removing genes from the analysis. We applied MARE on the amino acid dataset using the default settings and then reduced the nucleotide dataset accordingly. MARE reduced the nucleotide dataset from 240 to 65 genes, increasing the overall information content of the alignment from 0.31 to 0.45. We applied PartitionFinder 2 using linked branch lengths, the rcluster algorithm and BIC to select the statistically best-fit partitioning scheme and models of nucleotide substitution available in BEAST 2. We used an uncorrelated relaxed molecular clock model (Drummond et al., 2006) and a lognormal prior, with tree and clock model linked across partitions. Fossils included as terminals in the FBD analyses are provided in Supplementary Table 3. Because we did not include morphological data in our analysis to place the fossils in a "total evidence" dating framework *sensu* Ronquist et al. (2012), we assigned them to appropriate groups via monophyly constraints (Heath et al., 2014) according to a recent review of spider fly fossils by Gillung and Winterton (2017). The two Jurassic species of *Archocyrtus* were treated as stem acrocerids, while the Cretaceous-aged *Schlingeromyia minuta* was included within the crown Acroceridae. *Glaesoncodes completinervis* was treated as stem *Ogcodes* based on head and wing venation characters, while *Ogcodes exotica* was included in the crown *Ogcodes*. Finally, *Cyrtinella flavinigra* and *Villalites electrica* were placed in a clade containing *Cyrtus* and *Turbopsebius* also based on head and wing venation characters (Gillung and Winterton, 2017).

We ran the analysis for over 600 million generations with four incrementally heated chains and evaluated MCMC convergence and mixing in Tracer v1.6, ensuring that effective sample sizes (ESS) exceeded 200 for all parameters. We then resampled the phylogenetic trees at a lower frequency in LogCombiner v2.3.1 (BEAST package), with a burn-in of 30%. Finally, we summarized the subsampled trees in a maximum clade credibility tree using TreeAnnotator v2.3.1 (BEAST package), with mean heights as node heights. We further compared the effective prior (under the prior) and posterior distributions (with data included) of all the parameters to ensure that our analyses were not prior-sensitive and that the data were informative for the MCMC analyses (results not shown).

2.9. Data availability

Published AHE and transcriptome raw data for 44 species included herein is available from the NCBI SRA database (Bioprojects PRJNA325838). Transcriptome raw reads for the remaining six species will be available in the near future according to the 1KITE Project timeline (http://www.1kite.org/). Accession numbers for the published data and unique identifiers for the 1KITE unpublished transcriptomic data used here are provided in Supplementary Table 4. Individual loci used in this study can be obtained from the alignment files (Supplementary Files 1–4) in the Zenodo Database (https://doi.org/10.5281/zenodo.1289998) prior to the release of 1KITE transcriptomic raw data.

3. Results

3.1. Incongruence of nucleotide and amino acid-based phylogenies

The analyses of the nucleotide and amino acid datasets under a variety of tree estimation methods and dataset permutations resulted in two well-supported topologies, one based on amino acids and the other based on nucleotides (Fig. 1). The phylogeny based on nucleotides was well supported throughout, with a Bayesian posterior probability (PP) of 1 on each node, and only three nodes with maximum likelihood bootstrap values (BS) lower than 100% (Fig. 1A). The phylogeny based on the concatenated amino acid alignment was less supported overall, with 25% of nodes with PP and BS lower than 1.0 and 100%, respectively, and some of the poorly supported nodes located along the backbone of the tree (Fig. 1B). The multispecies coalescent (MSC) analysis using nucleotides resulted in a topology very similar to the one based on concatenation and was relatively well supported overall (Supplementary Fig. 1). The MSC topology based on amino acids (Supplementary Fig. 1) was congruent with the one based on the concatenated dataset (Fig. 1B), albeit with weak statistical support, with many of the particularly poorly supported nodes placed along the backbone.

The monophyly of Panopinae and its internal relationships represent the most significant difference between the phylogenies based on amino acids and nucleotides. In the topology based on nucleotides, Panopinae was recovered as monophyletic and sister to a clade including representatives of the former Acrocerinae (Fig. 1A). In contrast, the topology based on amino acids recovered a polyphyletic Panopinae, with two lineages formerly included in Acrocerinae nested within the subfamily (Fig. 1B). Establishing the evolutionary history of Panopinae has profound implications in the understanding of Acroceridae host usage. Species of Panopinae are unique among acrocerids in attacking heavy bodied, stout legged spiders in the Mygalomorphae, including tarantulas, trapdoor spiders, funnel-web spiders, among others. All other spider flies attack hosts in the Araneomorphae, such as jumping spiders, wolf spiders, orb-weavers, among many others (Gillung and Borkent, 2017). Assuming that Panopinae is monophyletic would result in a hypothesis for Acroceridae evolution where there was only one invasion of Mygalomorphae, while the alternative hypothesis of non-monophyly would require either two independent origins for the mygalomorph host life history, or the loss of this trait in some lineages (Fig. 1).

3.2. Exploratory analyses

We constructed nine supplementary datasets and used a plethora of additional analyses to explore the origins of the conflict between nucleotides and amino acids, and to indirectly access the reliability of the two alternative topologies. We removed 3rd codon positions to investigate whether heterogeneity in evolutionary rates across codon positions caused any error in our tree estimation based on nucleotides. The topology based on 1st and 2nd positions only was very similar to the one based on the whole nucleotide dataset, with minor differences in the relationships within some genera (Supplementary Fig. 2). Since there were no significant changes in relationships after the removal of 3rd codon positions, we included all codon positions in downstream analysis of nucleotide data to include as much information as possible.

Additionally, we used a CAT-GTR-G mixture model of base substitution (Lartillot and Philippe, 2004), which resulted in a topology that was highly discordant with the nucleotide topology under the homogeneous GTR + G model (Supplementary Fig. 3). The most striking difference was the position of *Ogcodes*, which was recovered as the sister group to Philopotinae under the CAT-GTR-G model (Supplementary Fig. 3). The analysis of amino acid data under the CAT-GTR-G mixture model resulted in a topology that is discordant with all other topologies we recovered based on either nucleotides or amino acids (Supplementary Fig. 3). Similar to the nucleotide topology using the mixture model, *Ogcodes* Latreille was recovered as sister to Philopotinae in the amino acid analysis (Supplementary Fig. 3B).

We also accounted for non-random distribution of missing data by excluding five taxa with low locus coverage. The reduced dataset consisted of 45 taxa (out of 50), which was then analyzed under Bayesian inference. The reduced nucleotide topology (Supplementary Fig. 4) was completely congruent with the topology including all 50 taxa (Fig. 1A), and is well supported overall, with every node having

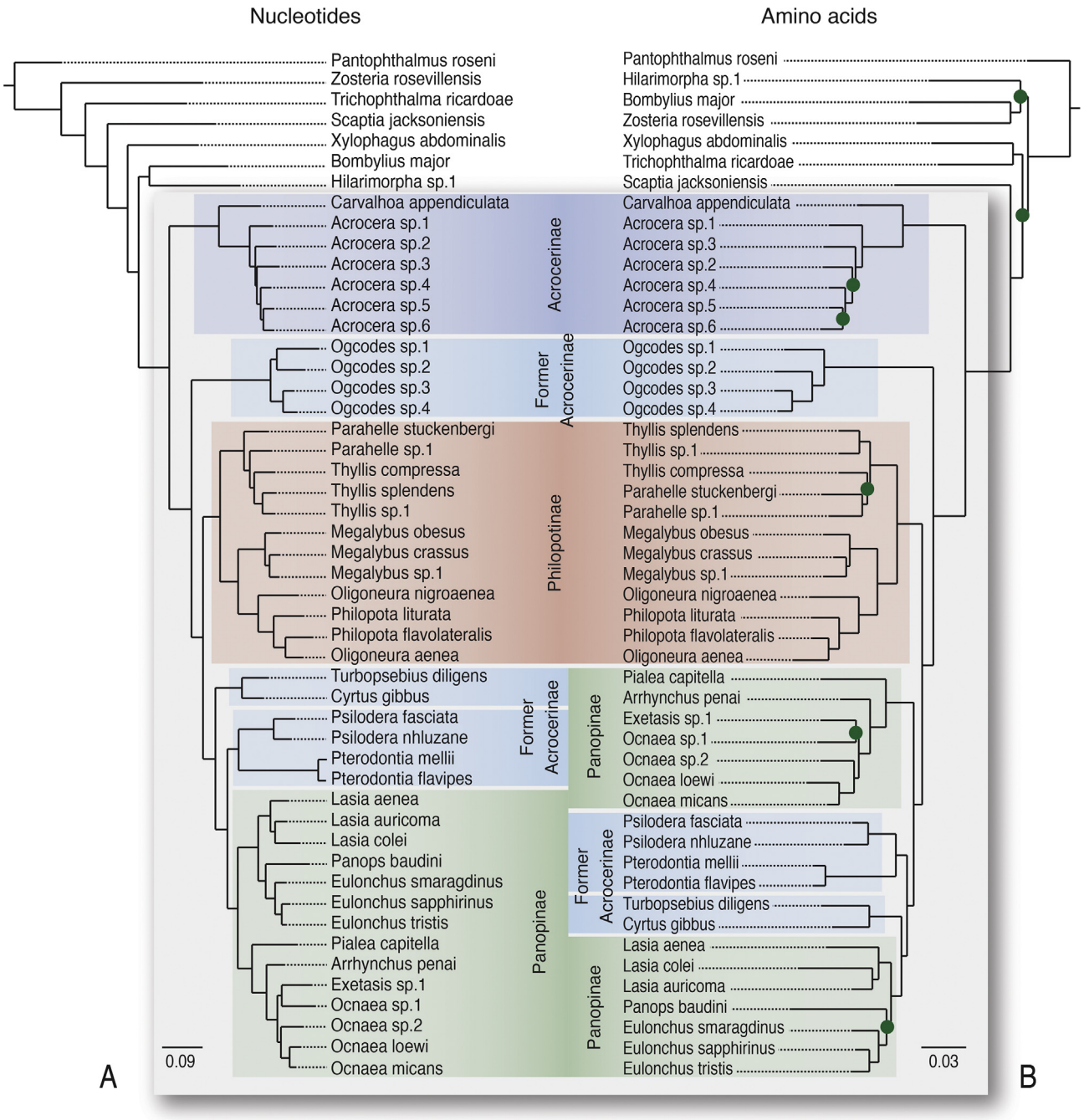

Fig. 1. Phylogeny of spider flies based on the nucleotide (A) and amino acid (B) alignments. Green circles indicate nodes with posterior probability lower than 0.99 and/or bootstrap values lower than 80.

posterior probability (PP) of 1. Conversely, the topology based on the reduced taxon set for amino acids (Supplementary Fig. 4) differs greatly from the topology based on the complete taxon set (Fig. 1B). These results suggest that non-random distribution of missing data had a strong influence on tree estimation using the amino acid dataset, whereas it had no apparent effect on tree estimation based on the nucleotide alignment.

We also investigated the effect of synonymous and non-synonymous information in our analysis. We degenerated nucleotides at codon positions that have the potential to undergo synonymous substitutions using the *degen1* coding approach (Regier et al., 2010; Zwick et al., 2012), and then estimated phylogenies using Bayesian inference. The resulting degenerate nucleotide topology is very similar to the topology based on amino acids, rendering Panopinae polyphyletic, and is relatively well supported overall, with only five nodes with PP lower than 1 (Supplementary Fig. 5). We also applied the *degen1* coding approach to the nucleotide dataset excluding the five taxa with low gene coverage, thus reducing the effect of non-random distribution of missing data. The resulting degenerate nucleotide topology, in turn, was very similar to the amino acid topology with reduced taxon set, rendering Panopinae paraphyletic (Supplementary Fig. 6).

Moreover, we explored the effects of substitution rate heterogeneity across loci using a simplified binning approach as proposed by Mirarab et al. (2014) (Borowiec, 2017; Winterton et al., 2018). We divided the entire set of 240 loci into three subsets of slow-, intermediate- and fast-evolving genes. We then estimated phylogenies under BI separately for the slow- and fast-evolving loci, discarding the intermediate subset of genes. The two topologies based on the nucleotide dataset (for fast- and

slow-evolving loci) are completely congruent with each other and highly concordant with the concatenated nucleotide topology (Supplementary Fig. 7). In contrast, the topologies based on the slow- and fast-evolving loci translated as amino acids differ greatly from one another and are highly discordant with the amino acid topology based on all loci (Supplementary Fig. 8). Also, the topologies based on slow and fast-evolving loci as amino acids are poorly supported overall, with some of the low posterior probability (PP) nodes located at the backbone (Supplementary Fig. 8).

To assess the phylogenetic support for the two conflicting hypotheses regarding the monophyly of Panopinae, we implemented a four-cluster analysis with likelihood mapping for the concatenated nucleotide and amino acid datasets, and for each locus separately (Fig. 2). We defined four taxon clusters, two of each containing taxa assigned to Panopinae, and the other two clusters containing taxa of the former Acrocerinae that were recovered nested within Panopinae in the analyses of amino acid data (Fig. 1A; Fig. 2A). Of the three possible unrooted topologies for the four-taxon clusters, only one results in a monophyletic Panopinae (Fig. 2A). Analysis of the concatenated amino acid dataset indicates stronger support for a non-monophyletic Panopinae, with 64.5% of evaluated quartets supporting this hypothesis, while monophyly of Panopinae based on the concatenated amino acid dataset is only supported by 22.7% of the evaluated quartets (Fig. 2C). In contrast, analysis of the concatenated nucleotide dataset indicates stronger support for a monophyletic Panopinae, with 52.6% of all quartets indicating this relationship (Fig. 2C). FcLM analysis of individual loci resulted in a similar scenario. For amino acids, 60.4% of loci supported a non-monophyletic Panopinae, with only 29.2% indicating its monophyly (Fig. 2B, Supplementary Table 1). For nucleotides, alternatively, 46.7% of loci support a monophyletic Panopinae (Fig. 2B, Supplementary Table 1).

We also evaluated whether sequence data in the amino acid and nucleotide datasets (with and without 3rd codon positions) have evolved under globally stationary, time-reversible and homogeneous conditions (SRH) using the software SymTest (Ababneh et al., 2006; Jermiin et al., 2008). Results indicate that sequences in the nucleotide dataset with 3rd codon positions are unlikely to have evolved under globally SRH conditions, since > 90% of Bowker's tests significantly rejected global symmetry (Fig. 3). The nucleotide dataset without 3rd codon positions, by contrast, suffered much less from such violations, with most pairwise comparisons supporting the hypothesis of homogeneity (Fig. 3). Additionally, results indicate that approximately 10% of sequences in the amino acid dataset are unlikely to have evolved under globally SRH conditions (Fig. 3), with deviations from SRH conditions in the amino acid dataset being much greater than in the nucleotide dataset without 3rd codon positions, but much smaller than in the nucleotide dataset with 3rd codon positions.

3.3. Relationships among Acroceridae lineages

In all resulting phylogenies, Acroceridae was recovered as monophyletic (Fig. 1, Supplementary Fig. 1). The diverse subfamily Acrocerinae was recovered as polyphyletic, consisting of four independent lineages; herein we refer to this non-monophyletic assemblage as the former Acrocerinae. The enigmatic *Carvalhoa* Koçak & Kemal and the cosmopolitan *Acrocera* Meigen were recovered in a clade sister to all other Acroceridae. The genus *Ogcodes* was recovered as sister to the remaining acrocerids (except *Carvalhoa* + *Acrocera*). This relationship was well supported in both amino acid and nucleotide trees regardless of the tree estimation method used. Subsequently, the next clade comprised the subfamilies Philopotinae, Panopinae and two independent lineages of the former Acrocerinae. Within the monophyletic Philopotinae, an early dichotomy was recovered, with *Parahelle* Schlinger and *Thyllis* Erichson in one clade, and *Megalybus* Philippi, *Oligoneura* Bigot and *Philopota* Wiedemann in the other (Fig. 1). The former acrocerine genera *Turbopsebius* and *Cyrtus* were placed in a clade subtending the two remaining lineages, one including the former acrocerine genera *Psilodera* and *Pterodontia*, and the other containing the subfamily Panopinae. Primarily in analyses using amino acids, however, Panopinae was not supported to be monophyletic. Within Panopinae, one basal dichotomy was recovered, with one lineage including *Lasia* Wiedemann, *Eulonchus* Gerstaecker and *Panops* Lamarck, and the other comprising *Pialea* Erichson, *Arrhynchus* Philippi, *Exetasis* Walker and *Ocnaea* Erichson.

3.4. Timing of Acroceridae evolution

We used a reduced nucleotide dataset of 65 loci to estimate a chronogram for spider flies. Fossilized birth-death (FBD) process divergence dating (Heath et al., 2014) performed here shows that the origin of crown spider flies dates back to the Upper Jurassic, at approximately 160 MYA (186–156 Ma 95% highest probability density interval, HPD) (Fig. 4, Supplementary Table 2). This new estimate for the age of spider flies is much younger than the 198 MYA estimate recovered in a previous study (Winterton et al., 2007). Our results indicate that the major lineages of Acroceridae were already present by the Upper Cretaceous, but the greatest amount of cladogenesis occurred during the Paleogene, with most genera present by the end of that period. A few genera, however, evolved later in the Miocene, approximately 20–10 Ma ago (Fig. 4). The 95% HPD values for each node are given in Supplementary Table 2, and the numbered nodes in the Acroceridae phylogeny are presented in Supplementary Fig. 9. All 12 spider fly fossils (Gillung and Winterton, 2017) were included as terminals in the dating analyses (Supplementary Table 3). The chronogram was very well supported overall, with all nodes (except for one) in the backbone of the tree with posterior probabilities (PP) of 1 (Fig. 4, Supplementary Table 2).

4. Discussion

4.1. Conflict among data types

We found surprising conflict between phylogenetic signal in the nucleotide and amino acid datasets, which resulted in two well-supported alternative hypotheses for spider fly evolution (Fig. 1). The fact that protein-coding gene sequence data (nucleotides) and their protein translations (amino acids) support conflicting phylogenies is highly significant since both types of data should have evolved under the same species tree as they are extracted from the same observations. The critical difference between the two topologies concerns the monophyly of the traditionally well-established and widely accepted subfamily Panopinae (Schlinger, 1981, 2003; Winterton et al., 2007). The subfamily is recovered as monophyletic in analyses using nucleotide data, and as polyphyletic using amino acids (Fig. 1).

4.2. Reliability of the two alternative hypotheses

The topology based on nucleotides was far more robust to perturbations of the dataset, with results consistent when taxa with low gene occupancy are removed, third codon positions are excluded, loci are sampled based on evolutionary rate, and multiple phylogeny estimation methods are used (BI, ML and MSC). The topology based on amino acids, on the other hand, changes substantially when the data is perturbed, with nodes of interest in the backbone varying considerably (Fig. 1, Supplementary Figs. 4 and 8). Moreover, the MSC analysis of the amino acid dataset suggests extensive levels of conflict among loci, a phenomenon that is further demonstrated in the extreme differences in tree topology if the fastest one third or slowest one third of the loci are analyzed separately (Supplementary Fig. 8).

We performed four-cluster likelihood mapping analysis (FcLM) to further understand the nature of the conflict among loci in the nucleotide and amino acid datasets. Results showed that the nucleotide

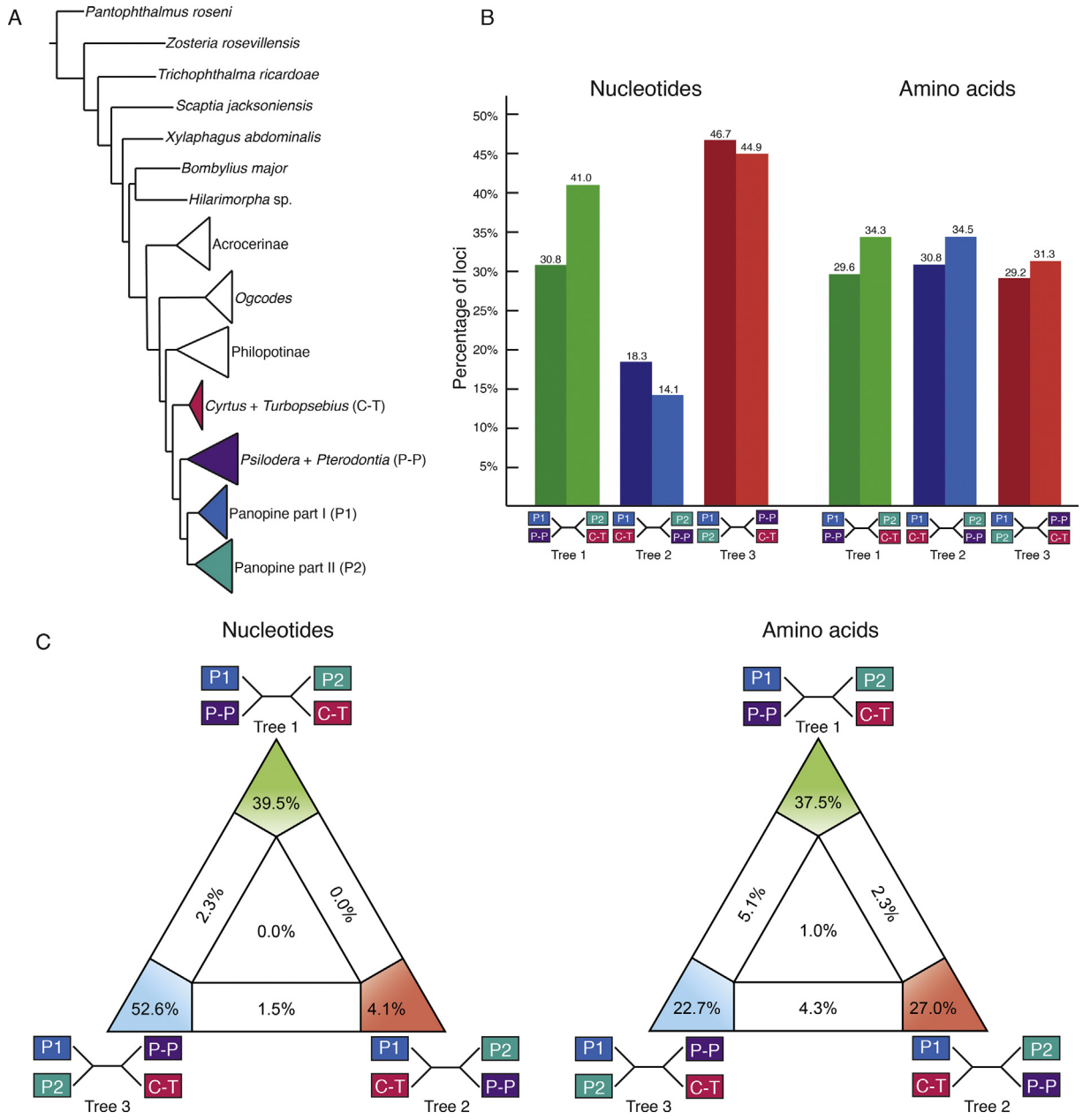

Fig. 2. Four-cluster likelihood mapping (FcLM) analyses results. A. Phylogram of Acroceridae based on nucleotides showing the four taxon clusters used. B. FcLM results for each individual locus in the nucleotide and amino acid datasets. Bars show the percentage of loci supporting each of the three possible unrooted topologies, with darker colors (bars on left) showing raw percentages, and lighter colors (bars on left) showing percentages weighted over relative support for each topology as shown in Supplementary Table 1. C. FcLM results for the concatenated nucleotide and amino acid alignments. Values at the corners indicate the percentage of fully resolved phylogenies for all possible quartets.

dataset had much stronger support for the preferred topology, while in the amino acid dataset there was roughly equal support for the three alternative topologies (Fig. 2). Even though one topology was preferred in the amino acid dataset, its weight was not much greater than the weight towards the other two topologies (Fig. 2). When nucleotide loci were analyzed individually, a clear majority of genes supported the same topology that was preferred in the concatenated analysis (Fig. 2, Supplementary Table 1). In the case of the nucleotide dataset, when topology preference for each locus was weighted over its relative strength of support (see Supplementary Table 1), the support for the preferred tree was greater than the second preferred tree, while the third topology was supported by only a handful of genes. In contrast, when amino acid loci were analyzed individually, support was roughly equally split over the three possible topologies (Fig. 2, Supplementary Table 1). In summary, the nucleotide data was less equivocal regarding the preferred topology in both concatenated and individual loci analyses, while there was more conflict as to which of the topologies were preferred using amino acid data (Fig. 2, Supplementary Table 1).

Results of SymTest indicate that the amino acid and nucleotide datasets including 3rd codon positions violated, at least to some degree, the assumption of global stationarity, reversibility and homogeneity (SRH conditions) (Fig. 3). Violation of SRH conditions was much greater in the nucleotide dataset including 3rd codon positions, but when 3rd codon positions were removed, fewer violations were observed in the nucleotide dataset than in the amino acid dataset (Fig. 3). Since we obtained virtually the same tree topology when analyzing the nucleotide dataset with or without 3rd codon positions, this indicates that violation of SRH conditions was unlikely to strongly impact on our

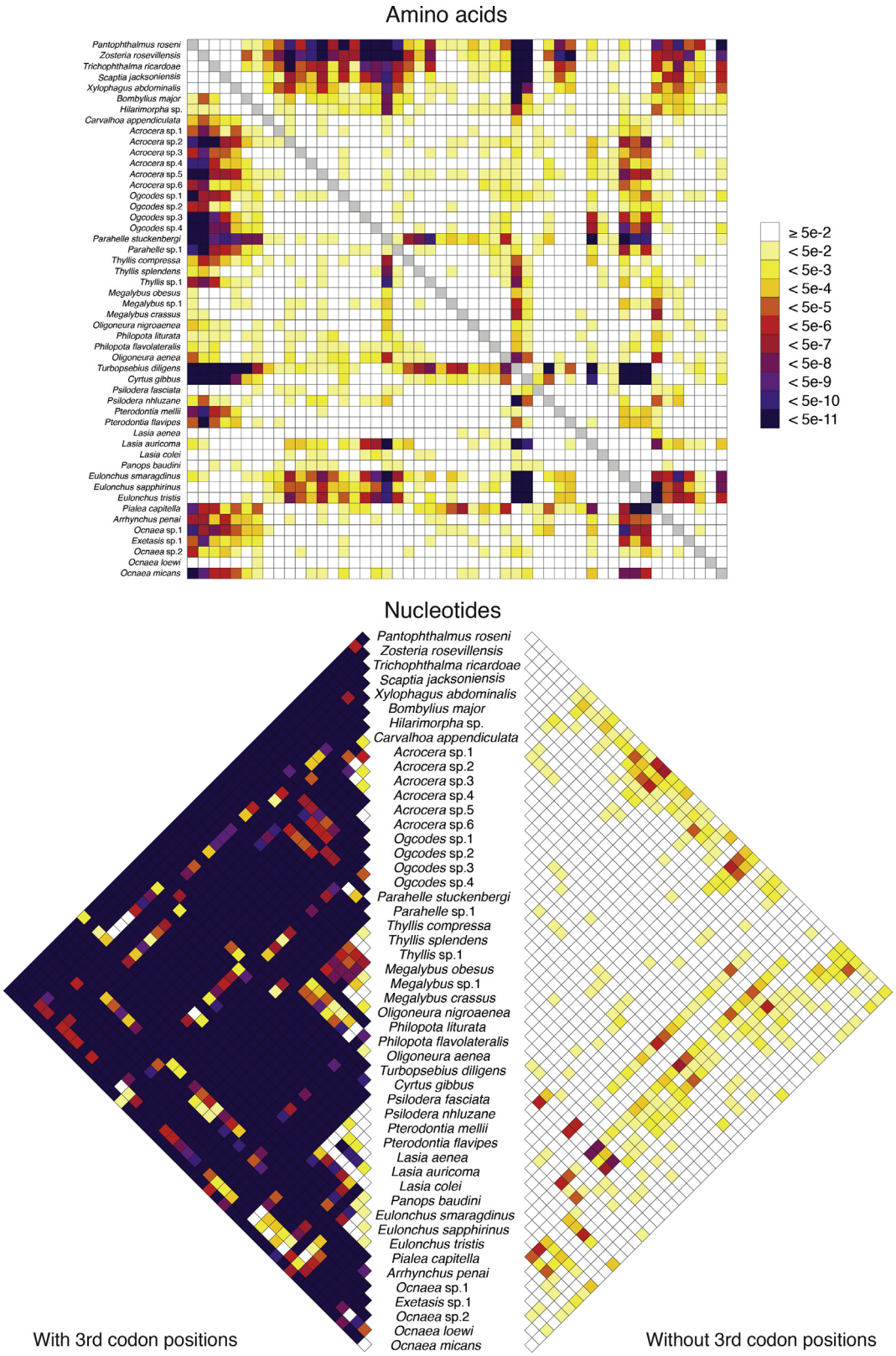

Fig. 3. Heat maps showing the results from pairwise comparison of aligned amino acid and nucleotide sequences (with and without 3rd codon positions) using Bowker's matched-pairs tests of symmetry. Cells in white specify p-values > 0.05, indicating that the corresponding pair of sequences seemingly does not violate the assumption of global stationarity, reversibility and homogeneity (SRH conditions).

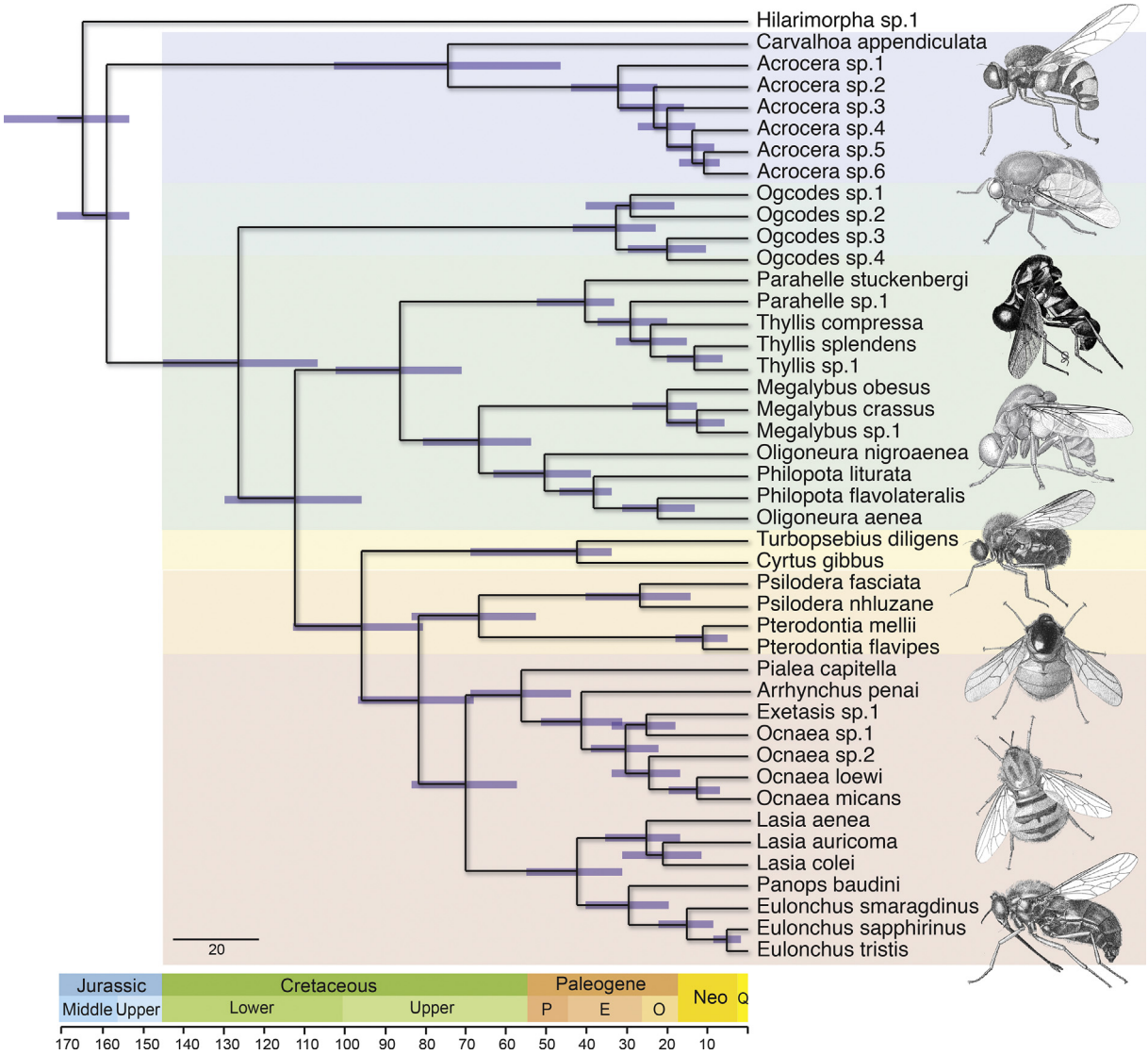

Fig. 4. Estimated divergence times among lineages of Acroceridae under the fossilized birth-death process, in BEAST 2. Scale is in MYA. Bars depict the 95% highest posterior probability density of each estimate. Mean ages and ranges are provided in Supplementary Table 3 and refer to nodes indicated in Supplementary Fig. 10.

results. Thus, the conflict between the amino acid and the nucleotide topologies is likely not linked to SRH violation as measured by SymTest. It is generally assumed that phylogeny estimation based on nucleotides generally performs worse than amino acids specifically because nucleotides tend to violate SRH conditions to a greater extent than amino acids (Zwick et al., 2012; Rota-Stabelli et al., 2013; Cox et al., 2014). Nonetheless, we found evidence supporting the opposite case in our study. The removal of 3rd positions from the nucleotide dataset heavily reduced violation of SRH conditions. Thus, if tree estimation based on nucleotides was affected by violation of SRH conditions while that based on amino acids was not, the expectation is that after the removal of 3rd positions the resulting topology should be congruent with the amino acid tree.

When synonymous changes in the nucleotide dataset were masked using the *degen1* approach, the topologies based on the original and degenerated datasets were critically discordant (Fig. 1, Supplementary Figs. 4–6). These results suggest that there may be conflict in phylogenetic signal originating from synonymous and nonsynonymous changes. We compared the overall number of variable sites in the original and degenerated nucleotide datasets and observed that the overall proportion of variable sites in 1st and 3rd codon positions decreased considerably when synonymous changes were masked. Overall proportion of variable sites decreased from 50% in the complete nucleotide alignment to only 14% in the degenerated dataset (Supplementary Fig. 10). Synonymous substitutions in 1st codon positions may be contributing the phylogenetic signal supporting the monophyly of Panopinae, because when synonymous changes are excluded, the resulting topology supports a non-monophyletic Panopinae (Supplementary Figs. 5 and 6).

Determining whether analysis of synonymous versus non-synonymous changes is likely to be more inaccurate is not trivial, especially in light of the absence of gross SRH violations in both cases, as measured by SymTest. We speculate that phenomena, including different selection regimes and different patterns of non-independence among sites, may result in nucleotide and amino acid sequences that subtly violate the assumptions of common phylogenetic models, which could affect inference based on synonymous and non-synonymous changes in different directions. Additionally, our FcLM results suggest differences in patterns of topological conflict among loci in the nucleotide and amino acid datasets. Gillung and Khouri et al. (in prep) are using posterior predictive simulation (Bollback, 2002; Doyle et al., 2015; Duchene et al., 2016) to evaluate absolute model fit to the current datasets and investigating whether model misspecification is the source of conflict among and within the datasets.

The conflict within the results based on amino acids may have biological or methodological causes, including, for instance, incomplete lineage sorting and poor model fit to some genes or subsets of data, respectively. In either case, this decreases the credibility of the topology inferred from the amino acid concatenated dataset. If conflict among gene trees is real and pervasive, not modelling it explicitly could result in inaccurate estimates of topology. Species tree estimation methods can account for some of the biological sources of conflict; however, despite favoring a topology similar to that inferred from the concatenated dataset, our ASTRAL results are inconclusive due to the low support for the nodes of interest.

Results based on analyses of nucleotides were more robust to alterations of data and different analytical methods. This implies that either the results are accurate, or that there is pervasive systematic error affecting all nucleotide analyses (and most genes or data subsets) in the same way. Violation of SRH conditions is thought to be the most common source of error disproportionately affecting nucleotide analyses. Given that our SymTest results suggest that this is not the case for our dataset, we prefer the hypothesis of a monophyletic Panopinae, until further investigation.

4.3. Patterns of Acroceridae evolution

This study comprises the first phylogenomic treatment of Acroceridae relationships, with molecular sequence data sampled across all major spider fly lineages. Analyses of nucleotide data converged upon a fully resolved and well-supported tree topology that is incongruent with traditional hypotheses based on morphology (Schlinger, 1987) and smaller sampling of molecular data (Winterton et al., 2007). Unprecedented aspects of our results include the placement of the morphologically derived, species-rich genus *Ogcodes* as sister to the rest of the Acroceridae other than *Acrocera* + *Carvalhoa*, and the non-monophyly of some traditionally well-established genera, including *Parahelle* and *Ocnaea* (Fig. 1). More importantly, the non-monophyly of Acrocerinae, already suggested by Winterton et al. (2007), indicates pervasive and strong discordance between traditional morphological systematics and molecular phylogenetic results.

Acrocerinae were polyphyletic in all of our analyses, with four independent clades. A clade composed of the enigmatic Chilean genus *Carvalhoa* and the cosmopolitan, species-rich *Acrocera* was recovered as sister to all other spider flies, in general agreement with previous molecular results (Winterton et al., 2007). Whilst adult *Acrocera* and *Carvalhoa* are morphologically similar to other Acrocerinae, having relatively small heads, bulbous bodies and reduced wing venation, their larval morphology and behaviour contrast with the rest of the family. Species in these two genera have unique associations with araneomorph spiders in the Haplogynae, while the remaining acrocerids attack Entelegynae araneomorph spiders or Mygalomorphae spiders (Gillung and Borkent, 2017). Additionally, the first instar planidial larvae of all other acrocerids have well sclerotized body segments with setae or scales allowing them to actively locomote via looping, leaping and flicking movements (King, 1916; Schlinger, 1960b, 2003). Instead, *Acrocera* first instar larvae lack both sclerotization and long setae, and only crawl (Overgaard Nielsen et al., 1999).

The placement of *Ogcodes* as sister to the remaining Acroceridae (excluding *Carvalhoa* + *Acrocera*) was recovered with strong support in all analyses irrespective of data type (nucleotides or amino acids), phylogeny inference method (BI, ML or MSC) and alternative taxon and gene sampling. The genus was previously placed in the former Acrocerinae based on morphology (Schlinger, 1987) and Sanger sequence data, albeit with low confidence (Winterton et al., 2007). This placement is justifiable as species of *Ogcodes* have small body size and reduced wing venation, highly apomorphic traits that are likely interrelated.

The two remaining clades of the former Acrocerinae comprise the genera *Turbopsebius* + *Cyrtus* and *Pterodontia* + *Psilodera*. Schlinger (1972) postulated the *Cyrtus–Opsebius* (including *Turbopsebius*) lineage of Acroceridae, and our results confirm their close relationship. The phylogenetic position of both *Pterodontia* and *Psilodera* has always been contentious (Schlinger, 1960a, 1972; Winterton et al., 2007). *Psilodera* was previously affiliated with the acrocerines *Pterodontia* and *Ogcodes*, although with weak statistical support (Winterton et al., 2007). Here, *Psilodera* and *Pterodontia* were recovered as the sister clade to the Panopinae in the nucleotide topology.

Monophyly of the bizarre Philopotinae has never been contested, and the clade was, unsurprisingly, recovered with strong statistical support in all of our analyses. Several morphological features define the subfamily, including enlarged postpronotal lobes forming a collar around the head and a distinct arched body shape (Schlinger, 1987). Our phylogenomic analyses also strongly support the internal arrangement of Philopotinae as proposed by Winterton et al. (2007), with two main clades recovered. The first clade includes the genera *Megalybus*, *Philopota* and *Oligoneura*, with *Parahelle* and *Thyllis* in the second clade.

In all nucleotide-based topologies, Panopinae are monophyletic with high statistical support. Overall, reciprocal monophyly of individual panopine genera is well supported, except for *Exetasis* and *Ocnaea*. Schlinger (1968) differentiated the two genera based on two weak morphological characters, distribution of the microtrichia on the wing membrane and the absence of the wing vein R_4 in *Exetasis*, though other authors have dissented. Our results indicate that the two genera should probably be synonymized.

4.4. Timeline of Acroceridae evolution and diversification

The origin of Acroceridae has been estimated by Wiegmann et al. (2003) at approximately 175–225 MYA, and by Winterton et al. (2007) at *ca.* 173–221 MYA. Our results indicate a much younger age for the origin of the family in the Middle to Late Jurassic (156–187 MYA). This difference in age estimates might be due to a different calibration approach used here (tip dating versus node dating), greater fossil sampling, and a revised, younger age estimate of Baltic amber (Aleksandrova and Zaporozhets, 2008). The age and plesiomorphic appearance of the oldest definitive spider flies, *Archocyrtus gibbosus* Ussatchov and *A. kovalevi* (Nartshuk), both described from late Jurassic fossil beds from Karatau, Kazakhstan (Ussatchov, 1968; Nartshuk, 1996), are consistent with a late Mesozoic origin and Cretaceous diversification of Acroceridae.

Whilst *Carvalhoa* + *Acrocera* diverged from the rest of the family relatively early (156–174 MYA, Middle Jurassic), the rest of Acroceridae radiated more recently, with the divergence of *Ogcodes* from the rest of Acroceridae occurring 45 million years later, in the Lower Cretaceous. Philopotinae diverged from Panopinae and remaining Acrocerinae in the Lower Cretaceous (97–131 MYA). Within Philopotinae, the New World and Oriental genera (*Megalybus*, *Philopota* and *Oligoneura*) diverged from Afrotropical genera (*Parahelle* and *Thyllis*) approximately 72–103 MYA, towards the Upper Cretaceous. Finally, Acrocerinae *partim* and Panopinae diverged during the Middle Cretaceous (82–114 MYA), with crown group Panopinae appearing during the Upper Cretaceous (*ca.* 98–69 MYA) (Fig. 4, Supplementary Table 2).

5. Conclusion

We applied a phylogenomic approach to resolve the phylogeny of spider flies, sampling molecular sequence data of 240 homologous genes from all major lineages. Analyses of supermatrices as well as species tree approaches converged upon a robust hypothesis of Acroceridae evolution based on nucleotides under a variety of analytical parameters.

Acroceridae is remarkable within Diptera as the only parasitoid group specialized in spiders, with remaining fly parasitoids mainly

attacking other insects. We took advantage of the recent advances in divergence time estimation (Heath et al., 2014) to propose a robust hypothesis for the timing of spider fly evolution, in which all known fossil acrocerids were included as terminals, with ages ranging from the Upper Jurassic to the Miocene (Gillung and Winterton, 2017). The lack of clarity concerning the position of Acroceridae within the Diptera tree of life, however, limits how we can interpret the timing of the diversification of this specialized group of spider endoparasitoids.

Although more sequence data have often been shown to help resolve difficult phylogenetic questions, our study of spider fly phylogeny shows that simply increasing the amount of data can in fact be detrimental if added sequences have properties that introduce conflict among data types. When large-scale data matrices are used to study challenging nodes in the tree of life, relatively subtle model violations may be sufficiently amplified to mislead analyses, and those violations may not be obvious in many datasets. Thus, the comparative and exploratory approach implemented here may be a desirable way to detect conflicting signal in phylogenomic analyses. Specifically, it is important to compare topologies based on both amino acids and nucleotides because, even though they represent merely alternative coding of the same underlying data, their statistical analyses are fundamentally different (Huelsenbeck et al., 2008). In particular, the customary use of empirical amino acid models in which all of the potentially free parameters are fixed to specific values may be a source of model violation. Also, the use of global exchangeability rates as implemented in the CAT + GTR + G model might introduce tremendous amounts of model misspecification, because under this model it is assumed that all partitions share the same exchangeability rates.

Our results further provide an insight into the question of data type choice in phylogenomics and the importance of analyzing data both as amino acids and nucleotides. Careful analyses of data are critical, especially when larger amounts of sequence data are becoming available for inclusion in phylogenetic studies. Exploratory analyses such as tests of compositional heterogeneity, posterior predictive approaches to assess absolute model fit (Bollback, 2002; Doyle et al., 2015; Duchene et al., 2016), or sensitivity of results to removal of sites likely to introduce systematic error (Salichos and Rokas, 2013; Goremykin et al., 2015) should become a part of the standard phylogenomics toolkit. In addition, future work on phylogenomics should focus on better understanding of how different biases affect different data sources.

Acknowledgements

This work was supported by the U.S. National Science Foundation (DEB-1144119 to SLW), and by the Brazilian National Council for Scientific and Technological Development (grant 209447/2013-3 to JPG). Thank you to the following people from the 1KITE team: Lars Podsiadlowski, Alexander Donath, Daniela Bartel, Sabrina Simon, Karen Meusemann, and to the 1KITE Antliophora group for providing transcriptome data (http://www.1kite.org/subprojects.html). Thank you to Lars Jermiin for making SymTest available for the authors. Thank you to Silvio S. Nihei and Carlos E. Lamas for the loan of material from the Museum of Sao Paulo (MZSP) for DNA sequencing. Also, thank you to Michelle Trautwein for contributing analytical organization and capability for the Wiegmann Lab. Thank you to Brian Cassel for his help with targeted enrichment and initial processing of sequences. Comments from Brendon B. Boudinot, Phillip S. Ward, Brian Moore, Karen Meusemann and Alfried Vogler refined some of the ideas presented here.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ympev.2018.08.007.

References

Ababneh, F., Jermiin, L.S., Ma, C., Robinson, J., 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22, 1225–1231.
Aberer, A.J., Kobert, K., Stamatakis, A., 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31, 2553–2556.
Adachi, J., Hasegawa, M., 1996. MOLPHY vol 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28, 1–150.
Adachi, J., Waddell, P.J., Martin, W., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50, 348–358.
Aleksandrova, G.N., Zaporozhets, N.I., 2008. Palynological characteristics of Upper Cretaceous and Paleogene deposits on the west of the Sambian Peninsula (Kaliningrad region), part 1. Stratigr. Geol. Correl. 16, 295–316.
Andrews, S., 2010. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Bininda-Emonds, O.R.P., 2007. Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evol. Bioinform. 3, 59–85.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.
Bollback, J.P., 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19, 1171–1180.
Borkent, C.J., Gillung, J.P., Winterton, S.L., 2016. Jewelled spider flies of North America: a revision and phylogeny of *Eulonchus* Gerstaecker (Diptera, Acroceridae). ZooKeys 619, 103–146.
Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4 e1660.
Borowiec, M.L., 2017. Convergent evolution of the army ant syndrome and congruence in big-data phylogenetics. BioRXiv. https://doi.org/10.1101/134064.
Borowiec, M.L., Lee, E.K., Chiu, J.C., Plachetzki, D.C., 2015. Dissecting phylogenetic signal and accounting for bias in whole-genome data sets: a case study of the Metazoa. Mol. Biol. Evol. 16, 987.
Bowker, A.H., 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43, 572–574.
Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 10 e1003537.
Branstetter, M.G., Danforth, B.N., Pitts, J.P., Faircloth, B.C., Ward, P.S., Buffington, M.L., Gates, M.W., Kula, R.R., Brady, S.G., 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27, 1019–1025.
Cady, A., Leech, R., Sorkin, L., Stratton, G., Caldwell, M., 1993. Acrocerid (Insecta: Diptera) life histories, behaviors, host spiders (Arachnida: Araneida), and distributional records. Can. Entomol. 125, 931–944.
Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Pääbo, S., Hasegawa, M., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47, 307–322.
Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., 2015. Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. U.S.A. 112, 14912–14917.
Cox, C.J., Li, B., Foster, P.G., Embley, T.M., Civán, P., 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst. Biol. 63, 272–279.
Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8, 783–786.
Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. Supplement 3 In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington DC, pp. 345–352.
Doyle, V.P., Young, R.E., Naylor, G.J.P., Brown, J.M., 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64, 824–837.
Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4 e88.
Duchene, S., Di Giallonardo, F., Holmes, E.C., 2016. Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. Mol. Biol. Evol. 33, 255–267.
Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.
Espeland, A., Breinholt, M., Willmott, J., Warren, K.R., Vila, A.D., Toussaint, R., Maunsell, E.F.A., Aduse-Poku, S.C., Talavera, K., Eastwood, G., et al., 2018. A comprehensive and dated phylogenomic analysis of butterflies. Curr. Biol. 28, 770–778.
Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
Fučíková, K., Lewis, P.O., Lewis, L.A., 2016. Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence evolution. Mol. Phylogenet. Evol. 98, 176–183.
Garrison, N.L., Rodriguez, J., Agnarsson, I., Coddington, J.A., Griswold, C.E., Hamilton, C.A., Hedin, M., Kocot, K.M., Ledford, J.M., Bond, J.E., 2016. Spider phylogenomics: untangling the spider tree of life. PeerJ 4 e1719.
Gillung, J.P., Winterton, S.L., 2011. New genera of philopotine spider flies (Diptera, Acroceridae) with a key to living and fossil genera. ZooKeys 127, 15–27.
Gillung, J.P., Nihei, S.S., 2016. Evolution of Philopotinae, with a revision and phylogeny of the New World spider fly genus *Philopota* Wiedemann (Diptera, Acroceridae). Zool. J. Linnean. Soc. 176, 707–780.
Gillung, J.P., Borkent, C.J., 2017. Death comes on two wings: a review of dipteran natural

enemies of arachnids. J. Arachn. 45, 1–19.
Gillung, J.P., Winterton, S.L., 2017. A review of fossil spider flies (Diptera: Acroceridae) with descriptions of new genera and species from Baltic Amber. J. Syst. Palaeontol. 16, 325–350.
Goremykin, V.V., Nikiforova, S.V., Cavalieri, D., Pindo, M., Lockhart, P., 2015. The root of flowering plants and total evidence. Syst. Biol. 64, 879–891.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652.
Haddad, S., Shin, S., Lemmon, A.R., Lemmon, E.M., Svacha, P., Farrell, B.D., Slipinski, A., Windsor, D., McKenna, D.D., 2018. Anchored hybrid enrichment provides new insights into the phylogeny and evolution of longhorned beetles (Cerambycidae). Syst. Ent. 43, 68–89.
Hamilton, C.A., Lemmon, A.R., Lemmon, E.M., Bond, J.E., 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC Evol. Biol. 16, 212.
Heath, T.A., Huelsenbeck, J.P., Stadler, T., 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. U.S.A. 111, E2957–E2966.
Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919.
Huelsenbeck, J.P., Joyce, P., Lakner, C., Ronquist, F., 2008. Bayesian analysis of amino acid substitution models. Philos. Trans. R. Soc. B 363, 3941–3953.
Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.
Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231.
Jermiin, L., Ho, S.Y.W., Ababneh, F., Robinson, J., Larkum, A.W., 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53, 638–643.
Jermiin, L.S., Jayaswal, V., Ababneh, F., Robinson, J., 2008. Phylogenetic model evaluation. In: Keith, J.M. (Ed.), Bioinformatics, Volume 1: Data, Sequence Analysis, and Evolution. Humana Press, Totowa, pp. 331–364.
Jia, F., Lo, N., Ho, S.Y.W., 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS One 9 e95722.
Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
King, J.L., 1916. Observations on the life history of Pterodontia flavipes Gray (Diptera). Ann. Entomol. Soc. Am. 9, 309–321.
Kocot, K.M., Struck, T.H., Merkel, J., Waits, D.S., Todt, C., Brannock, P.M., Weese, D.A., Cannon, J.T., Moroz, L.L., Lieb, B., et al., 2016. Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst. Biol. 66, 256–282.
Kozlov, A.M., Aberer, A.J., Stamatakis, A., 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31, 2577–2579.
Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B.M., Wägele, J.W., Misof, B., 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front. Zool. 7, 10.
Kutty, S.N., Wong, W.H., Meusemann, K., Meier, R., Cranston, P.S., 2018. A phylogenomic analysis of Culicomorpha (Diptera) resolves the relationships among the eight constituent families. Syst. Ent. 35, 823–836.
Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A., 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14, 82.
Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34, 772–773.
Lartillot, N., Philippe, H., 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Syst. Biol. 21, 1095–1109.
Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7, S4.
Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286.
Le, S.Q., Dang, C.C., Gascuel, O., 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29, 2921–2936.
Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.
Mayrose, I., Friedman, N., Pupko, T., 2005. A gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21, 151–158.
Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop, New Orleans.
Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52.
Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T., 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346, 250463–1250463.
Misof, B., Misof, K.A., 2009. Monte Carlo approach successfully identifies randomness of multiple sequence alignments: a more objective approach of data exclusion. Syst. Biol. 58, 21–34.
Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K., Meusemann, K., 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinf. 14, 348.
Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767.
Nabholz, B., Künstner, A., Wang, R., Jarvis, E.D., Ellegren, H., 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. 28, 2197–2210.
Nartshuk, E.P., 1996. A new fossil acrocerid fly from the Jurassic beds of Kazakhstan (Diptera: Acroceridae). Zoosystematica Rossica 4, 313–315.
Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.
Overgaard Nielsen, B., Funch, P., Toft, S., 1999. Self-injection of a dipteran parasitoid into a spider. Naturwissenschaften 86, 530–532.
Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., et al., 2017. Evolutionary history of the Hymenoptera. Curr. Biol. 27, 1013–1018.
Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., et al., 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinf. 18, 111.
Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. U.S.A. 112, 15402–15407.
Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–573.
Rambaut, A., Suchard, M.A., Xie, D., Drummond, A.J., 2014. Tracer v1.6, Available from http://tree.bio.ed.ac.uk/software/tracer/.
Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.L., Harshman, J., Huddleston, C.J., Kingston, S., et al., 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66, 857–879.
Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.
Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.
Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D., Rasnitsyn, A.P., 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61, 973–999.
Rota-Stabelli, O., Lartillot, N., Philippe, H., Pisani, D., 2013. Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst. Biol. 62, 121–133.
Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol. Biol. Evol. 30, 197–214.
Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signal. Nature 497, 327–331.
Schlinger, E.I., 1960a. A review of the South African Acroceridae (Diptera). Ann. Natal Museum 14, 459–504.
Schlinger, E.I., 1960b. A review of the genus *Eulonchus* Gerstaecker. Part I. The species of the smaragdinus group (Diptera: Acroceridae). Ann. Am. Entomol. Soc. 53, 416–422.
Schlinger, E.I., 1968. A revision of *Arrynchus* Philippi and a key to the genera of the *Ocnaea* branch of the Panopinae (Diptera). Rev. Chil. Entomol. 6, 47–54.
Schlinger, E.I., 1972. New east Asian and American genera of the "*Cyrtus-Opsebius*" branch of the Acroceridae (Diptera). Pacific Insects 14, 409–428.
Schlinger, E.I., 1981. Acroceridae. In: McAlpine, J.F., Peterson, B.V., Shewell, G.E., Teskey, H.J., Vockeroth, J.R., Wood, D.M. (Eds.), Manual of Nearctic Diptera. Vol. 1. Agriculture Canada Research Branch, Monograph 27, Ottawa, pp. 575–584.
Schlinger, E.I., 1987. The biology of Acroceridae (Diptera): true endoparasitoids of spiders. In: Nentwig, W. (Ed.), Ecophysiology of Spiders. Springer-Verlag, Germany, pp. 319–327.
Schlinger, E.I., 2003. Acroceridae, spider endoparasitoids. In: Goodman, S.M., Benstead, J.P. (Eds.), The Natural History of Madagascar. University of Chicago Press, Chicago, pp. 734–740.
Schlinger, E.I., Gillung, J.P., Borkent, C.J., 2013. New spider flies from the Neotropical Region (Diptera, Acroceridae) with a key to New World genera. Zookeys 270, 59–93.
Shaffer, H., Minx, P., Warren, D., Shedlock, A.M., Thomson, R.C., Valenzuela, N., Abramyan, J., Badenhorst, D., Biggar, K.K., Borchert, G.M., et al., 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14, R28.
Shin, S., Clarke, D.J., Lemmon, A.R., Lemmon, E.M., Aitken, A.L., Haddad, S., Farrell, B.D., Marvaldi, A.E., Oberprieler, R.G., McKenna, D.D., 2017. Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. Mol. Biol. Evol. 35, 823–836.
Shin, S., Bayless, K.M., Winterton, S.L., Dikow, T., Lessard, B.D., Yeates, D.K., Wiegmann, B.M., Trautwein, M.D., 2018. Taxon sampling to address an ancient rapid radiation: a supermatrix phylogeny of early brachyceran flies (Diptera). Syst. Ent. 43, 277–289.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Streicher, J.W., Schulte, J.A., Wiens, J.J., 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65, 128–145. https://academic.oup.com/sysbio/article/65/1/128/2461451.

Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 94, 6815–6819.

Sullivan, J., Swofford, D.L., Naylor, G.J.P., 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16, 1347–1356.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, 609–612.

Ussatchov, D.A., 1968. New Jurassic Asilomorpha (Diptera) fauna from Karatau. Entomol. Rev. 47, 617–628.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666.

Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., Li, C., White, S., Xiong, Z., Fang, D., et al., 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat. Genet. 45, 701–706.

Waterhouse, R.M., Tegenfeldt, F., Zdobnov, E.M., Kriventseva, E.V., 2013. OrthoDB: a hierarchical catalog of animal fungal and bacterial orthologs. Nucleic Acids Res. 41, 358–365.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

Wiegmann, B.M., Thorne, J.L., Yeates, D.K., Kishino, H., 2003. Time flies: A new molecular time-scale for fly evolution without a clock. Syst. Biol. 52, 745–756.

Wiegmann, B.M., Trautwein, M.D., Winkler, I.S., Barr, N.B., Kim, J.W., Lambkin, C., Bertone, M.A., Cassel, B.K., Bayless, K.M., Heimberg, A.M., et al., 2011. Episodic radiations in the fly tree of life. Proc. Natl. Acad. Sci. U.S.A. 108, 5690–5695.

Winterton, S.L., Wiegmann, B.M., Schlinger, E.I., 2007. Phylogeny and Bayesian divergence time estimations of small-headed flies (Diptera: Acroceridae) using multiple molecular markers. Mol. Phylogenet. Evol. 43, 808–832.

Winterton, S.L., Gillung, J.P., 2012. A new species of spider fly in the genus *Sabroskya* Schlinger from Malawi, with a key to Acrocerinae world genera (Diptera, Acroceridae). Zookeys 171, 1–15.

Winterton, S.L., Lemmon, A.R., Gillung, J.P., Garzon, I.J., Badano, D., Bakkes, D.K., Breitkreuz, L.C.V., Engel, M.S., Lemmon, E.M., Liu, X., et al., 2018. Evolution of lacewings and allied orders using anchored phylogenomics (Neuroptera, Megaloptera, Raphidioptera). Syst. Entomol. 43, 330–354.

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evolut. 11, 367–372.

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford.

Yeates, D.K., Meusemann, K., Trautwein, M., Wiegmann, B., Zwick, A., 2016. Power, resolution and bias: recent advances in insect phylogeny driven by the genomic revolution. Curr. Opin. Insect Sci. 13, 16–23.

Young, A.D., Lemmon, A.R., Skevington, J.H., Mengual, X., Ståhls, G., Reemer, M., Jordaens, K., Kelso, S., Lemmon, E.M., Hauser, M., et al., 2016. Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). BMC Evol. Biol. 16, 143.

Zwick, A., Regier, J.C., Zwickl, D.J., 2012. Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. PLoS One 7 e47450.

Please cite this article in press as: Peters et al., Evolutionary History of the Hymenoptera, Current Biology (2017), http://dx.doi.org/10.1016/j.cub.2017.01.027

Current Biology
Report

Evolutionary History of the Hymenoptera

Ralph S. Peters,[1,24,25,*] Lars Krogmann,[2] Christoph Mayer,[3] Alexander Donath,[3] Simon Gunkel,[4] Karen Meusemann,[3,5,6] Alexey Kozlov,[7] Lars Podsiadlowski,[8] Malte Petersen,[3] Robert Lanfear,[9,10] Patricia A. Diez,[11] John Heraty,[12] Karl M. Kjer,[13] Seraina Klopfstein,[14] Rudolf Meier,[15] Carlo Polidori,[16] Thomas Schmitt,[17] Shanlin Liu,[18,19,20] Xin Zhou,[21,22] Torsten Wappler,[4] Jes Rust,[4] Bernhard Misof,[3] and Oliver Niehuis[3,5,23,24,*]

[1]Center of Taxonomy and Evolutionary Research, Arthropoda Department, Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany
[2]Entomologie, Staatliches Museum für Naturkunde Stuttgart, 70191 Stuttgart, Germany
[3]Center for Molecular Biodiversity Research, Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany
[4]Steinmann Institut für Geologie, Mineralogie und Paläontologie, 53115 Bonn, Germany
[5]Department of Evolutionary Biology and Ecology, Institute for Biology I (Zoology), University of Freiburg, 79104 Freiburg (Brsg.), Germany
[6]Australian National Insect Collection, CSIRO National Research Collections Australia (NRCA), Acton, ACT 2601, Australia
[7]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany
[8]Institute of Evolutionary Biology and Ecology, University of Bonn, 53121 Bonn, Germany
[9]Ecology, Evolution and Genetics, Research School of Biology, Australian National University, Canberra, ACT 2601, Australia
[10]School of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia
[11]Centro de Investigaciones y Transferencia de Catamarca, CITCA-CONICET/UNCA, 4700 Catamarca, Argentina
[12]Department of Entomology, University of California, Riverside, Riverside, CA 92521, USA
[13]Department of Biological Sciences, Rutgers University, Newark, NJ 07102, USA
[14]Naturhistorisches Museum der Burgergemeinde Bern, 3005 Bern, Switzerland
[15]Department of Biological Sciences and Lee Kong Chian Natural History Museum, National University of Singapore, Singapore 117543, Singapore
[16]Instituto de Ciencias Ambientales (ICAM), Universidad de Castilla-La Mancha, 45071 Toledo, Spain
[17]Department of Animal Ecology and Tropical Biology, University of Würzburg, 97074 Würzburg, Germany
[18]China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen, Guangdong Province, 518083, People's Republic of China
[19]BGI-Shenzhen, Shenzhen, Guangdong Province, 518083, People's Republic of China
[20]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark
[21]Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, People's Republic of China
[22]Department of Entomology, China Agricultural University, Beijing 100193, People's Republic of China
[23]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
[24]These authors contributed equally
[25]Lead Contact
*Correspondence: r.peters@leibniz-zfmk.de (R.S.P.), oliver.niehuis@biologie.uni-freiburg.de (O.N.)
http://dx.doi.org/10.1016/j.cub.2017.01.027

SUMMARY

Hymenoptera (sawflies, wasps, ants, and bees) are one of four mega-diverse insect orders, comprising more than 153,000 described and possibly up to one million undescribed extant species [1, 2]. As parasitoids, predators, and pollinators, Hymenoptera play a fundamental role in virtually all terrestrial ecosystems and are of substantial economic importance [1, 3]. To understand the diversification and key evolutionary transitions of Hymenoptera, most notably from phytophagy to parasitoidism and predation (and vice versa) and from solitary to eusocial life, we inferred the phylogeny and divergence times of all major lineages of Hymenoptera by analyzing 3,256 protein-coding genes in 173 insect species. Our analyses suggest that extant Hymenoptera started to diversify around 281 million years ago (mya). The primarily ectophytophagous sawflies are found to be monophyletic. The species-rich lineages of parasitoid wasps constitute a monophyletic group as well. The little-known, species-poor Trigonaloidea are identified as the sister group of the stinging wasps (Aculeata). Finally, we located the evolutionary root of bees within the apoid wasp family "Crabronidae." Our results reveal that the extant sawfly diversity is largely the result of a previously unrecognized major radiation of phytophagous Hymenoptera that did not lead to wood-dwelling and parasitoidism. They also confirm that all primarily parasitoid wasps are descendants of a single endophytic parasitoid ancestor that lived around 247 mya. Our findings provide the basis for a natural classification of Hymenoptera and allow for future comparative analyses of Hymenoptera, including their genomes, morphology, venoms, and parasitoid and eusocial life styles.

RESULTS AND DISCUSSION

We sequenced whole-body transcriptomes of 167 species of Hymenoptera and selected outgroups and supplemented our

dataset with sequenced and annotated genomes of five hymenopterans and a beetle (for details, see Supplemental Experimental Procedures and Data S1A–S1D). Our study includes 54 families of Hymenoptera, representing all major superfamilies. The phylogenetic inferences are based on the analysis of 1.5 million amino acid and 3.0 million nucleotide positions, respectively, derived from 3,256 single-copy protein-coding genes (Data S1E) and inferred by using a combination of domain-, gene-, and codon position-based data partition schemes to improve the fitting of the applied substitution models. Considering the taxonomic and molecular sampling, this is the most comprehensive dataset ever generated for investigating phylogenetic relationships within Hymenoptera or any other insect group. The dataset was furthermore used to estimate divergence times with an independent-rates as well as with a correlated-rates molecular clock approach (Data S1H) and a validated set of 14 fossils (Data S1F).

The inferred phylogenetic relationships and divergence time estimates were used to assess where in the phylogeny of Hymenoptera, when in their geological history, and how often major evolutionary transitions took place. Specifically, we studied the switch from feeding on plants to feeding on an insect host (parasitoidism), the formation of a wasp waist, the evolution of a venomous stinger to subdue mobile hosts, the evolution of eusociality, and the switch from hunting prey to collecting pollen. These evolutionary transitions are partially reflected by the historic classification of Hymenoptera: sawflies (“Symphyta”) are those Hymenoptera that lack the wasp waist that characterizes all remaining Hymenoptera (Apocrita), “Parasitica” encompasses the primarily parasitoid Apocrita that lack a stinger, and Aculeata comprises the stinging wasps, ants, and bees (Anthophila) [1]. Yet, how many major lineages each of these groups encompasses has been controversial for decades [4–11].

The results of our phylogenomic study received strong support in all analyses, unless stated otherwise, and alter previous ideas regarding the evolutionary history of Hymenoptera (Figure 1B; for full results and detailed experimental procedures, see Figure S1, Supplemental Experimental Procedures, and additional figures deposited at Mendeley Data, http://dx.doi.org/10.17632/s5j2f62z3d.2). According to our analyses, extant Hymenoptera started to diversify between the Carboniferous and the Triassic (95% confidence interval [CI]: 329–239 million years ago [mya]; mean: 281 mya; node 1 [n.1] in Figure 1B), with the oldest currently known Hymenoptera fossils being from the Triassic, ~224 million years old [8]. Previous studies suggested this divergence to have occurred between the sawfly lineage Xyeloidea and the remaining Hymenoptera [5, 7–11], whereas our analysis identified a much more inclusive clade of sawflies (Eusymphyta; n.2) that also contains Pamphilioidea and Tenthredinoidea as closest relatives of all remaining Hymenoptera (Unicalcarida). These superfamilies had been thought to form a paraphyletic grade [5, 7, 9, 11]. Instead, they represent an unexpected and previously unrecognized major radiation of primarily ectophytophagous insects that comprises more than 7,000 described species [1]. We estimate the first diversification of the extant eusymphytan lineages to have occurred 276–157 mya (mean 212 mya). Note that Eusymphyta were corroborated as the sister group of all remaining Hymenoptera when additionally scrutinizing the analyzed molecular data for conflicting phylogenetic signal (Supplemental Experimental Procedures). Given the novelty and importance of our finding, we anticipate that it will significantly influence future research on Hymenoptera relationships, and we encourage researchers to further assess this particular phylogenetic hypothesis in future studies, for example by extending the taxon sampling within Eusymphyta and the outgroup.

A clade Eusymphyta representing the extant sister lineage of all remaining Hymenoptera (Unicalcarida) has profound consequences for inferring ground-plan characters of Hymenoptera. For example, Hymenoptera were previously thought to have been ancestrally ectophytophagous, based on the assumption that eusymphytans form a paraphyletic assemblage. Considering that the sister group of Hymenoptera (Aparaglossata) was ancestrally likely predacious [12], the inferred relationship between Eusymphyta and Unicalcarida implies that the most recent common ancestor of Hymenoptera could have been ecto- or endophytophagous. A sister group relationship between Eusymphyta and Unicalcarida furthermore implies that the remarkable ability of male Hymenoptera to restore diploidy in their muscle cells was already present in the last common ancestor of all Hymenoptera (with a secondary loss in Xyelidae), or that this feature evolved at least twice (in Unicalcarida and Tenthredinoidea) [13]. Finally, the unexpected finding that the turnip sawfly, *Athalia rosae* (Tenthredinoidea), whose genome has recently been sequenced by the i5K initiative [14], is a representative of the sister lineage of all remaining Hymenoptera will improve our understanding of the genetic composition of the most recent common ancestor of Hymenoptera: genomic features shared between the turnip sawfly and species of Unicalcarida with sequenced genomes (e.g., *Nasonia* parasitoid wasps, ants, bees) were likely inherited from their common ancestor.

In agreement with earlier studies [9, 10], we found a single origin of the endophytic sawfly lineages (i.e., Cephoidea, Orussoidea, Siricoidea, and Xiphydrioidea; n.3), which form a paraphyletic grade, in which Orussoidea (parasitoid woodwasps) represent the closest relatives of Apocrita (n.4). Morphological data have suggested a sister group relationship of Orussoidea and Apocrita (Vespina) [6, 15], but results from analyzing molecular data have been inconsistent [7, 9]. Our analyses provide strong support for the monophyly of Vespina and of Apocrita (n.5) and imply that the bulk of primarily parasitoid wasps are descendants of a single endophytic parasitoid ancestor that lived in the Permian or in the Triassic (CI: 289–211 mya; mean: 247 mya). Contrary to earlier hypotheses of sawfly relationships (see [10]), we identified Cephoidea, and not Siricoidea and/or Xiphydrioidea, as the closest extant relatives of Vespina (n.6), a result only recently suggested [7].

The evolution of the wasp waist, a constriction between the first and the second abdominal segment greatly improving the maneuverability of the abdomen’s rear section, including the ovipositor, was a major innovation in the evolution of Hymenoptera that undoubtedly contributed to the rapid diversification of Apocrita (n.5) [6]. Our analysis is the first to persuasively demonstrate that the most diverse parasitoid wasp lineages (i.e., Ceraphronoidea, Ichneumonoidea, and Proctotrupomorpha) constitute a natural group (Parasitoida; n.7) whose astonishing radiation was likely triggered by further optimization of the parasitoid lifestyle and related traits (e.g., endoparasitoidism,

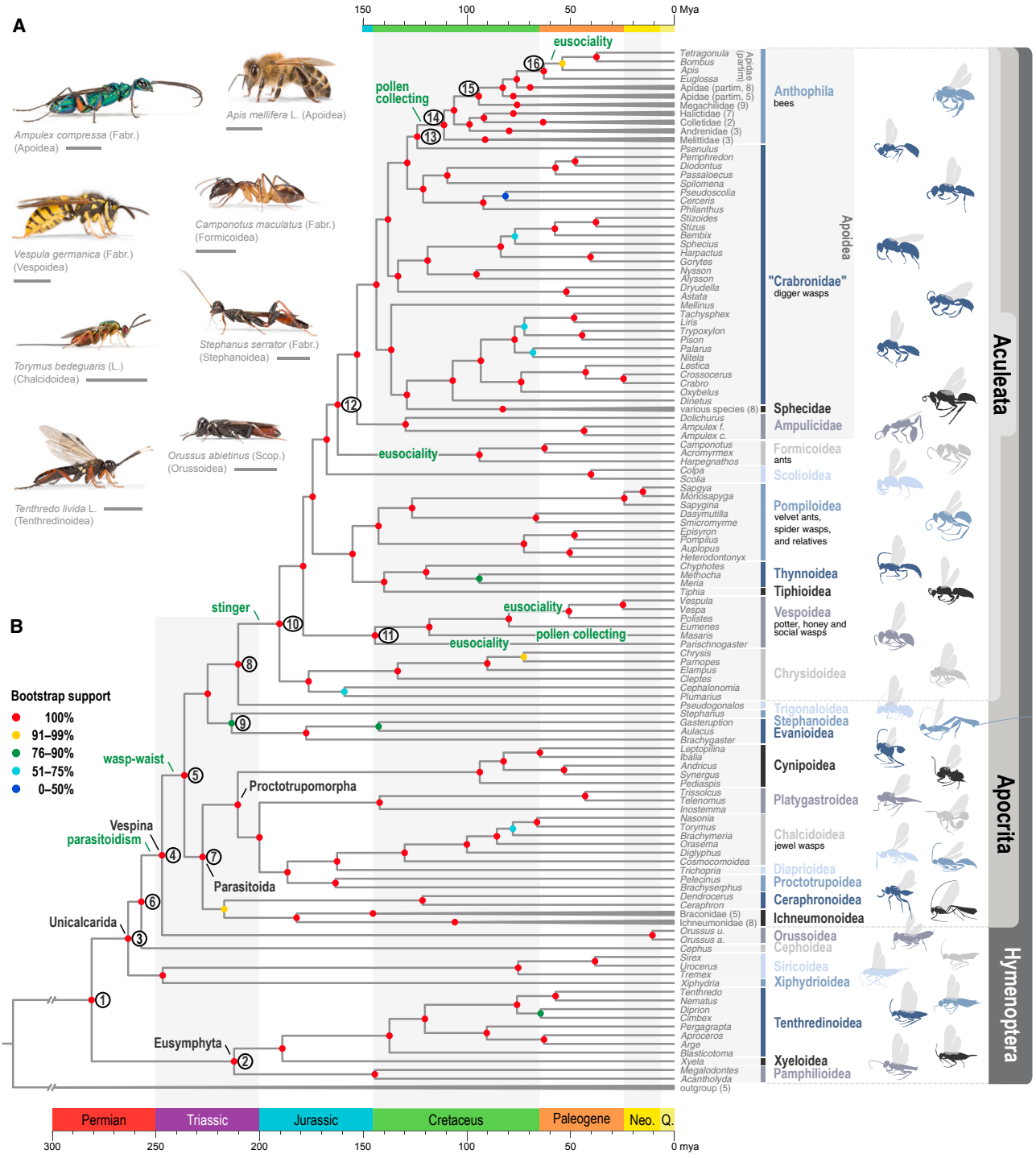

Figure 1. Evolutionary History of the Hymenoptera

(A) Representatives of sawflies, wasps, ants, and bees. Scale bars represent 5 mm.

(B) Phylogenetic relationships and divergence time estimates of Hymenoptera. Key evolutionary events are indicated at the respective clades (note that only the major eusocial lineages are considered). The tree was inferred under the maximum-likelihood optimality criterion, analyzing 1,505,514 amino acid sites and applying a combination of protein domain- and gene-specific substitution models. Divergence times were estimated with an independent-rates molecular clock approach and considering 14 validated fossils. Triangular branches cover multiple species (number of species in parentheses) whose relationships are shown in detail in Figure S1. Nodes with circled numbers are referred to in the main text.

miniaturization), which allowed for successfully attacking a variety of new hosts. We estimate the beginning of the group's radiation at 266–195 mya (mean: 228 mya), only a few million years after Parasitoida separated from the remaining Apocrita (CI: 276–203 mya; mean: 236 mya). The early radiation of Parasitoida thus falls within a time period when the parasitoids' major host lineages (e.g., Hemiptera, Holometabola) also started to diversify [16].

We identified the enigmatic Trigonaloidea as the closest extant relatives of Aculeata with strong node support (n.8), a hypothesis only recently put forth [7, 9]. Evanioidea, which had also been discussed as a possible sister group of Aculeata [5, 10, 17, 18], cluster with Stephanoidea (n.9). Node support for this relationship is low, however, and it needs to be investigated further in future studies that include additional types of characters and samples of Megalyroidea, a lineage that we were unable to sequence. Note that in contrast to Aculeata, the Evanioidea, Stephanoidea, and Trigonaloidea have all remained species-poor. The identification of the closest relatives of Aculeata will be important for better understanding which traits (e.g., venoms) fostered the diversification of the stinging wasps.

Our analysis sheds new light on the phylogeny of Aculeata (n.10), whose early diversification occurred 224–160 mya (mean: 190 mya). Chrysidoids are confirmed as the sister group of all remaining Aculeata [19]. We corroborate the artificial nature of the former superfamily "Vespoidea" (i.e., all Aculeata except Apoidea and Chrysidoidea) [5], which comprises four major lineages that are paraphyletic with respect to Apoidea [20]. The potter, honey, and social wasps (Vespoidea sensu Pilgrim et al. [20]: Vespidae; n.11) were identified as the sister lineage of all remaining non-chrysidoid Aculeata. However, the phylogenetic position of the species-poor Rhopalosomatidae (Vespoidea sensu Pilgrim et al. [20]), an aculeate wasp family that we were unable to sequence and possible sister lineage of Vespidae, remains controversial [9, 10, 20]. The inferred phylogenetic relationships within Vespidae suggest two independent origins of eusociality, a previously fiercely contested hypothesis [21, 22]. In agreement with an earlier phylogenomic study [23], we inferred ants (Formicoidea) as being the closest extant relatives of Apoidea (n.12) in all of our analyses, except when applying a Bayesian approach, which suggested ants plus scoliid wasps (Scolioidea, possibly including also the family Bradynobaenidae [20], which we were unable to sequence) as being sister to Apoidea (figure deposited at Mendeley Data, http://dx.doi.org/10.17632/s5j2f62z3d.2). We estimate the last common ancestor of ants and Apoidea to have lived in the Jurassic or the Cretaceous (CI: 192–136 mya; mean: 162 mya).

We located the phylogenetic origin of bees (Anthophila) within the apoid wasp family "Crabronidae" (n.13), which our study shows to be an artificial construct comprising five major lineages. The crabronid wasp lineage in our study most closely related to bees is the species-poor tribe Psenini. This result substantiates the idea that the switch from a predatory to a herbivorous lifestyle was a key to the tremendous diversification of bees [24]. We estimate the origin of bees to have been in the Cretaceous (CIs: 147–93 mya; means: 124 and 111 mya), a result that is consistent with a close temporal link between the diversifications of bees and angiosperms [24]. Melittid bees were identified as the sister lineage of all remaining Anthophila (n.14), which implies that short-tongued bees do not represent a natural group. In contrast, we confirmed long-tongued bees (i.e., Apidae and Megachilidae) to constitute a natural entity (n.15) [24]. We also found the eusocial apid bee lineages to be monophyletic, corroborating the hypothesis that eusociality has evolved once, not twice, in corbiculate (pollen basket) bees (n.16) [25].

Our study confirms the power of phylogenomic approaches for deciphering difficult-to-resolve arthropod phylogenetic relationships [12, 16, 26, 27] by yielding well-supported answers to some of the most pressing questions regarding the evolutionary history of the sawflies, wasps, ants, and bees. We provide strong evidence for understanding the phylogenetic relationships among all major lineages of Hymenoptera, and we were able to date the individual divergence events, both paramount for deciphering the tempo and mode of diversification of ecologically, economically, sociobiologically, and/or pharmaceutically relevant traits of interest (e.g., gene repertoires, haplodiploidy and sex determination, eusociality, chemosensation, and venoms). Finally, our study offers the basis for establishing a natural classification of the insect order Hymenoptera.

EXPERIMENTAL PROCEDURES

We sequenced the transcriptomes of 134 species of Hymenoptera using Illumina HiSeq 2000 sequencing technology (Data S1A–S1C). We complemented our dataset by including previously published transcriptomes of 29 Hymenoptera and four Neuropteroidea [16, 28]. Finally, we considered the official gene sets of five Hymenoptera and the flour beetle *Tribolium castaneum* (Data S1D). All paired-end reads were assembled with SOAPdenovo-Trans-31kmer (version 1.01) [29], the assembled transcripts were filtered for possible contaminants, and the raw reads and filtered assemblies were submitted to the NCBI SRA and TSA archives. We searched the assemblies with the software Orthograph (version beta4) [28] for transcripts of 3,260 protein-coding genes that the OrthoDB v7 database [30] suggested to be single-copy in Hymenoptera and Neuropteroidea (outgroup) by applying the best reciprocal hit criterion. Orthologous transcripts were aligned with MAFFT (version 7.017) [31] at the translational (amino acid) level. All multiple sequence alignments (MSAs) were quality assessed and, if necessary, improved and masked using the procedure outlined by Misof et al. [16]. The resulting MSAs were concatenated to a supermatrix that we simultaneously partitioned based on a combination of Pfam protein domains and genes [16]. The phylogenetic information content of each partition was assessed with MARE (version 0.1.2-rc) [32], and all uninformative partitions were removed. We subsequently used PartitionFinder (developer versions 2.0.0-pre2, 2.0.0-pre9, and 2.0.0-pre10) [33] to simultaneously infer a partition scheme and proper amino acid substitution models for analyzing each partition with the rcluster algorithm. We applied the same partition scheme when analyzing the corresponding supermatrix at the transcriptional (nucleotide) level, except that we modeled the first and second codon position of each partition separately (note that we excluded the hypervariable third codon position from our analyses). Phylogenetic trees were reconstructed with ExaML (versions 3.0.15 and 3.0.17) [34], conducting 50 independent tree searches per supermatrix. Node support was inferred with the bootstrap method [35]. Decisive datasets were used for testing the possible impact of missing data at the partition level on the inferred phylogenetic tree [36], and four-cluster likelihood mapping was used for assessing the phylogenetic signal for alternative phylogenetic relationships [37]. Permutation tests allowed assessing the impact of heterogeneous amino acid sequence composition, non-stationarity of substitution processes, and non-random distribution of missing data on the inferred phylogenetic tree [16]. We additionally conducted phylogenetic inferences in a Bayesian framework, using ExaBayes [38] with its default settings, enabling automatic substitution model detection and applying the same data partitioning scheme that we used in analyses under the maximum-likelihood optimality criterion. We analyzed three independent runs with four coupled Markov chain Monte Carlo

chains and 200,000 generations each. The consense tool (part of the ExaBayes software package) was used to obtain a consensus tree based on the extended majority rule method (MRE), discarding the first 25% of the sampled topologies as burn-in. Divergence times were calibrated using 14 fossils (Data S1F), selected following best-practice recommendations [39] and representing extant lineages distributed across the entire Hymenoptera Tree of Life. Divergence times were estimated with mcmctree in conjunction with codeml (both part of the PAML software package, version 4.9) [40]. We analyzed a subset of the amino acid and of the nucleotide supermatrix, both comprising only sites that had amino acids or nucleotides present in at least 95% of the species, both with an independent-rates model and with a correlated-rates model (Figure 1B; Data S1H) and sampling parameters previously assessed for convergence of results.

Data Resources

Data reported in this paper have been published in Mendeley Data and are available at http://dx.doi.org/10.17632/trbj94zm2n.2 (inferred matrices and statistics) and http://dx.doi.org/10.17632/s5j2f62z3d.2 (figures). All sequencing data are available at NCBI via the Umbrella BioProject accession number NCBI: PRJNA183205 ("The 1KITE project: evolution of insects").

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure, Supplemental Experimental Procedures, and one dataset and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.01.027.

AUTHOR CONTRIBUTIONS

B.M., L.K., O.N., and R.S.P. conceived the study. C.P., J.H., K.M., K.M.K., L.K., O.N., P.D., R.M., R.S.P., S.K., and T.S. collected or provided samples. A.D., K.M., L.P., O.N., R.S.P., S.L., and X.Z. sequenced, assembled, and processed the transcriptomes. A.K., C.M., K.M., M.P., O.N., R.L., and R.S.P. phylogenetically analyzed the transcriptomes. J.R., L.K., O.N., R.S.P., S.G., and T.W. are responsible for the dating of the inferred phylogeny. All authors contributed to the writing of the manuscript, with L.K., O.N., and R.S.P. taking the lead.

ACKNOWLEDGMENTS

The presented data are the result of the collaborative efforts of the 1KITE consortium. The sequencing and assembly of the 1KITE transcriptomes were funded by BGI through support to the China National GeneBank. We thank S. Blank, A. Dorchin, J. Gusenleitner, V. Mauss, C. Schmid-Egger, and M. Schwarz for help with identification of samples and R. Allemand, E. Altenhofer, E. van den Berghe, A. Blanke, J. de Boer, J. Chille, A. Dorchin, T. Eltz, M. Fierke, R. Glatz, K. Kantner, M. Kivan, K. Kraaijeveld, S. Leonhardt, M. Neumann, M. Niehuis, G. Reder, K. Riede, M. Shaw, N. Schiff, K.-H. Schmalz, J. Schmidt, P. Schüle, K. Schütte, J. Steidle, N. Szucsich, D. Tagu, and D. Yeates for providing valuable samples. We are grateful to S. Brown, D. Gilbert, J. Liebig, and R. Waterhouse for providing information required for the transcript orthology prediction. We thank V. Achter, S. Bank, D. Bartel, A. Böhm, H. Escalona, O. Hlinka, T. Pauli, S. Simon, A. Stamatakis, V. Winkelmann, the Cologne High Efficient Operating Platform for Science (CHEOPS) at the Regionales Rechenzentrum Köln (RRZ), and the CSIRO HPC Dell PowerEdge M620 Linux Cluster Systems for computing time and/or bioinformatic support. We furthermore acknowledge the Gauss Centre for Supercomputing e. V. for funding computing time on the GCS Supercomputer SuperMUC at the Leibniz Supercomputing Centre (LRZ). We thank A. Stamatakis for implementing the four-cluster likelihood quartet mapping feature in ExaML. We acknowledge the Amt für Umwelt, Verbraucherschutz und Lokale Agenda of Bonn, Hessen Forst, the Israeli Nature and National Parks Protection Authority, the Mercantour National Park Service, and the Struktur- und Genehmigungsbehörde Süd and the Struktur- und Genehmigungsdirektion Nord (both Rhineland Palatinate) for granting permission to collect samples. K.M. thanks O. Hlinka (IM&T), D. Yeates (CSIRO), and the Schlinger Endowment to the CSIRO National Research Collections Australia for support. H. Goulet and the Natural History Museum (London) kindly granted permission to use published drawings as blueprints for illustrating our main figure. U. Vaartjes kindly helped with illustrating the main figure. A.D., B.M., C.M., J.R., L.P., M.P., O.N., R.S.P., and S.G. were supported by the Leibniz Graduate School for Genomic Biodiversity Research. C.P. was supported by the Universidad de Castilla-La Mancha and the European Social Fund (ESF). O.N. and T.S. acknowledge the German Research Foundation (DFG) for supporting parts of this study (NI 1387/1-1; SCHM 2645/2-1).

Received: September 24, 2016
Revised: December 13, 2016
Accepted: January 16, 2017
Published: March 23, 2017

REFERENCES

1. Grimaldi, D.A., and Engel, M.S. (2005). Evolution of the Insects (Cambridge University Press).
2. Aguiar, A.P., Deans, A.R., Engel, M.S., Forshage, M., Huber, J.T., Jennings, J.T., Johnson, N.F., Lelej, A.S., Longino, J.T., Lohrmann, V., et al. (2013). Order Hymenoptera. Zootaxa *3703*, 51–62.
3. Quicke, D.L.J. (1997). Parasitic Wasps (Chapman & Hall).
4. Dowton, M., and Austin, A.D. (1994). Molecular phylogeny of the insect order Hymenoptera: apocritan relationships. Proc. Natl. Acad. Sci. USA *91*, 9911–9915.
5. Sharkey, M.J. (2007). Phylogeny and classification of Hymenoptera. Zootaxa *1668*, 521–548.
6. Vilhelmsen, L., Mikó, I., and Krogmann, L. (2010). Beyond the wasp-waist: structural diversity and phylogenetic significance of the mesosoma in apocritan wasps (Insecta: Hymenoptera). Zool. J. Linn. Soc. *159*, 22–194.
7. Heraty, J., Ronquist, F., Carpenter, J.M., Hawks, D., Schulmeister, S., Dowling, A.P., Murray, D., Munro, J., Wheeler, W.C., Schiff, N., and Sharkey, M. (2011). Evolution of the hymenopteran megaradiation. Mol. Phylogenet. Evol. *60*, 73–88.
8. Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D.L., and Rasnitsyn, A.P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. *61*, 973–999.
9. Sharkey, M.J., Carpenter, J.M., Vilhelmsen, L., Heraty, J., Liljeblad, J., Dowling, A.P.G., Schulmeister, S., Murray, D., Deans, A.R., Ronquist, F., et al. (2012). Phylogenetic relationships among superfamilies of Hymenoptera. Cladistics *28*, 80–112.
10. Klopfstein, S., Vilhelmsen, L., Heraty, J.M., Sharkey, M., and Ronquist, F. (2013). The hymenopteran tree of life: evidence from protein-coding genes and objectively aligned ribosomal data. PLoS ONE *8*, e69344.
11. Malm, T., and Nyman, T. (2015). Phylogeny of the symphytan grade of Hymenoptera: new pieces into the old jigsaw (fly) puzzle. Cladistics *31*, 1–17.
12. Peters, R.S., Meusemann, K., Petersen, M., Mayer, C., Wilbrandt, J., Ziesmann, T., Donath, A., Kjer, K.M., Aspöck, U., Aspöck, H., et al. (2014). The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol. Biol. *14*, 52.
13. Aron, S., de Menten, L., Van Bockstaele, D.R., Blank, S.M., and Roisin, Y. (2005). When hymenopteran males reinvented diploidy. Curr. Biol. *15*, 824–827.
14. i5K Consortium (2013). The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J. Hered. *104*, 595–600.
15. Rasnitsyn, A.P., and Quicke, D.L.J. (2002). History of Insects (Kluwer Academic Publishers).
16. Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science *346*, 763–767.

17. Peters, R.S., Meyer, B., Krogmann, L., Borner, J., Meusemann, K., Schütte, K., Niehuis, O., and Misof, B. (2011). The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. *9*, 55.
18. Zimmermann, D., and Vilhelmsen, L. (2016). The sister group of Aculeata (Hymenoptera) – evidence from internal head anatomy, with emphasis on the tentorium. Arthropod Syst. Phylogeny *74*, 195–218.
19. Brothers, D.J. (1999). Phylogeny and evolution of wasps, ants and bees (Hymenoptera, Chrysidoidea, Vespoidea and Apoidea). Zool. Scr. *28*, 233–249.
20. Pilgrim, E.F., Von Dohlen, C.D., and Pitts, J.P. (2008). Molecular phylogenetics of Vespoidea indicate paraphyly of the superfamily and novel relationships of its component families and subfamilies. Zool. Scr. *37*, 539–560.
21. Hines, H.M., Hunt, J.H., O'Connor, T.K., Gillespie, J.J., and Cameron, S.A. (2007). Multigene phylogeny reveals eusociality evolved twice in vespid wasps. Proc. Natl. Acad. Sci. USA *104*, 3295–3299.
22. Pickett, K.M., and Carpenter, J.M. (2010). Simultaneous analysis and the origin of eusociality in the Vespidae (Insecta: Hymenoptera). Arthropod Syst. Phylogeny *68*, 3–33.
23. Johnson, B.R., Borowiec, M.L., Chiu, J.C., Lee, E.K., Atallah, J., and Ward, P.S. (2013). Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. Curr. Biol. *23*, 2058–2062.
24. Cardinal, S., and Danforth, B.N. (2013). Bees diversified in the age of eudicots. Proc. Biol. Sci. *280*, 20122686.
25. Romiguier, J., Cameron, S.A., Woodard, S.H., Fischman, B.J., Keller, L., and Praz, C.J. (2016). Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. Mol. Biol. Evol. *33*, 670–678.
26. Garrison, N.L., Rodriguez, J., Agnarsson, I., Coddington, J.A., Griswold, C.E., Hamilton, C.A., Hedin, M., Kocot, K.M., Ledford, J.M., and Bond, J.E. (2016). Spider phylogenomics: untangling the spider tree of life. PeerJ *4*, e1719.
27. Fernández, R., Edgecombe, G.D., and Giribet, G. (2016). Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. Syst. Biol. *65*, 871–889.
28. Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., and Niehuis, O. (2017). Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics *18*, 111.
29. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics *30*, 1660–1666.
30. Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M., and Kriventseva, E.V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. *41*, D358–D365.
31. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
32. Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K., and Meusemann, K. (2013). Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics *14*, 348.
33. Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. *14*, 82.
34. Kozlov, A.M., Aberer, A.J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics *31*, 2577–2579.
35. Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M., and Stamatakis, A. (2010). How many bootstrap replicates are necessary? J. Comput. Biol. *17*, 337–354.
36. Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., et al. (2014). Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. *31*, 239–249.
37. Strimmer, K., and von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. USA *94*, 6815–6819.
38. Aberer, A.J., Kobert, K., and Stamatakis, A. (2014). ExaBayes: massively parallel bayesian tree inference for the whole-genome era. Mol. Biol. Evol. *31*, 2553–2556.
39. Parham, J.F., Donoghue, P.C.J., Bell, C.J., Calway, T.D., Head, J.J., Holroyd, P.A., Inoue, J.G., Irmis, R.B., Joyce, W.G., Ksepka, D.T., et al. (2012). Best practices for justifying fossil calibrations. Syst. Biol. *61*, 346–359.
40. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects

Daniel Dowling[1,*], Thomas Pauli[1], Alexander Donath[1], Karen Meusemann[1,2,3], Lars Podsiadlowski[4], Malte Petersen[1], Ralph S. Peters[5], Christoph Mayer[1], Shanlin Liu[6,7], Xin Zhou[8,9] Bernhard Misof[1], and Oliver Niehuis[1,*]

[1]Centre for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Bonn, Germany

[2]Evolutionary Biology & Ecology, Institute for Biology I, University of Freiburg, Freiburg (Brsg.), Germany

[3]Australian National Insect Collection, CSIRO National Research Collections Australia, Acton, ACT, Australia

[4]University of Bonn, Institute of Evolutionary Biology and Ecology, Bonn, Germany

[5]Arthropod Department, Zoological Research Museum Alexander Koenig, Bonn, Germany

[6]China National GeneBank, BGI-Shenzen, Shenzhen, Guangdong Province, China

[7]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

[8]Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing, China

[9]College of Food Science and Nutritional Engineering, China Agricultural University, Beijing, China

*Corresponding authors: E-mails: ddowlin@tcd.ie; o.niehuis@zfmk.de.

Accepted: November 23, 2016

Abstract

RNA interference (RNAi) refers to the set of molecular processes found in eukaryotic organisms in which small RNA molecules mediate the silencing or down-regulation of target genes. In insects, RNAi serves a number of functions, including regulation of endogenous genes, anti-viral defense, and defense against transposable elements. Despite being well studied in model organisms, such as *Drosophila*, the distribution of core RNAi pathway genes and their evolution in insects is not well understood. Here we present the most comprehensive overview of the distribution and diversity of core RNAi pathway genes across 100 insect species, encompassing all currently recognized insect orders. We inferred the phylogenetic origin of insect-specific RNAi pathway genes and also identified several hitherto unrecorded gene expansions using whole-body transcriptome data from the international 1KITE (1000 Insect Transcriptome Evolution) project as well as other resources such as i5K (5000 Insect Genome Project). Specifically, we traced the origin of the double stranded RNA binding protein R2D2 to the last common ancestor of winged insects (Pterygota), the loss of Sid-1/Tag-130 orthologs in Antliophora (fleas, flies and relatives, and scorpionflies in a broad sense), and confirm previous evidence for the splitting of the Argonaute proteins Aubergine and Piwi in Brachyceran flies (Diptera, Brachycera). Our study offers new reference points for future experimental research on RNAi-related pathway genes in insects.

Key words: evolution, RNA interference, r2d2, argonaute, dicer.

Introduction

RNA interference (RNAi), also known as RNA silencing, refers to a set of molecular processes in which small RNA (sRNA) molecules (i.e., siRNA, miRNAs, and piRNAs) target and silence or down-regulate the expression of specific nucleic acids (Ha and Kim 2014). The core components of RNAi pathways are Argonaute proteins, which associate with the sRNAs and silence specific target nucleic acids (Meister 2013). The Argonaute and sRNA complex is termed the RNA induced silencing complex (RISC). The RISC uses complementary base pairing of the sRNA to identify the target RNA molecules. Argonaute proteins can silence their targets, certain Argonautes cleave the target mRNA while others affect their targets using alternative mechanisms (Ketting 2011). RNAi pathways differ in number of ways including the exact proteins involved, sRNAs involved, and target RNAs. For instance

the siRNA pathway targets dsRNA of viral origin while the piRNA pathway primarily targets transposons (Meister 2013; Czech and Hannon 2016).

RNAi interference pathways are found throughout eukaryotic organisms and are thought to be present in the last common ancestor of extant eukaryotes. RNAi may have originated as a means of anti-viral defense (Shabalina and Koonin 2008). Other RNAi functions, such as gene regulation, are thought to have evolved later (Shabalina and Koonin 2008). While the basic structure of RNAi pathways and involved proteins are similar throughout eukaryotes, substantial gene duplication and gene loss has occurred in multiple lineages (for examples see: Campbell et al. 2008; Tomoyasu et al. 2008; Jaubert-Possamai et al. 2010; Lewis et al. 2016). In insects, three main RNAi pathways are involved in gene regulation and defense against viruses and transposable elements (Obbard et al. 2009). The origin and evolution of the genes involved in these three pathways is not well documented. Therefore, we screened transcriptome assemblies of 100 insect species for ten core RNAi pathway genes and present the most comprehensive overview of the evolution and distribution of these core RNAi pathways in insects and related arthropods. In addition to the ten core RNAi genes, we also searched for transcripts of *Sid-1*, a gene associated with the systemic spread of RNAi between the cells of *Caenorhabditis elegans* (Winston et al. 2002).

Studies on model organisms show that duplication and loss of core RNAi pathway genes have occurred multiple times. For instance, the number of paralogous genes coding for Argonaute proteins varies throughout eukaryotes: humans have eight genes coding for Argonaute proteins, *Drosophila melanogaster* has five, *Arabidopsis thaliana* has ten, while the nematode *C. elegans* has 26 Argonaute proteins (Hutvagner and Simmard 2008; Siomi and Siomi 2009). This observed duplication of core RNAi pathway genes might be correlated with a diversification (Hutvagner and Simmard 2008) and functional specialization of the RNAi pathways (Mukherjee et al. 2013). In insects, the duplication of core RNAi genes led to three largely separate RNAi pathways, each using different proteins and sRNA molecules (Obbard et al. 2009). Each of the three RNAi pathways has a particular class of sRNAs that associates with a specific Argonaute protein to form a RISC, which targets and silences specific gene expression. The three insect RNAi pathways are briefly outlined below.

(1) The micro-RNA (miRNA) pathway is involved in the regulation of gene expression. miRNA molecules originate in the nuclear genome. Immature miRNAs are processed by the proteins Drosha and Pasha in the nucleus and then exported to the cytoplasm (Ghildiyal and Zamore 2009). In the cytoplasm, the miRNAs are further processed by Dicer1 and its co-factor Loquacious (Ghildiyal and Zamore 2009). The fully mature miRNAs are loaded into Argonaute1 to form the RISC of the miRNA pathway.

(2) The small-interfering-RNA (siRNA) pathway, sometimes referred to as just RNAi, has two functions. The first is a means of anti-viral defense. Here dsRNA of viral origin (produced either inside or outside of the cell) is processed by the protein Dicer2 and the dsRNA binding protein R2D2 into small interfering RNAs (siRNAs) (Meister 2013). Subsequently, the siRNAs are loaded into Argonaute2 to form a RISC, which silences viral gene expression. The second function of the siRNA pathway is as a defense against transposable elements (e.g., transposons) in the genome. The transcribed transposon RNA is processed by Dicer2 and Loquacious (rather than R2D2) to form mature siRNAs (Czech et al. 2008). The siRNAs form RISC with Argonaute2, which silences the expression of transposons to prevent their further transposition in the genome (Czech et al. 2008).

(3) The piwi-interacting RNA pathway is involved in defense against the transposition of transposons in the germline (Siomi et al. 2011). In *Drosophila*, this pathway involves multiple Argonaute proteins of the Piwi sub-clade (i.e., Argonaute3, Aubergine, and Piwi) (Aravin et al 2007). Primary piRNAs are generated through cleavage transposon transcripts by the nuclease zucchini, thereby generating Piwi-interacting RNAs (piRNAs). These primary piRNAs are loaded into the Piwi proteins, resulting in transposon transcripts being further targeted and silenced. This creates a feedback loop, in which the cleavage of a transcript generates secondary piRNAs that target the same transcript (Meister 2013). This is called "the ping–pong amplification loop" (Aravin et al. 2007; Siomi et al. 2011).

RNAi effects were first observed in the 1990s (Napoli et al. 1990) with an explanatory mechanism proposed in 1998 (Fire et al. 1998) (for a historical overview see Sen and Blau 2006). An RNAi system in an organism can be exploited by the experimental introduction of double-stranded RNA. This allows researchers to silence specific genes and elucidate their function (Bellés 2010). Furthermore, RNAi-based technologies have great potential applications as tools for the management, control, and even protection of important insect species (Scott et al. 2013). Further applications of RNAi include novel therapies against disease (Bumcrot et al. 2006) and development of crops that are resistant to pest insects (Baum et al. 2007; Mao et al 2007; Price and Gatehouse 2008; Huvenne and Smagghe 2010). Although experimentally induced RNAi has been shown to silence the target genes in many insect species, the efficacy of RNAi is known to vary significantly between species (Terenius et al. 2011).

Differences of RNAi efficacy among insects could be partially explained by diversity in the RNAi pathway genes present in different lineages. Studies on insects whose genomes have been sequenced show that the number of core RNAi pathway genes varies between different major insect groups, along with gene duplications apparently occurring in several lineages. For example, in mosquitoes multiple *Argonaute* paralogous gene copies have been identified. Both *Aedes aegypti* (two copies of *Argonaute1*) and *Culex pipiens* (two copies of

Argonaute2) have multiple copies of *Argonaute* genes (Campbell et al. 2008). The red flour beetle, *Tribolium castaneum*, also has two paralogs of both, *Argonaute2* and *R2D2* (Tomoyasu et al. 2008). In the pea aphid, *Acyrthosiphon pisum*, multiple copies of miRNA (gene-regulatory) pathway genes have been described (two paralogs each of *Argonaute1, Loquacious*, and *Dicer1*, and four paralogs of *Pasha*) (Jaubert-Possamai et al. 2010). While the genes, proteins, and overall mechanism of the RNAi system are well studied in model insect species, the distribution and evolution of core RNAi genes across the broad scale diversity of insects will be explored in this study.

Material and Methods

Data Used

To infer the distribution, duplication, and loss of core RNAi pathway genes in insects, we screened assemblies of transcriptomes of 100 insect species (supplementary table S3, Supplementary Material online), a subset of the transcriptomes published by Misof et al. (2014) for ten RNAi pathway genes involved in the three main insect RNAi pathways. We selected genes coding for three major protein families involved in insect RNAi: Argonaute proteins, Rnase III proteins, and dsRNA binding proteins. Additionally, we also searched for *Sid-1*, a gene associated with the systemic spread of RNAi between cells. We follow Misof et al. (2014) and use "insect/s" as a synonym for all hexapods, including the orders Protura (coneheads), Diplura (two-pronged bristletails), and Collembola (springtails). We additionally searched the official gene sets (proteins) of seven arthropod species—five insects, one chelicerate, and one crustacean: *Apis mellifera* and *Nasonia vitripennis* (Hymenoptera), *Acyrthosiphon pisum* (Hemiptera), *Bombyx mori* (Lepidoptera), *Tribolium castaneum* (Coleoptera), *Ixodes scapularis* (Chelicerata), and *Daphnia pulex* (crustaceans, Branchiopoda) (supplementary table S1, Supplementary Material online).

We substantiated the hypothesis of R2D2 being a derived feature of pterygote insects by screening the draft genomes of Hrabe's Jmping Bristletail (*Machilis hrabei*; Archaeognatha; https://www.hgsc.bcm.edu/arthropods/hrabes-jumping-bristletail-genome-project; last accessed November 30, 2016) and Silvestri's Northern Forcepstail (*Catajapyx aquilonaris*; Diplura; https://www.hgsc.bcm.edu/arthropods/silvestris-northern-forcepstail-genome-project; last accessed November 30, 2016).

Gene Identification

To identify putative orthologs of the ten RNAi-related pathway genes and *Sid-1*, we first translated the assembled transcripts of each transcript library into all six possible reading frames using the exonerate tool fastatranslate (Slater and Birney 2005; version 2.2). We subsequently used resulting amino acid sequences to create BLAST-searchable databases in Geneious 7.1.5 (Biomatters, Auckland, New Zealand; Kearse et al. 2012). We additionally obtained the official gene sets (protein sets) of seven arthropod species, for which full genomes are available, and generated seven separate BLAST-searchable databases in Geneious (for details, see supplementary table S1, Supplementary Material online). To determine the timing of duplication of Dicer genes we also searched the genomes of two spider species (Sanggaard et al. 2014), the African social velvet spider (*Stegodyphus mimosarum*) and the Brazilian white-knee tarantula (*Acanthoscurria geniculata*), and one centipede (*Strigamia maritima*) (Chipman et al. 2014) for *Dicer* orthologs. To determine if *Sid-1/Tag-130* homologs were present in Diptera we BLAST searched (tBLASTn) the genome assemblies of three dipteran species. Species selected were: *Drosophila pseudoobscura* (GenBank assembly accession: GCA_001014495.1), *Aedes aegypti* (GCA_001014885.1), and *Anopheles gambiae* (GCA_001542645.1).

We used ten amino acid sequences involved in RNAi pathways known from *Drosophila melanogaster* and one amino acid sequence (Sid-1), which is absent in *Drosophila*, but is known from *B. mori* as query sequences (supplementary table S2, Supplementary Material online). All sequences were downloaded from the NCBI protein database. We used each of the eleven amino acid sequences as a query for blastp (BLAST program suite, Altschul et al. 1990) and searched within Geneious against local BLAST databases created from the 100 transcriptomes and the seven official gene sets. We removed false positives (nonorthologous homologs) by searching each hit with blastp against the NCBI nonredundant protein database (http://blast.ncbi.nlm.nih.gov/Blast.cgi; last accessed November 30, 2016). We only considered a transcript to be an ortholog and derived from a given RNAi pathway gene when it was found as best reciprocal hit.

Generation of Gene Trees

All identified amino acid sequences of a given RNAi pathway protein were aligned using the Geneious alignment tool (using the Geneious alignment algorithm; Kearse et al. 2012). We used the deer tick (*I. scapularis*, Chelicerata) and the crustacean branchiopod (*D. pulex*) as outgroups. Short sequences (< 50% of the proteins consensus length) were removed from the alignments. We visually inspected the alignments and manually corrected them for obvious misalignments. For six alignments, we inferred a gene tree applying the maximum likelihood optimality criterion as implemented in PhyML (Guindon et al. 2010; version 3.0) with the following parameters: substitution model: WAG + G, proportion of invariant sites: 0 (fixed), substitution rate categories: 4, alpha-shape parameter: estimated, optimization parameters: topology/length/rate. Statistical tree robustness was assessed in PhyML via bootstrapping (1,000 bootstrap replicates).(supplementary figures S1–S6, Supplementary Material online).

Sid-1/Tag-130 Identification

It has been suggested that putative insect orthologs of *C. elegans* Sid-1 are in fact orthologous with the *C. elegans* protein tag-130 (Tomoyasu et al. 2008). To test this, we recreated the multiple sequence alignment using *C. elegans* Sid-1 and tag-130 amino acid sequences. Using the multiple sequence alignment, we recreated the *Sid-1* gene tree to determine if our putative insect orthologs clustered more closely to *C. elegans* Sid-1 or tag-130. The method used was the same as for the other gene trees.

Ancestral State Reconstruction

To infer gains and losses of orthologs of core RNAi pathway genes throughout insect evolutionary history, we used the ancestral state reconstruction package of Mesquite (Maddison and Maddison 2008; version 3.02). We used the number of genes found in each species as character states and a phylogenetic tree adapted from Misof et al. (2014). The ancestral states were reconstructed using maximum parsimony. Note that Mesquite does not allow ancestral state reconstruction under the Dollo parsimony (Maddison and Maddison 2008) optimality criterion, which penalizes the loss and subsequent regain of a character. Thus, certain figures (supplementary figs. S7–S17, Supplementary Material online) appear to show the loss of gene in one lineage and its subsequent re-evolution in a descendant lineage.

To independently infer contraction and expansion of *Argonaute* genes we used the CAFE 3.0 (Han et al. 2013). As input we used selected *Argonaute1*, *Argonaute2*, *Piwi/Aubergine*, and *Argonaute3* as gene families and provided the number of homologs belonging to each gene family and a ultrametric phylogenetic tree of all species (adapted from Misof et al, 2014). We specified that CAFE 3.0 search for an optimal λ value. We did not specify that λ varies.

Testing for Evidence of Positive Selection in Specific Genes

To determine whether or not positive selection was acting on certain core RNAi pathway genes, we used the package codeML in the program PAML (version 4.8; Yang 2007). codeML calculates the ratio of nonsynonymous substitutions to synonymous substitutions (ω).

We selected two genes to test. The first was *R2D2* in beetles (Coleoptera). Duplicate copies of *R2D2* previously identified in *Tribolium castaneum* were identified in three beetle species. We tested for evidence of positive selection in all branches of the beetle clade comprising *Gyrinus marinus*, *Aleochara curtula*, and *Meloe violaceus* (note that we only detected one copy of *R2D2* in *Lepicerus* sp.).

R2D2 was not found in several Lepidoptera transcriptomes suggesting that it was lost in members of this group. We hypothesized that the double-stranded RNA binding protein Loquacious may fulfill the role of *R2D2* in species which have lost *R2D2*. We tested Loquacious for evidence of positive selection in the branches within the Lepidoptera clade comprising *Nemophora degeerella*, *Yponomeuta evonymellus*, *Zygaena fausta*, and *Parides eurimedes*. Evidence of positive selection in Loquacious in specific branches of Lepidoptera would suggest that it underwent rapid evolution and may taking the role ordinarily taken by *R2D2*.

For both genes (i.e., *R2D2* and *Loquacious*), we generated multiple sequence alignments on the nucleotide level with the amino acid alignments as guidance using PAL2NAL (version 14) (Suyama et al. 2006). We applied a branch site model, in which ω is allowed to vary among both sites and branches, to test for positive selection in specified branches. For both genes we used the gene trees created above as input trees for the codeML analyses. We used two models: one in which ω varies on our branch of interest (alternative model) and one in which ω is fixed for each branch (null models). Models settings for null model were: model = 2, NSsites = 2, fix_kappa = 0, kappa = 2, fix_omega = 1, omega = 1. Model settings for alternative model were: model = 2, NSsites = 2, fix_kappa = 0, kappa = 2, fix_omega = 0, omega = 1. We tested for statistically significant difference between the two models using a Likelihood Ratio Test (LRT) with one degree of freedom.

Transcriptome Completeness Assessment

To assess transcriptome assembly completeness, we used BUSCO version 1.1b (Simão et al. 2015) to search for a set of 2,675 conserved genes that are near-universal single-copy orthologs in arthropods. These genes serve as a benchmark for genome or transcriptome completeness and are found as single copies in the majority (95%) of arthropod genomes in the OrthoDB database (Kriventseva et al. 2015). BUSCO uses a combination of BLAST (Camacho et al. 2009), profile Hidden Markov Models generated with HMMER 3 (Eddy 2011), and a gene model refinement procedure (Stanke et al. 2004) to identify and discriminate genes which are present, duplicated, fragmented, or missing in the searched transcriptome. As transcriptomes only contain a subset of the total genes present in the genome we expect that not all 2,675 BUSCO genes will be found.

Results

Our systematic search for core genes directly involved in RNA interference pathways (five in the miRNA pathway, three in the siRNA pathway, and two in the piRNA pathway) in whole-body transcript libraries of 100 insect species revealed putative orthologs of at least one gene from each of the three RNA silencing pathways in all 32 studied insect orders. We found a complete set of ten genes in 13 of all studied orders. We furthermore found putative orthologs of *Sid-1*, a gene associated with systemic RNAi, in 25 out of the 32 insect orders. Finally, analysis of the 100 transcriptomes indicated gene

duplication and gene loss events in multiple lineages and species (fig. 1). While transcriptomes can be used to identify the RNAi genes, they do not allow us to conclusively state that a gene is missing from the genome.

miRNA Pathway Genes

We identified orthologs of five miRNA pathway genes known from *Drosophila* (Obbard et al. 2009) in our studied insect species: we found *Argonaute1*, *Dicer1*, *Loquacious*, *Drosha*, and *Pasha* in the transcriptomes of 67, 66, 87, 79, and 80 insect species, respectively, representing all major lineages (table 1). Consistent with this observation, ancestral reconstruction using Mesquite (v. 3.02) suggests that all five miRNA pathway genes were present in the last common ancestor of insects. Possible duplicates of *Dicer1* and *Pasha* were found in the transcriptomes of *Planococcus citri* (Hemiptera; two *Dicer 1* and three *Pasha*) and *Essigella californica* (Hemiptera; two *Dicer1* and two *Pasha*) (fig. 1).

siRNA Pathway Genes

Of all three currently known core genes involved in the siRNA pathway of *Drosophila*, we identified orthologs of *Argonaute2*, *Dicer2*, and *R2D2* in the assembled transcripts of 94, 80, and 68 species, respectively (table 1), again representing the major insect lineages. However, we did not find *R2D2* in any of the primary wingless insect (nonpterygote) species. We found possible duplicates of *Argonaute2* in the transcript assemblies of the following species: *Tanzaniophasma* sp. (Mantophasmatodea), *Peruphasma schultei* (Phasmatodea), *Prorhinotermes simplex* (Isoptera), *Xenophysella greensladeae* (Hemiptera), *Pseudomallada prasinus* (Neuroptera), and *Panorpa vulgaris* (Mecoptera).

Table 1

Orthologs of the Members of the Three Different RNAi Pathways Identified in 100 Investigated Insect Transcriptomes (subset of data published by Misof et al. 2014)

Gene	Pathway	Present	Duplicates
Argonaute1	miRNA	67	0
Dicer1	miRNA	66	2
Loquacious	miRNA	87	0
Drosha	miRNA	79	0
Pasha	miRNA	80	2
Argonaute2	siRNA	94	6
Dicer2	siRNA	80	0
R2D2	siRNA	68	3
Aubergine/Piwi	piRNA	89	28
Argonaute3	piRNA	51	1
Sid-1/Tag-130	Systemic RNAi	68	7

NOTE.—The present column shows the number of transcriptomes (out of 100) in which a putative ortholog was found. The duplicates column shows the number of transcriptomes (out of 100) in which more than one putative ortholog for a given gene was identified. For this study, we used assembly version 2 of all transcriptomes, released in October 2015.

We identified two copies of *R2D2* in *Meloe violaceus*, *Aleochara curtula*, and *Gyrinus marinus* (Coleoptera). Ancestral state reconstruction using Mesquite (v. 3.02) suggests that *R2D2* was present in the last common ancestor of Pterygota. Ancestral state reconstruction using CAFE 3.0 indicates that *Argonaute2* was present in two copies in the last common ancestor of insects. Subsequently, in winged insects one copie was lost while in wingless insects *Argonaute2* was duplicated.

piRNA Pathway Genes

The piRNA system of *Drosophila melanogaster* involves three Argonaute proteins of the Piwi family (Argonaute3, Piwi, and Aubergine). We identified both Piwi and Aubergine only in Diptera (three species: *Bombylius major*, *Lipara lucens*, and *Triarthria setipennis*) (fig. 1). Outside of Diptera, we found orthologs of either *Piwi/Aubergine* in the transcript assemblies of 85 species (table 1), representing all major insect lineages. Consistent with this observation, ancestral state reconstruction generated with Mesquite (v. 3.02) suggests that homologs of both *Piwi/Aubergine* and *Argonaute3* were present in the last common ancestor of insects, with *Piwi/Aubergine* present in multiple copies (integers between two and five were equally likely). Ancestral state reconstruction with CAFE 3.0 indicates that two copies of *Piwi/Aubergine* were present in last common ancestor of insects as well as two copies of *Argonaute3*. Furthermore the duplications of *Piwi/Aubergine* in several insect clades (e.g., Diptera and Hemiptera) were suggested to be independent gene expansions. We found multiple copies of *Piwi/Aubergine* in the transcriptomes of 25 nondipteran species (fig. 1). We found orthologs of *Argonaute3* in transcriptome data of 51 species, representing major insect lineages except many polyneopteren groups encompassing Isoptera, Blattodea, Mantodea, Grylloblattodea, Mantophasmatodea, Phasmatodea, and Embioptera. While we found a possible transcript of *Argonaute3* in one species of the insect order Grylloblattodea (ice crawlers), *Grylloblatta bifratrilecta*, the length of the transcript was too short to unambiguously assess orthology. Finally, we found multiple copies of *Argonaute3* in *Anurida maritima* (Collembola).

Systemic RNAi

Phylogenetic analysis of putative insect *Sid-1* orthologs indicates that they form a clade distinct from *C. elegans Sid-1* and *Tag-130*. We also identified *Tag-130* protein domains in many insect putative *Sid-1* orthologs. We identified putative orthologs of *Sid-1/Tag-130* in the transcriptomes of 68 species, representing almost all major insect lineages except species belonging to Antliophora (i.e., Diptera, Mecoptera, and Siphonaptera). We found multiple copies of *Sid-1/Tag-130* in the transcriptomes of 13 species, in particularly in Collembola, with two present in *Sminthurus viridis*, three in *Folsomia candida*, four in *Pogonognathellus* sp., and three in *Anurida maritima*. Multiple copies of *Sid-1/Tag-130* were also

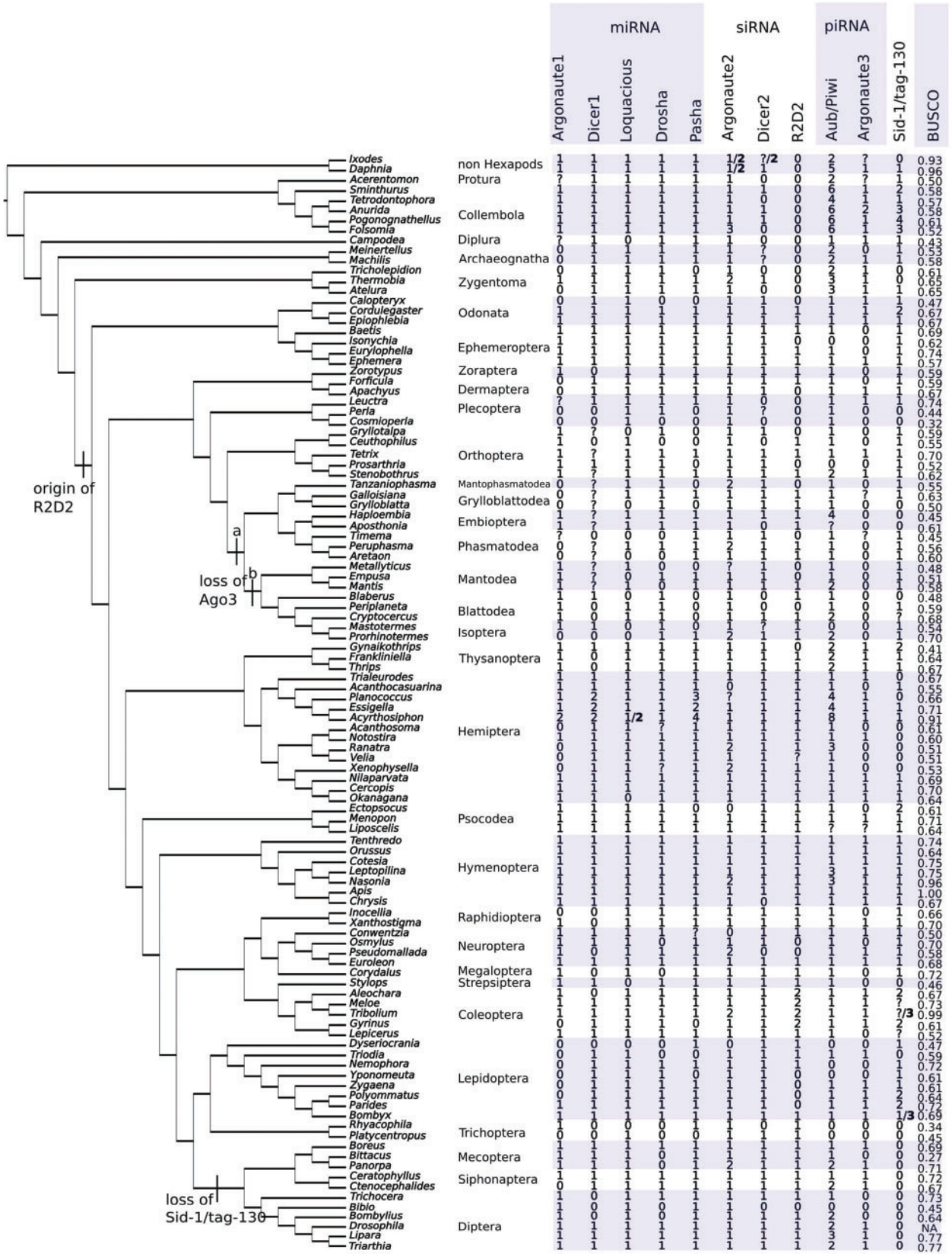

Fig. 1.—Distribution of ten core RNA interference genes and the systemic RNA interference gene *Sid-1* in insects. The number of copies of each gene found using our methodology is noted in the table. Genes whose presence could not be conclusively verified or denied are marked with a question mark (?). Genes which were not found are marked with a zero (0). In some genome species known genes were not recovered. Where this is the case we include the known number of genes in bold after the number we have identified. Tree topology is based on that of Misof et al. (2014). Note that Blattodea is considered paraphyletic.

identified in *Cordulegaster boltonii* (Odonata), *Gynaikothrips ficorum* (Thysanoptera), *Ectopsocus briggsi* (Psocodea), *Aleochara curtula* and *Gyrinus marinus* (both Coleoptera), and *Polyommatus icarus* and *Parides eurimedes* (both Lepidoptera). Ancestral state reconstruction suggests that *Sid-1/Tag-130* was present in the last common ancestor of insects.

Dicer Genes in Other Arthropods

For both spider species we found multiple contigs homologous with insect Dicer proteins (see supplementary material). Both Dicer1 and Dicer2 returned many of the same contigs. Therefore we could not conclusively determine if the spider Dicers were orthologs of insect Dicer1, Dicer2, or orthologous with both. In both spiders, the resulting sequences had a higher identity with Dicer1. In the centipede we found two sequences homologous with insect Dicers. Both sequences were returned as BLAST hits for both Dicer1 and Dicer2 queries. Both hits shared a higher identity with Dicer1.

Evidence of Positive Selection

We found no evidence for positive selection in the two candidate genes *R2D2* and *Loquacious* along any of the investigated branches (i.e., branches within Coleoptera and branches within Lepidoptera). However, it is important to note that evidence for positive selection may have been missed due to the small number of nucleotide sequences analyzed.

Discussion

RNAi is an important biological process in insects (and other eukaryotes) and serves a range of biological functions. Manipulation of RNAi systems is a potentially lucrative field of research with numerous applications. Our results show that the genes of the three major insect RNAi pathways identified in *Drosophila melanogaster* are present in all insect orders. Our analysis indicates that in different insect lineages RNAi-related pathway genes have been duplicated and, in some cases, have potentially been lost more frequently than previously known. Duplications may lead to subfunctionalization or neofunctionalization in RNAi pathways and could explain observed differences in the efficacy of RNAi across different insect groups. Loss of core RNAi-related genes may also explain observed decreases in RNAi efficacy in certain lineages.

Using whole-body transcriptomes of mostly adult insects (supplementary table S3, Supplementary Material online) to detect presence or absence of genes has limitations. As the transcriptome only contains genes expressed at the time of the insect's death (e.g., frozen with liquid Nitrogen), the respective transcriptome may lack genes only expressed at specific developmental stages. Moreover, gene expression restricted to specific tissues could have caused low transcript abundance in whole-body transcriptomes. We therefore cannot distinguish between a gene which may have been lost and one that was not (or very lowly) expressed. Therefore, we also searched for the eleven genes in several published official gene sets (supplementary table S1, Supplementary Material online).

Our results indicate/imply that the evolution of RNAi pathways in insects is a gradual and complex process. Insects inherited a complete RNAi system from their common ancestor and, over time, diversified and expanded this original system. One striking example of this is the evolution of the dsRBP R2D2 in the winged insects. This provided winged insects with two complementary and parallel RNAi pathways—miRNA and siRNA. We infer numerous expansions of argonaute proteins involved in the piRNA pathway in insects. Duplicate copies of *Piwi/Aubergine* were found in 28 of 100 transcriptomes. In comparison, we did not identify any duplicates of *Argonaute1* (argonaute protein of the miRNA pathway) in a single transcriptome. In flies a similar pattern has been observed in which multiple copies of *Piwi/Aubergine* are frequently observed while *Argonaute1* duplications are not (Lewis et al. 2016). As we used transcriptomes we cannot conclusively state that a gene is lost from a species (the gene in question may not have been expressed at the time the transcriptome was generated). However, we do observe several intriguing patterns which suggest that certain components have indeed been lost in specific lineages. One example is *Sid-1/Tag-130* which appears to have been lost in flies and their close relatives (i.e., Antliophora). Another putative loss event is observed in a large clade of hemimetabolous insects, the Dictyoptera (Mantodea, Blattodea, and Isoptera) which appear to have lost *Argonaute3*. Like *Piwi/Aubergine*, *Argonaute3* is involved in the piRNA pathway and its apparent loss poses a curious counter example to the multiple expansions of this pathway observed in other lineages. Our results underscore the diversity of RNAi systems observed in insects and hint at the complex evolutionary histories which must have brought them into being.

Origin of R2D2

The three core proteins of the anti-viral RNAi pathway are Argonaute2, Dicer2, and R2D2. The siRNAs involved in this pathway originate from exogenous dsRNA (e.g., from viruses). The pathway is, therefore, sometimes termed the exo-siRNA pathway. It is the pathway exploited when RNAi is experimentally induced. R2D2 is a double-stranded RNA binding protein (dsRBP) necessary for loading siRNAs into RISC (Liu et al. 2003, 2006). Orthologs of *R2D2* have been identified in several insects including *Drosophila* (Liu et al. 2003), *Tribolium* (Tomoyasu et al. 2008), and the crop pest *Bemisia tabaci* (whitefly) (Uphadhyay et al. 2013). To date, *R2D2* has not been identified outside of insects. We identified orthologs of *R2D2* in all orders of winged insects (Pterygota). However, we

neither found *R2D2* in apterygote insect orders (12 transcriptomes in total) nor in outgroup taxa (*Ixodes scapularis* and *Daphnia pulex*). Ancestral state reconstruction correspondingly suggests that *R2D2* is a derived feature (autapomorphy) of Pterygota. It also suggests that core RNAi proteins duplicated gradually and involved a series of independent gene duplication events rather than a single whole-scale duplication of the RNAi pathway.

While *R2D2* is seemingly absent in primary wingless insects, the other core siRNA pathway genes (i.e., *Argonaute2* and *Dicer2*) are present. Additionally, we could not find *R2D2* in the draft genomes of Hrabe's Jumping Bristletail (*Machilis hrabei*) and Silvestri's Northern Forcepstail (*Catajapyx aquilonaris*). These two genomes have been sequenced and are currently analyzed by researchers of the i5K initiative (i5K Consortium 2013). The absence of R2D2 does not necessarily mean that these species lack a functional exo RNAi pathway. It is possible that the corresponding gene from the miRNA pathway (*Loquacious*) could compensate for R2D2 in these species.

An alternative siRNA pathway (known as the endo-siRNA pathway) involving the proteins Argonaute2, Dicer2, and the dsRBP Loquacious is known from *Drosophila* (Czech et al. 2008; Okamura et al. 2008). This pathway is likely involved in the down-regulation of transposons in somatic cells (Chung et al. 2008). We identified orthologs of Loquacious in all primary wingless insects but *Campodea augens* (Diplura). This suggests that primary wingless insects have a complete siRNA pathway. It remains to be investigated whether or not the siRNA pathway in primary wingless insects involves siRNA of exogenous (e.g., viruses) or endogenous (e.g., transposons) origin or both.

Duplication of *R2D2* has been previously found in *Tribolium* (Tomoyasu et al. 2008). We found evidence of multiple R2D2 homologs in other beetle transcriptomes; however, we were unable to determine if duplication of R2D2 occurred once in beetles or multiple times independently. We tested *R2D2* orthologs in five beetle species to infer evidence of positive selection acting on these genes. While *R2D2* is one of the most rapidly evolving genes in *Drosophila* (Obbard et al. 2006), we did not find any evidence for positive selection in beetles.

R2D2 in Lepidoptera

An *R2D2* ortholog has been identified in the silk moth (*Bombyx mori*). However, it is expressed at very low rates (Swevers et al. 2011). We could not identify *R2D2* in the transcriptomes of four investigated species of Lepidoptera, suggesting that in these species *R2D2* is either expressed at a very low level or is entirely absent. All four investigated species of Lepidoptera belong to the large group of Ditrysia, which includes the vast majority of Lepidoptera, including *B. mori*. The four species belong to four families within Ditrysia (Yponomeutidae, Zygaenidae, Lycaenidae, and Papilionidae). While the number of families investigated is small, they represent the broader diversity of Ditrysia. The consistent pattern observed and the congruency with published results (Swevers et al. 2011) suggests that *R2D2* may be expressed at a low level or is entirely absent in all members of Ditrysia.

The low level of expression of the *R2D2* gene observed in *B. mori* has been suggested as a response to the domestication of this species and subsequent decrease in frequency of viral infection (Swevers et al. 2011). Our results, however, suggest that the R2D2 protein is not (or is generally very lowly) expressed in members of Ditrysia (and, thus, the majority of the Lepidoptera). This implies that loss or low expression of R2D2 significantly predates the domestication of *B. mori*. The possibility that *R2D2* is expressed at low concentrations in Ditrysia may partially explain the variable success observed in experimentally inducing RNAi in Lepidoptera under laboratory conditions (Terenius et al. 2011). It may also have implications for developing RNAi-based crop protections against pest species within Lepidoptera.

Piwi/Aubergine in Diptera

In insects, the piRNA pathway acts as a defense against transposons in the germ line. Unlike in other RNA silencing pathways (miRNA and endo- and exo-siRNA), Dicer proteins are not involved. Additionally, the piRNA pathway uses Argonaute proteins of the Piwi family rather than those of the Ago family (i.e., Argonaute1 and Argonaute2). In the model species *D. melanogaster*, three Piwi proteins (Piwi, Aubergine, and Argonaute3) take part in the piRNA pathway. Argonaute3 and Aubergine operate in a loop (termed the ping–pong amplification loop) which alternately are cleaving sense and anti-sense transcripts. Piwi binds to the resulting piRNAs generated by the loop (Aravin et al. 2007; Siomi et al. 2011). In *Tribolium castaneum*, only two Piwi proteins are present: an ortholog of Argonaute3 and one corresponding to Aubergine/Piwi (Tomoyasu et al. 2008). The mosquitoes *Aedes aegypti* and *Culex pipiens* have large expansions of Piwi proteins with seven and six copies of the *Aubergine* gene, respectively (Campbell et al. 2008). In mosquitoes expansion of *Piwi* genes has been suggested to be a response to increased transposon content in the genome (Campbell et al. 2008).

The split between *Aubergine* and *Piwi* occurred 182–156 million years ago in a common ancestor of brachyceran flies (Lewis et al. 2016). In Brachycera, Piwi plays a role in heterochromatin formation (Chambeyron and Seitz 2014). Our results are consistent with the evidence that the *Piwi/Aubergine* split occurred in the most recent common ancestor Brachycera. We also investigated the transcript assembly of a representative of Bibionomorpha, which are considered to be the closest relatives of the Brachycera, *Bibio marci*, but did not find any orthologs of *Piwi* and *Aubergine* in this species. The BUSCO value of *B. marci* was only 0.45 (all species

mean = 0.6, all species median = 0.61) which suggests a relatively incomplete transcriptome. Because we could not identify orthologs of several other target genes in this species either, this possibly indicates that the transcriptome may have been of inferior quality. Thus, our data are inconclusive in respect of whether the split between *Piwi* and *Aubergine* occurred in the last common ancestor of Brachycera or whether it occurred earlier in the dipteran phylogeny.

In Diptera numerous independent duplications of *Argonaute3* and *Piwi/Aubergine* have also been identified (Lewis et al. 2016). These duplications have been suggested as a response to genomic parasites (e.g., transposons) (Lewis et al. 2016). Our results suggest that the *Piwi/Aubergine* gene has also been duplicated numerous times independently in other insect groups such as Hemiptera, Thysanoptera, and Hymenoptera. Whether this is a response to a high frequency of transposons in the genomes of the analyzed species or whether the duplication has led to new functionality remains to be investigated. The genomes currently sequenced and analyzed in context of the i5K initiative (i5K Consortium 2013) will provide the basis for such investigations.

Loss of Sid-1/Tag-130 in Antliophora

Sid-1 is a transmembrane protein associated with the systemic spread of the RNAi response in the nematode *C. elegans* (Winston et al. 2002). *Drosophila* species lack both orthologs of the gene *Sid-1* and a systemic RNAi response. In other insects, such as *Tribolium*, *Sid-1* like genes have been identified (Tomoyasu et al. 2008). The particular role of the Sid-1 protein in insects, however, remains uncertain. Our analysis could not distinguish if the insect *Sid-1* like genes are orthologous with either *C. elegans Sid-1* or *Tag-130*. We identified orthologs of *Sid-1/Tag-130* in species of most insect orders, but were unable to detect transcripts of *Sid-1/Tag-130* in the analyzed transcriptomes of dipteran species. This corroborates the idea that this gene is absent in flies and relatives. Intriguingly, we did not find orthologs of *Sid-1/Tag-130* in other members of Antliophora (i.e., Mecoptera—scorpion flies in a broader sense—and Siphonaptera—fleas), either. This suggests that *Sid-1/Tag-130* was already lost in the last common ancestor of this species rich endopterygote insect lineage.

Conclusion

Using transcriptomic data of 100 insect species, we have gained new insights into the evolution of RNAi pathways in this highly diverse animal group. We show that RNAi related pathway genes are found in all insect orders. Our results suggest several novel gene expansions and indicate the distribution of core RNAi pathway genes in numerous nonmodel organisms. Additionally, we have identified certain key evolutionary events including the origin of R2D2 in pterygote insects and the loss of *Sid-1* in Diptera.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This manuscript has been enabled by the 1KITE consortium and the i5K initiative. The sequencing and assembly of the 1KITE transcriptomes were funded by the BGI through support of the China National GenBank. Alexander Donath, Bernhard Misof, Oliver Niehuis, Ralph S. Peters, and Lars Podsiadlowski furthermore acknowledge the Leibniz association for installing the graduate school Genomic Biodiversity Research, in which the present study arose. Karen Meusemann acknowledges the Schlinger Foundation for funding. We especially thank Stephen Richards and Richard Gibbs of the Baylor College of Medicine Human Genome Sequencing Center for granting access to i5K pilot data prior to their official publication.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Aravin AA, Hannon GJ, Brennecke J. 2007. The piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764.

Bellés X. 2010. Beyond *Drosophila*: RNAi in vivo and functional genomics in insects. Annu Rev Entomol. 55:111–128.

Baum JA, et al. 2007. Control of coleopteran insect pests through RNA interference. Nat Biotechnol. 25:1322–1326.

Bumcrot D, Manoharan M, Koteliansky V, Sah DWY. 2006. RNAi therapeutics: a potential new class of pharmaceutical drugs. Nat Chem Biol. 2:711–719.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Campbell CL, Black WC, Hess AM, Foy BD. 2008. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. BMC Genomics 18:425.

Chambeyron S, Seitz H. 2014. Insect small non-coding RNA involved in epigenetic regulations. Curr Opin Insect Sci. 1:1–9.

Chipman AE, et al. 2014. The first myriapod genome sequenced reveals conservative gene content and genome organisation in the centipede *Strigamia maritima*. PLoS Biol. 12(11):e1002005.

Chung W, Okamura K, Martin R, Lai EC. 2008. Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. Curr Biol. 18:798–802.

Czech B, Hannon GJ. 2016. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. Trends Biochem Sci. 41(4):324–337.

Czech B, et al. 2008. An endogenous small interfering RNA pathway in *Drosophila*. Nature 453:798–802.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7:e1002195.

Fire A, et al. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811.

Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. Nat Rev Genet. 10:94–108.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):1–37.

Ha M, Kim VN. 2014. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 15:509–524.
Han M, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30(8):1987–1997.
Hutvagner G, Simmard MJ. 2008. Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol. 9:22–32.
Huvenne H, Smagghe G. 2010. Mechanism of dsRNA uptake in insects and potential of RNAi for pest control: a review. J Insect Physiol. 56:227–235.
i5K Consortium. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 104:595–600.
Jaubert-Possamai S, et al. 2010. Expansion of the miRNA pathway in the hemipteran insect *Acrythosiphon pisum*. Mol Biol Evol. 27:979–987.
Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.
Ketting RF. 2011. The many faces of RNAi. Dev Cell 20(2):148–161.
Kriventseva EV, et al. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 43:D250–D256.
Lewis SH, Salmela H, Obbard DJ. 2016. Duplication and diversification of Dipteran Argonaute genes, and the evolutionary divergence of Piwi and Aubergine. Genome Biol Evol. 8(3):507–518..
Liu Q, et al. 2003. R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. Science 301:1921–1925.
Liu X, Jiang F, Kalidas S, Smith D, Liu Q. 2006. Dicer-2 and R2D2 coordinately bind siRNA to promote assembly siRISC complexes. RNA 12:1514–1520.
Maddison WP, Maddison DR. 2008. Mesquite: a modular system for evolutionary analysis. Version 3.02. http://mesquiteproject.org
Mao Y, et al. 2007. Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol. Nat Biotechnol. 25:1307–1313.
Meister G. 2013. Argonaute proteins: functional insights and emerging roles. Nat Rev Genet. 14(7):447–459.
Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.
Mukherjee K, Campos H, Kolaczkowski B. 2013. Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. Mol Biol Evol. 30:627–641.
Napoli C, Lemieux C, Jorgensen B. 1990. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-supression of homologous genes in trans. Plant Cell 2:279–289.
Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Phil Trans R Soc B. 364:99–115.
Obbard DJ, Jiggins FM, Halligan DL, Little TJ. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. Curr Biol. 16:580–585.
Okamura K, Balla S, Martin R, Liu N, Lai EC. 2008. Two distinct mechanisms generate endogenous siRNA from bidirectional transcription in *Drosophila melanogaster*. Nat Struct Mol Biol. 15:581–590.
Price DRG, Gatehouse JA. 2008. RNAi mediated crop protection against insects. Trends Biotechnol. 26:393–400.
Sanggaard KW, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. Nat Commun. 5:3765.
Scott JG, et al. 2013. Towards the elements of successful insects RNAi. J Insect Physiol. 59:1212–1221.
Sen GL, Blau HM. 2006. A brief history of RNAi: the silence of the genes. FASEB J. 20:1293–1299.
Shabalina SA, Koonin EV. 2008. Origins and evolution of eukaryotic RNA interference. Trends Ecol Evol. 23:578–587.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.
Siomi H, Siomi M. 2009. On the road to reading the RNAi code. Nature 457:396–404.
Siomi MC, Sato K, Pezic D, Aravin AA. 2011. Piwi-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol. 12:246–258.
Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.
Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32:W309–W312.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into corresponding codon alignments. Nucleic Acids Res. 34:w609–w612.
Terenius O, et al. 2011. RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. J Insect Physiol. 57:231–245.
Swevers L, Liu J, Huvenne H, Smagghe G. 2011. Search for limiting factors in the RNAi pathway in silkmoth tissues and the Bm5 cell line: the RNA-binding proteins R2D2 and Translin. PLoS One 6:e20250.
Tomoyasu Y, et al. 2008. Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in *Tribolium*. Genome Biol. 9:R10.
Uphadhyay SK, et al. 2013. siRNA machinery in whitefly (*Bemisia tabaci*). PLoS One 8:e83692.
Winston WM, Molodowitch C, Hunter CP. 2002. Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. Science 295:2456–2459.
Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Associate editor: Daniel Sloan

Molecular Phylogenetics and Evolution 116 (2017) 213–226

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Transcriptome and target DNA enrichment sequence data provide new insights into the phylogeny of vespid wasps (Hymenoptera: Aculeata: Vespidae)

Sarah Bank[a,1], Manuela Sann[a,b,1], Christoph Mayer[a], Karen Meusemann[a,b], Alexander Donath[a], Lars Podsiadlowski[c], Alexey Kozlov[d], Malte Petersen[a], Lars Krogmann[e], Rudolf Meier[f], Paolo Rosa[g], Thomas Schmitt[h], Mareike Wurdack[b,h], Shanlin Liu[i,j,k], Xin Zhou[l,m], Bernhard Misof[a], Ralph S. Peters[n,*], Oliver Niehuis[a,b,*]

[a] Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany
[b] Department of Evolutionary Biology and Ecology, Institute for Biology I (Zoology), University of Freiburg, Hauptstraße 1, 79104 Freiburg, Germany
[c] Institute of Evolutionary Biology and Ecology, University of Bonn, An der Immenburg 1, 53121 Bonn, Germany
[d] Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
[e] Department of Entomology, State Museum of Natural History, Rosenstein 1, 70191 Stuttgart, Germany
[f] National University of Singapore, 14 Science Dr 4, Singapore 117543, Singapore
[g] Via Belvedere 8/d, 20044 Bernareggio MI, Italy
[h] Department of Animal Ecology and Tropical Biology, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany
[i] China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen, Guangdong Province 518083, People's Republic of China
[j] BGI-Shenzhen, Shenzhen, Guangdong Province 518083, People's Republic of China
[k] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark
[l] Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, People's Republic of China
[m] Department of Entomology, China Agricultural University, Beijing 100193, People's Republic of China
[n] Center of Taxonomy and Evolutionary Research, Arthropoda Department, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany

ARTICLE INFO

Keywords:
Eumeninae
Raphiglossinae
Zethinae
RNA-seq
Transcriptomics
Phylogenomics

ABSTRACT

The wasp family Vespidae comprises more than 5000 described species which represent life history strategies ranging from solitary and presocial to eusocial and socially parasitic. The phylogenetic relationships of the major vespid wasp lineages (*i.e.*, subfamilies and tribes) have been investigated repeatedly by analyzing behavioral and morphological traits as well as nucleotide sequences of few selected genes with largely incongruent results. Here we reconstruct their phylogenetic relationships using a phylogenomic approach. We sequenced the transcriptomes of 24 vespid wasp and eight outgroup species and exploited the transcript sequences for design of probes for enriching 913 single-copy protein-coding genes to complement the transcriptome data with nucleotide sequence data from additional 25 ethanol-preserved vespid species. Results from phylogenetic analyses of the combined sequence data revealed the eusocial subfamily Stenogastrinae to be the sister group of all remaining Vespidae, while the subfamily Eumeninae turned out to be paraphyletic. Of the three currently recognized eumenine tribes, Odynerini is paraphyletic with respect to Eumenini, and Zethini is paraphyletic with respect to Polistinae and Vespinae. Our results are in conflict with the current tribal subdivision of Eumeninae and thus, we suggest granting subfamily rank to the two major clades of "Zethini": Raphiglossinae and Zethinae. Overall, our findings corroborate the hypothesis of two independent origins of eusociality in vespid wasps and suggest a single origin of using masticated and salivated plant material for building nests by Raphiglossinae, Zethinae, Polistinae, and Vespinae. The inferred phylogenetic relationships and the open access vespid wasp target DNA enrichment probes will provide a valuable tool for future comparative studies on species of the family Vespidae, including their genomes, life styles, evolution of sociality, and co-evolution with other organisms.

* Corresponding authors at: Center of Taxonomy and Evolutionary Research, Arthropoda Department, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany (R.S. Peters). Department of Evolutionary Biology and Ecology, Institute for Biology I (Zoology), University of Freiburg, Hauptstraße 1, 79104 Freiburg, Germany (O. Niehuis).
E-mail addresses: r.peters@leibniz-zfmk.de (R.S. Peters), oliver.niehuis@biologie.uni-freiburg.de (O. Niehuis).
[1] These two authors contributed equally to this work. Authors in alphabetic order.

http://dx.doi.org/10.1016/j.ympev.2017.08.020
Received 21 April 2017; Received in revised form 17 August 2017; Accepted 31 August 2017
Available online 06 September 2017

1. Introduction

Vespids (Hymenoptera: Vespidae) represent a well-characterized group of more than 5000 described species of stinging wasps (Aculeata) (Carpenter, 1982; Brothers and Carpenter, 1993; Pickett and Carpenter, 2010). Most vespid wasp species are solitary and exhibit a predatory lifestyle providing their offspring with larvae of either moths (Lepidoptera), beetles (Coleoptera), or sawflies (Hymenoptera: Tenthredinidae) (Iwata, 1976; Krombein, 1979; Carpenter and Cumming, 1985; Budriene, 2003). Species of the subfamily Masarinae show a behavioral switch to collecting pollen and nectar as food source for their offspring (Gess, 1996). Besides solitary forms, vespids encompass obligatorily and facultatively eusocial species, presocial forms, and social parasites (Crespi and Yanega, 1995; Hunt, 2007; Archer, 2012). These extraordinary behavioral features have fueled many studies on the evolution of sociality within insects, but the basic question, how often eusociality evolved within Vespidae, has still remained controversial due to conflicting hypotheses regarding the phylogenetic relationships among major vespid wasp lineages (*e.g.*, Carpenter, 1982, 2003; Schmitz and Moritz, 1998; Hines et al., 2007; Pickett and Carpenter, 2010).

Phylogenetic inferences regarding vespid relationships have primarily been based on analyzing morphological and behavioral characters (*e.g.*, Carpenter, 1982, 1987, 1988a,b, 1991, 1993, 1996; Carpenter and Cumming, 1985; Carpenter and Rasnitsyn, 1990; Vernier, 1997; Gess, 1998; Krenn et al., 2002; Arévalo et al., 2004; Carpenter and Perera, 2006; Hermes et al., 2013; Perrard et al., 2017). The results from these studies led to the widely accepted recognition of six subfamilies, whose phylogenetic relationships are hypothesized to be as follows: Euparagiinae + (Masarinae + (Eumeninae + (Stenogastrinae + (Polistinae + Vespinae)))) (see Fig. 1, left diagram). According to this system, the three eusocial groups Stenogastrinae, Polistinae, and Vespinae constitute a monophylum, which implies that eusociality evolved only once in the family Vespidae.

The phylogenetic relationships of vespid wasps inferred from molecular sequence data are largely incongruent with those based on morphological and behavioral traits (Schmitz and Moritz, 1998; Hines et al., 2007; Peters et al., 2017), implying two origins of eusociality and challenging the monophyly of the Eumeninae (see Fig. 1, right diagram). Studying DNA sequence data of a mitochondrial and of a nuclear ribosomal gene, Schmitz and Moritz (1998) were the first to show that Polistinae and Vespinae are likely more closely related to Eumeninae than to Stenogastrinae. However, the authors' conclusions were rejected by Carpenter (2003) who argued that a combined analysis of the molecular sequence data with available morphological and behavioral trait information supports the traditional concept of vespid wasp relationships. Yet, Hines et al. (2007) inferred the same phylogenetic relationships as Schmitz and Moritz (1998) by studying a set of four nuclear encoded genes (including one analyzed also by Schmitz and Moritz (1998)) and a significantly improved taxon sample. The conclusions drawn by Hines et al. (2007) were later contradicted by Pickett and Carpenter (2010). In a recent phylogenomic study of all major lineages of Hymenoptera (Peters et al., 2017), Stenogastrinae were proposed to be sister group of the remaining Vespidae. However, as this study included only few representatives of major vespid wasp lineages, it did not assess the phylogenetic position of the enigmatic eumenine tribe Zethini, which Hines et al. (2007) inferred as sister lineage of Polistinae and Vespinae (but see also Pickett and Carpenter, 2010).

Leaving the controversy about the phylogenetic position of Stenogastrinae aside, the phylogenetic relationships inferred by Hines et al. (2007) challenged the concept of monophyletic Eumeninae, the largest vespid wasp subfamily comprising more than 3500 species (Pickett and Carpenter, 2010). The Eumeninae *sensu* Carpenter (1982) (or Eumenidae, as the group was formerly given family status; Richards, 1962) unites the former subfamilies Eumeninae, Raphiglossinae, and Zethinae. Recently, Hermes et al. (2013) conducted a comprehensive phylogenetic study by analyzing morphological characters of species of the above three lineages. The results led the authors to subdivide the subfamily Eumeninae into three tribes: Eumenini (including part of the former Eumeninae), Odynerini (including the remaining part of the former Eumeninae), and Zethini (comprising the former Raphiglossinae and Zethinae). While the results of Hines et al. (2007) are compatible with two monophyletic tribes Eumenini and Odynerini within a subfamily Eumeninae, they argued that Zethini are more closely related to Polistinae and Vespinae than to the remaining Eumeninae (but see also Pickett and Carpenter, 2010). Therefore, Hines et al. (2007) suggested granting Zethini again subfamily status. However, the taxonomic sampling available to Hines et al. (2007) did not include samples of the species-poor former subfamily Raphiglossinae. It thus remained unclear whether these should be included in the subfamily Zethinae.

The lack of a robust phylogeny of vespids and in particular of the subfamily Eumeninae is not only a major obstacle for the stability of the classification of vespid wasps, but also for interpreting the group's evolutionary history. A poor understanding of the vespid wasp phylogenetic relationships makes it furthermore difficult to understand the evolution of those cleptoparasites and parasitoids (*e.g.*, cuckoo wasps; Hymenoptera: Chrysididae) that use vespid wasps as hosts (Kimsey and Bohart, 1991; Wurdack et al., 2015). In the present study, we address the most pressing and unresolved questions regarding phylogenetic relationships within the vespid family and establish a basis for future investigations that rely on a robust phylogeny of Vespidae and its subordinated groups. We seek to achieve this goal by two means: (1) simultaneous phylogenetic analyses of transcript and enriched target nucleotide sequence data of a total of 49 vespid wasp species covering all major lineages, except for Euparagiinae and Gayellini (Masarinae) that we were unable to sequence, to (1a) reassess the hypothesis of eusociality having evolved twice in the family Vespidae and (1b) evaluate the monophyly of the subfamily Eumeninae as well as of its tribes; and (2) design and publish a universal set of baits for enrichment of more than 900 single-copy protein-coding genes from Next Generation Sequencing (NGS) libraries of vespids to foster future in-depth phylogenomic analyses in subordinated vespid wasp lineages.

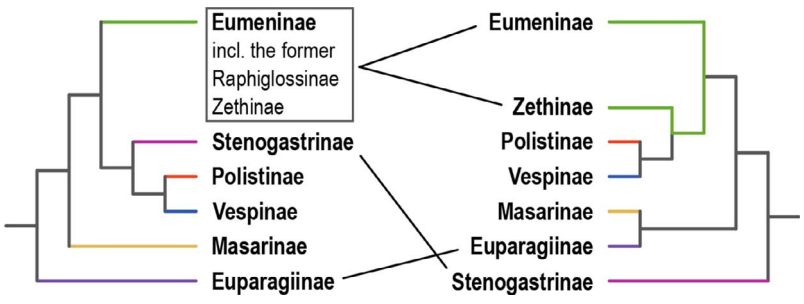

Fig. 1. Conflicting hypotheses on vespid subfamily relationships. The left cladogram was obtained by Carpenter (1982) from studying morphological data, the cladogram on the right was obtained by Hines et al. (2007) from studying molecular data. Carpenter (1982) inferred Stenogastrinae as the sister group to Polistinae + Vespinae and included the former subfamilies Raphiglossinae and Zethinae in the subfamily Eumeninae. Hines et al. (2007) inferred Stenogastrinae as sister group to all remaining Vespidae and found Zethinae to be the sister group of Polistinae and Vespinae. The position of the former subfamily Raphiglossinae remained unclear as they were not included in the study by Hines et al. (2007). Branch color-codes adopted from Hines et al. (2007) indicate subfamilies in the classificatory system of Vespidae proposed by Carpenter (1982): blue (Vespinae), green ("Eumeninae"), pink (Stenogastrinae), red (Polistinae), yellow (Masarinae). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

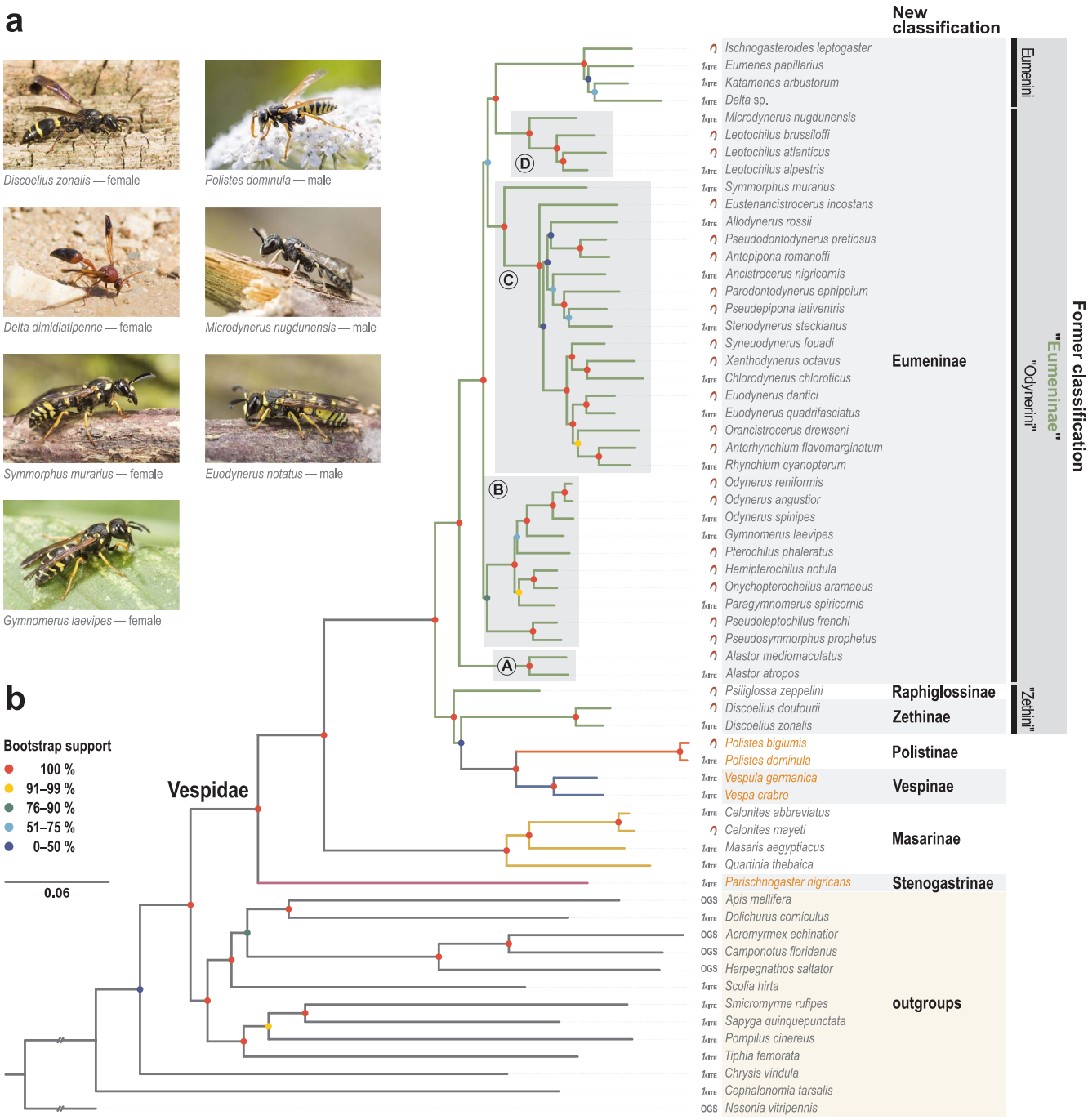

Fig. 2. Vespid wasps and their phylogenetic relationships. (A) Representatives of vespid wasps analyzed in the present investigation. All photographs by O. Niehuis. (B) Phylogenetic relationships of major vespid wasp lineages and proposed changes of the taxonomic classification at the subfamily level. The tree was inferred with ExaML, analyzing transcript (1KITE) and enriched (horseshoe magnet) genomic nucleotide sequences plus corresponding nucleotide sequences from five genome projects (OGS) on the translational level (dataset A1/a; 1,004,596 amino acid sites, 511 partitions, see Section 2.9 and Table 1). Support values are inferred from 150 non-parametric bootstrap replicates. The phylogenetic tree was rooted with *Nasonia vitripennis*. Note that the branches connecting *N. vitripennis* with the rest of the topology have been truncated (//). Capitalized letters (A–D) specify clades referred to in the main text. Species names printed in orange letters indicate that the species is eusocial. Branch color-codes adopted from Hines et al. (2007) indicate subfamilies in the classificatory system of Vespidae proposed by Carpenter (1982): blue (Vespinae), green ("Eumeninae"), pink (Stenogastrinae), red (Polistinae), yellow (Masarinae). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Material and methods

2.1. Taxon sampling and sample preservation

We studied a total of 49 species of vespid wasps representing the subfamilies Eumeninae (40 species, including three of the tribe Zethini and representing both the former Raphiglossinae and the former Zethinae), Masarinae (four species), Polistinae (two species), Stenogastrinae (one species), and Vespinae (two species) (Fig. 2A and B; Supplementary Table 1). Our sampling did not include samples of the vespid wasp lineages Euparagiinae and Gayellini (Masarinae). We also included available transcriptomes of one species of each of the following aculeate wasp families for outgroup comparison (Peters et al., 2017): Ampulicidae, Bethylidae, Chrysididae, Mutillidae, Pompilidae, Sapygidae, Scoliidae, and Tiphiidae (Supplementary Table 1). Finally, we incorporated genomic sequences of the three ant species *Camponotus*

floridanus, *Harpegnathos saltator* (Bonasio et al., 2010), *Acromyrmex echinatior* (Nygaard et al., 2011), the honeybee *Apis mellifera* (Honeybee Genome Sequencing Consortium, 2006), and the jewel wasp *Nasonia vitripennis* (Werren et al., 2010). Note that a recently published study, which became available to us after having had completed our analyses, provided new evidence that the wasp family Rhopalosomatidae, which was not part of our taxonomic sampling, is likely the extant sister lineage of Vespidae (Branstetter et al., 2017; see also Pilgrim et al., 2008).

All wasps were hand-collected with an insect net. Samples collected for enriching the DNA of target genes were preserved and stored in 96% ethanol at −20 °C. Samples collected for transcriptome sequencing were transferred into 2 ml Eppendorf vials containing 0.5 ml of RNAlater (Qiagen GmbH, Hilden, Germany) and were immediately ground with a disposable plastic pestle. Each Eppendorf vial was subsequently filled up to the lid with additional RNAlater and stored at 4 °C for subsequent procedures. Due to the destructive nature of the sample preservation in RNAlater, we preferentially sampled species that are easily identifiable in the field. However, in one instance (*Delta* sp.) the species of a collected sample remained unclear. We exclusively collected adult wasps and focused our sampling on representatives of Central European genera, as their transcriptomes were meant to facilitate future enrichment of target DNA of species occurring especially in this geographic region.

2.2. Transcriptome sequencing, assembly, and contamination check

RNA extraction, NGS library preparation, and sequencing of the prepared libraries on Illumina HiSeq sequencers followed the protocols given by Peters et al. (2017) and were conducted by BGI-Shenzhen (China). All cDNA libraries were paired-end (PE) sequenced on Illumina HiSeq2000 sequencing platforms (Illumina Inc., San Diego, CA, USA) with a read length of 150 base pairs (bp). Per species, we obtained about 2.5 Gbp of raw sequence data.

All raw reads were trimmed, assembled, and screened for possible contaminant sequences (which were then removed) as described by Peters et al. (2017). Both raw reads and the assembled transcriptomes are deposited at the Sequence Read Archive (SRA), respectively the Transcriptome Shotgun Assembly (TSA) of the National Center for Biotechnology Information (NCBI) under the Umbrella BioProject accession PRJNA183205 ("The 1KITE project: evolution of insects") (Supplementary Table 2).

2.3. Identification and alignment of single-copy genes in the sequenced transcriptomes

We identified contigs of putative single-copy genes in the transcriptome assemblies with Orthograph version 0.5.6 (https://github.com/mptrsen/Orthograph/; Petersen et al., 2017). The applied ortholog set comprised 3260 genes listed by OrthoDB version 7 (Waterhouse et al., 2013) to be single-copy in Holometabola. For the orthology identification in Orthograph, we used the official gene sets of six reference species with well-sequenced and annotated genomes (*A. echinatior*, Official Gene Set (OGS) version 3.8, Nygaard et al., 2011; *C. floridanus* and *H. saltator*, each OGS version 3.3, Bonasio et al., 2010; *A. mellifera*, OGS version 3.2, Honeybee Genome Sequencing Consortium, 2006; *N. vitripennis*, OGS version 2.0, Werren et al., 2010; *Tribolium castaneum*, OGS version 3.0, Tribolium Genome Sequencing Consortium, 2008). For details on the ortholog set and the applied Orthograph settings, see Peters et al. (2017). We included all five hymenopterans, whose amino acid and nucleotide sequences were part of the ortholog set in our analyses, while data of the flour beetle *T. castaneum* was only considered when identifying orthologous transcripts. The amino acid and nucleotide sequences of all 37 species whose transcriptomes (32 species) or official gene sets (five species) we exploited were further processed by removing terminal stop codons and masking internal stop codons with 'X' and 'NNN' in the amino acid and nucleotide and sequences, respectively.

The orthologous amino acid sequences of each of the 3260 single-copy genes were aligned with MAFFT version 7.123 (Katoh and Standley, 2013) applying the L-INS-i alignment algorithm. The resulting alignments were checked for outlier amino acid sequences and underwent a refinement procedure described by Misof et al. (2014) except for one difference: when aligning outlier amino acid sequences to respective best matching amino acid sequences of a reference species, we called MAFFT L-INS-i with the "–addfragments" option, since this method is especially suited for aligning short amino acid sequences to an existing alignment. Refined alignments were rechecked for outlier amino acid sequences and remaining outliers were permanently removed from the amino acid alignments as well as from the corresponding nucleotide sequence datasets. We subsequently deleted all gap-only sites (columns) from the resulting amino acid alignments. Finally, we inferred nucleotide sequence alignments from the nucleotide sequence datasets with a modified version of Pal2Nal version 14.1 (Suyama et al., 2006; see Misof et al., 2014 for details on the modification), using the amino acid sequence alignments as blueprints.

2.4. Design of baits for enriching genomic DNA of target genes

In order to enlarge our taxonomic sampling, we not only used 24 vespid wasp transcriptomes, but also included additional 25 ethanol-preserved vespid species from which nucleotide sequence data was sampled by enriching and sequencing a set of 913 single-copy genes. For this purpose, we exploited the nucleotide sequence alignments of the orthologous single-copy protein-coding genes (Section 2.3) to design baits for target DNA enrichment (see Section 2.5). We analyzed the aligned transcript sequences of 23 out of the 24 sequenced vespid wasps with the BaitFisher software, version 1.2.7 (Mayer et al., 2016). Note that the transcriptome of *Paragymnomerus spiricornis* was not yet available when bait design was conducted. Using BaitFisher, aligned transcripts were split into individual coding sequence (CDS) sections using the honeybee gene models (OGS version 3.2) and the corresponding genome assembly (version 4.5) as a guide (Elsik et al., 2014). We specified a bait length of 120 bp to optimize the probes for the SureSelect^{XT2} Target Enrichment System (Agilent Technologies) for enriching target DNA. Based on preliminary results from phylogenetic analyses of amino acid sequence data obtained from 24 transcriptome assemblies, we demanded that the nucleotide sequence of at least one representative of each of the following taxonomic groups (each group is enclosed by parentheses) was present in full length in all candidate bait regions with the length of the tiling design: (*Parischnogaster nigricans*), (*Quartinia thebaica*), (*Masaris aegyptiacus*), (*Celonites abbreviatus*), (*Discoelius zonalis*), (*Polistes dominula*), (*Vespa crabro*, *Vespula germanica*), (*Alastor atropos*), (*Allodynerus rossii*), (*Microdynerus nugdunensis*, *Leptochilus alpestris*), (*Delta* sp., *Eumenes papillarius*, *Katamenes arbustorum*), (*Gymnomerus laevipes*, *Odynerus spinipes*), (*Symmorphus murarius*), (*Ancistrocerus nigricornis*, *Stenodynerus steckianus*), (*Chlorodynerus chloroticus*, *Euodynerus quadrifasciatus*, *Rhynchium cyanopterum*). Depending on the length of the nucleotide sequence alignment suitable for bait design, we designed seven, five, three, or one bait(s) per CDS, with an offset between consecutive baits of 20 bp. Baits were inferred using the heuristic implementation of the unweighted Hamming 1-center DNA sequence search algorithm, specifying a maximum Hamming distance of 0.15 for clustering nucleotide sequences (Mayer et al., 2016).

BaitFisher designs baits at every potential start position of a bait region. The resulting redundancy (*i.e.*, having more bait start positions than required for realizing a specific tiling design at a given locus) was useful, since we used BaitFilter version 1.0.5 (part of the BaitFisher package) to search for and exclude suggested baits that possibly enrich no-target loci. BaitFisher was run with the following options: "-m blast-l –blast-min-hit-coverage-of-baits-in-tiling-stack 0.84 –blast-first-hit-evalue 0.000001". With these options, BaitFilter searched with the aid

of Blast+ software suite version 2.2.29 (Camacho et al., 2009) all potential baits against a reference genome, in this study an early draft genome assembly of the spiny mason wasp, *O. spinipes* (unpublished data). The blast result was used first to remove bait sets at start positions at which not at least one bait of a given bait stack showed a hit coverage of at least 84% with the target sequence in the reference genome. Second, we removed bait sets at start positions if one (or more) bait(s) in the bait set exhibited a significant sequence similarity with more than one position in the reference genome (*i.e.*, the best and second best hit had e-values smaller than 0.000001; see the BaitFisher and BaitFilter manual for more details). Finally, we used BaitFilter in a separate run to choose the optimal bait region among all remaining bait regions. Specifically, we chose the start position within a given CDS region at which the highest number of transcript sequences was available for designing baits. With BaitFisher, we assessed different tiling designs for 120-bp-long baits, since not all CDS regions contain sufficiently long and suitable alignment segments which can host full tiling designs and contain all required taxa: (1) seven baits tiled across 240 bp with an offset of 20 bp between baits, (2) five baits tiled across 200 bp with an offset of 20 bp between baits, (3) three baits tiled across 160 bp with an offset of 20 bp between baits, (4) two baits tiled across 140 bp with an offset of 20 bp between baits, or (5) a single bait. If tiling designs of different lengths fit into a given CDS, we chose the tiling design with the highest number of tiled baits.

2.5. Target DNA enrichment, sequencing, assembly, and contamination check

Genomic DNA (gDNA) was extracted from muscle tissue of 25 vespid wasp species (Supplementary Table 1) using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) and eluted in 100 µl nuclease-free water. Quality and quantity of the extracted gDNA was assessed with a Fragment Analyzer (Advanced Analytical Technologies GmbH, Heidelberg, Germany) and a Quantus Fluorometer (Promega, Fitchburg, Wisconsin, USA) (Supplementary Table 3).

During library preparation, we followed the SureSelect^{XT2} Target Enrichment System Protocol for Illumina Paired-End Multiplexed Sequencing Version E1 published in June 2015 by Agilent Technologies Inc., with some minor modifications. First, gDNA was cut into fragments of 150–400 bp using the Next dsDNase Fragmentase Kit (New England Biolabs Inc., Ipswich, USA) by incubating 100 ng gDNA with 2 µl NEB Next dsDNA Fragmentase and 2 µl 10x Fragmentase Reaction Buffer v2 for 20–25 min. The fragmented gDNA was purified with AMPure XP beads (Beckman Coulter GmbH, Krefeld, Germany) in a ratio of 1:1. Purified fragmented gDNA was subsequently eluted in 30 µl nuclease-free water. The quality and quantity of the fragmented gDNA was assessed by using again a Fragment Analyzer and a Quantus Fluorometer. In the library preparation steps "End Repair", "A-tailing", "Ligation of indexed adapter", and "Pre-amplification of indexed libraries", we reduced the reaction volume specified in Agilent's protocol (pages 43–54 for 100 ng DNA samples) by 50%. For the pre-amplification reaction, we applied the following PCR program: initial denaturation temperature of 98 °C for 2 min, followed by 12 cycles of 30 s at 98 °C, 30 s at 60 °C, and 60 s at 72 °C, followed by a 10 min final extension at 72 °C.

For enriching the target gDNA in the indexed libraries, we continued following the procedure outlined in Agilent's SureSelect^{XT2} Target Enrichment System Protocol for Illumina Paired-End Multiplexed Sequencing Version E1 published in June 2015 (pages 55–74) with minor modifications. Briefly, we used a SureSelect^{XT2} Custom 59.1 Mbp capture library comprising 49,226 different baits (Supplementary Table 4) and pooled the indexed libraries before the hybridization reaction as follows: one pool (A) comprised the libraries of 14 samples [plus two additional ones not included in this study] (~1.5 µg in total), with each library contributing 93 ng. Two additional pools comprised the indexed libraries of six [plus two not included in this study] (B1) and of seven [plus one not included in this study] (B2) samples (each with ~750 ng in total), again with each library contributing 93 ng. For more information, see Supplementary Table 5. The two specimens of the species *Odynerus reniformis* and *Eustenancistrocerus inconstans* were enriched twice, once in a 16-samples pool and once in an 8-samples pool for quality control reasons. After pooling the libraries, the total volume of the pools was reduced to 7.0 µl (pool of 14 [total 16] samples) and 3.5 µl (pools of seven [total eight] and six [total eight] samples) with a SpeedVac R SPD 111V (ThermoFisher Scientific, Waltham, MA; USA). Hybridization with the baits was allowed for 48 h at 65 °C in a GeneAmp PCR System 2720. We then initiated the physical separation of the target DNA fragments from the remaining DNA fragments by adding 50 µl Dynabeads MyOne Streptavidin T1 beads and incubating the mixture for 30 min at room temperature. After washing of the beads, the captured DNA was re-suspended in 30 µl nuclease-free water and post-amplified in an on-bead PCR reaction. For the post-amplification, we followed Agilent's protocol by applying the recommended PCR cycling program for a capture library size of > 1.5 Mb with a slightly increased cycle number: initial denaturation temperature of 98 °C for 2 min, followed by 12 cycles of 30 s at 98 °C, 30 s at 60 °C, and 60 s at 72 °C, followed by a 10 min final extension at 72 °C. We purified the amplicons with AMPure XP beads in a ratio of 1:0.7 to remove oligonucleotide primer dimers and to further select for fragments with a size between 200 and 500 bp. Each of the three processed library pools was eluted in 30 µl nuclease-free water and checked for quality and quantity with a Fragment Analyzer and a Quantus Fluorometer.

The three pools of enriched gDNA libraries were sequenced on an Illumina NextSeq 500 Serious sequencer (Illumina Inc., San Diego, CA, USA) with 150 bp PE generating about 0.7 Gbp of raw data per sample (the total amount of raw data of the twice-sequenced samples, *O. reniformis* and *E. inconstans*, was 1.27 Gbp and 1.54 Gbp, respectively). All obtained raw reads were trimmed with Trimmomatic version 0.35 (Bolger et al., 2014) and *de novo*-assembled with IDBA-UD version 1.1.1 (Peng et al., 2012) as described by Mayer et al. (2016). Finally, we searched all contigs sequenced on the same lane against each other using the program blastn of the Blast+ software suite version 2.2.31 (Camacho et al., 2009) in order to identify possible contaminant contigs. Contigs identified as contaminants were removed following the procedure outlined in Mayer et al. (2016), except that we selected a 10-fold expression difference between contigs rather than a 2-fold difference for distinguishing between contaminants and non-contaminants (see Mayer et al., 2016 for details).

2.6. Post-processing of assembled gDNA sequences

We used Orthograph version 0.5.6 (Petersen et al., 2017) to search the assembled gDNA data for contigs containing sections of enriched target genes. Orthograph concatenates by default contigs referring to different CDS regions of the same gene and provides the predicted amino acid and corresponding coding nucleotide sequences. However, since Orthograph is optimized to process cDNA rather than gDNA sequences, it translates into intronic sequence sections if possible by chance. This can severely bias downstream analyses, because sequences obtained from applying target DNA enrichment could share a small fraction of erroneously predicted amino acid residues (in contrast to sequences obtained from transcriptome sequencing; see Section 2.2). To remove such erroneously predicted sequence sections, we first mapped the predicted amino acid sequences of the target genes onto the aligned amino acid sequences of transcript origin using MAFFT version 7.273 (Katoh and Standley, 2016) applying the L-INS-i alignment algorithm. This was done by selecting the following alignment options: (1) "*–add*" for adding sequence fragments to an existing multiple sequence alignment, (2) "*–keeplength*" to not allow adding gaps to an existing multiple sequence alignment by removing any extra amino acids from the added sequences, and (3) "*–mapout*" to record information about where amino

acids were removed from the added sequences in order to keep the alignment length fixed. We subsequently used the recorded information of how the extra amino acid sequences were mapped onto the amino acid alignments to edit the nucleotide sequences of the enriched exons and remove corresponding codons. Finally, we aligned the corrected nucleotide sequences of the 25 added vespid wasps to the transcript nucleotide sequences with a modified version of Pal2Nal version 14.1 (Suyama et al., 2006; see Misof et al., 2014 for details on the modification), using the amino acid sequence alignments from the preceding step as blueprints. Next, we identified with custom Perl scripts the individual CDS sections in the amino acid sequence alignments, using the honeybee gene models of OGS version 3.2 in the draft genome assembly version 4.5 as a guide (Elsik et al., 2014). We then removed all amino acid residues that were aligned to non-target CDS sections from sequences obtained via target DNA enrichment. We additionally and conservatively removed with custom Perl scripts all amino acid sequence sections covering less than 95% of the honeybee target exon sequence in each multiple sequence alignment from sequences obtained via target DNA enrichment to ensure that no erroneously translated intronic sequence sections, which could bias the phylogenetic analyses, remained in the dataset. The nucleotide sequence alignments were subsequently processed accordingly, using the amino acid sequence alignments as blueprints and custom Perl scripts.

2.7. Enrichment statistics

We calculated the base coverage depth of all full-length or near full-length target exons as well as of the bait-binding sites on each enriched exon by mapping the raw reads onto the respective contig with the software segemehl version 0.2.0 (Hoffmann et al., 2009, 2014). The mapped data was subsequently exploited with SAMtools version 1.2 (Li et al., 2009) to infer base-coverage depth estimates of specific sequence sections (target coding exons and bait-binding sites). We further assessed the extent to which target DNA was enriched, applying the approach suggested by Mayer et al. (2016) for analyzing species with known genome size. While the genome size of none of the enriched species is currently known, we hypothesized that the genome sizes of *P. dominula* (246.3 Mbp; Standage et al., 2016) and *O. spinipes* (197.1 Mbp, inferred by analyzing the k-mer coverage distribution in paired-end sequenced libraries of this species; Niehuis, pers. comm.) are reasonable estimates to those of *Polistes biglumis* and of *Odynerus angustior* and *O. reniformis* for which we estimated the enrichment success. We acknowledge that the genome size even of closely related species can differ. However, significant genome size discrepancies should (in most instances) result in vastly disparate enrichment coefficient estimates when assessing different species, while similar enrichment coefficient estimates would be consistent with the idea of similar genome sizes of these species. Following Mayer et al. (2016), we compared the average base-coverage depth (C_t) of the bait-binding sites on sequenced target exons to the average base-coverage depth (C_g) expected for the sequenced and assembled fragments of a given genome in the absence of enrichment. C_g was calculated by dividing the total number of nucleotides considered for assembling the library of a respective species by the estimated size of the species' genome. Since we applied different tiling designs for enriching target loci, we also investigated whether or not the tiling design had an impact on the base-coverage depth of the bait-binding sites of enriched target exons, using the base-coverage estimates inferred with SAMtools. However, to reduce edge effects (*i.e.*, the base-coverage depth of one exon influencing the base-coverage depth of a flanking target exon), we restricted our analyses to genes for which we enriched only a single coding exon.

2.8. Phylogenetic analyses of transcript sequences and genes from official gene sets

All amino acid alignments were searched for sequence sections showing random similarity or ambiguously aligned residues with Aliscore version 1.2 (Misof and Misof, 2009; Kück et al., 2010). Aliscore was run with default parameters exept for using the '-e' option to cope with transcript sequence alignments containing many gaps (see Meusemann et al., 2010) and the '–r' option set to 10^{27} to compare all sequence pairs in each sliding window.

We decided to apply a protein domain-based partitioning scheme to improve the fit of substitution models for the amino acid and nucleotide sequence data, as suggested by Misof et al. (2014) when studying comparable transcriptome sequence data. We identified protein domains, families, and clans in each predicted (unmasked) transcript alignment on the amino acid level, exploiting information from the protein family databases Pfam-A (release 28; Finn et al., 2014) and Pfam-B (release 27; Finn et al., 2014). Domains were searched for with the aid of PfamScan software version 1.5 (released 2013-10-15, Finn et al., 2014) and HMMER version 3.1b2 (Eddy, 2011) as outlined in Misof et al. (2014) and Peters et al. (2017). The two Pfam databases were separately used to search for domains in the multiple sequence alignment (MSA) of each gene, and the domain with the highest number of hits across all species' sequences in the MSA was selected as the dominant domain. To merge the results of both databases and to avoid overlapping domains, we gave Pfam-A annotations priority over Pfam-B annotations. Please note that we did not consider any of the enriched target gene sequences when searching for protein domains (see Section 2.9). We then merged the coordinates received from the protein domain identification with the information on sites suggested to be removed by Aliscore. We deleted respective sections and concatenated the data blocks into a supermatrix on the amino acid level and generated a corresponding supermatrix on the nucleotide level. During this process, terminal gap symbols ('-') were masked for each data block with 'X' and 'N' in the amino acid and the nucleotide alignments, respectively. All sequence sections were concatenated according to the domain identification as follows: (i) sequence segments identified as Pfam-A domains belonging to the same clan were concatenated to clan-specific data blocks, (ii) sequence segments identified as the same Pfam-A domain (not associated with any clan) were concatenated to Pfam-A domain-specific data blocks, (iii) sequences segments identified as Pfam-B domains were concatenated to Pfam-B domain-specific data blocks, and (iv) sequence segments without any domain annotation were concatenated to the gene-specific data blocks.

The information content within the amino acid supermatrix was evaluated for each data block with the software MARE version 0.1.2-rc (Misof et al., 2013). All data blocks with zero information content were removed. In order to minimize non-random distribution of missing data, we only kept those data blocks that included sequences of each of the 37 species (*i.e.*, 32 species whose transcriptomes we analyzed plus five reference species). We kept the corresponding data blocks from the nucleotide supermatrix.

Having protein domain-based data blocks at hand, we next conducted a two-step heuristic approach to search for both an optimal partitioning scheme and best-fitting substitution models to the inferred partitions. To reduce the complexity of this task, we restricted the search for the best partitioning scheme to a subset of substitution models. Thus, we searched with PartitionFinder version 2.0.0 pre-release 10 (http://www.robertlanfear.com/partitionfinder/; Lanfear et al., 2014, 2016) in combination with RAxML 8.2.4 (Stamatakis, 2014) with the settings '–raxml –weights 1,1,0,1 –rcluster-max 10000 –rcluster-percent 100 –all-states –min-subset-size 50' for the best partitioning of the amino acid supermatrix, allowing only two different substitution models to be used (*i.e.*, LG + G and LG + G + F). Once the best partitioning scheme was found, we assessed in a second step the best fitting model for each partition. This was done with the help of the corrected Akaike information criterion (AICc; Hurvich and Tsai, 1989) and by comparing the fit of the following substitution models: WAG + G, WAG + G + F, BLOSUM62 + G, BLOSUM62 + G + F, DCMUT + G, DCMUT + G + F, JTT + G, JTT + G + F, LG + G, LG + G + F, LG4X. We

used exactly the same partitioning scheme, except that we additionally treated the three codon positions within each partition of the above inferred partitioning scheme as separate partitions, when analyzing the supermatrix at the nucleotide level and applied the GTR + G model to all partitions.

Phylogenetic relationships were inferred by applying the Maximum Likelihood (ML) optimality criterion as implemented in the software ExaML version 3.0.15 (Kozlov et al., 2015). We conducted 50 tree searches: 25 using randomized stepwise addition parsimony starting trees and 25 using completely random starting trees. All starting trees were inferred with RAxML version 8.2.7 (Stamatakis, 2014). The tree with the best log-likelihood score among the 50 evaluated ones was considered to be the best to reflect the phylogenetic hypotheses supported by the analyzed dataset.

We assessed support values for phylogenetic relationships by a partitioned non-parametric bootstrap analysis using a total of 50 (amino acid data set) and 100 (nucleotide data set) bootstrap replicates with ExaML version 3.0.15 (Kozlov et al., 2015). We determined whether or not the number of bootstrap replicates was sufficient for assessing support values for different hypothesis by applying the *a posteriori* bootstopping criterion (Pattengale et al., 2010) implemented in RAxML version 8.2.7 (Weighted Robinson Foulds distance building an extended majority-rule (MRE) consensus tree (autoMRE, threshold [0.03], with 1000 permutations; Stamatakis, 2014)). Bootstrap support values were mapped onto the two inferred best ML phylogenetic trees (one of which is based on the analysis of amino acids [analysis scheme A0], the other one is based on the analysis of nucleotides [analysis scheme N0]), which were subsequently drawn with FigTree version 1.4.3 (Rambaut, 2016) and rooted with the parasitoid wasp *N. vitripennis*. Exported vector graphics were edited with Inkscape version 0.91. Given that FigTree (and other tree visualization software) suffers from a significant software bug resulting in bootstrap support values being assigned to wrong nodes after re-rooting of a tree (Czech et al., 2017), we manually checked all bootstrap support values in the inferred illustrations.

We also searched for rogue taxa in the topologies inferred from analyzing the partitioned amino acid and the partitioned nucleotide sequence data, using the software RogueNaRok version 1.0 (Aberer et al., 2013) with the same wide array of settings as applied and specified by Peters et al. (2017). However, none of the species showed rogue behavior in the phylogenetic analyses.

To assess whether or not the dataset contained conflicting signal that is not obvious from the two inferred phylogenetic trees and to evaluate whether or not confounding signal due to compositional heterogeneity across taxa and/or non-random distribution of missing data (see Dell'Ampio et al., 2014) had an impact on the support of specific phylogenetic hypotheses, we applied the Four-Cluster Likelihood Mapping method (FcLM) on the original amino acid supermatrix as well as on permuted versions of it, following the strategy suggested by Misof et al. (2014). For more information on the approach, please consult Strimmer and von Haeseler (1997), Misof et al. (2014), as well as the legend to Supplementary Fig. 4. Note that we used the LG substitution matrix (Le and Gascuel, 2008) for permuting the supermatrices. FcLM was used to evaluate whether *D. zonalis* (the only representative of the tribe Zethini and of the former subfamily Zethinae whose transcriptome we sequenced), is closer related to Polistinae + Vespinae or to Eumeninae (excl. *D. zonalis*) (Supplementary Table 6). FcLM was done with ExaML version 3.0.17 (Kozlov et al., 2015) on the original amino acid supermatrix, using parsimony start trees, and applying the partitioning scheme and substitution models inferred when analyzing the complete supermatrix at the amino acid level. For the permutation approach, we used the same software and partition scheme, but replaced the original supermatrix with random data inferred with the aid of the LG substitution matrix (we consequently applied LG substitution model across all partitions when analyzing the permuted matrices via FcLM). Results were visualized in simplex graphs, using a custom Perl script.

2.9. Phylogenetic analysis of transcript and enriched target coding sequences

We analyzed nucleotide sequences obtained via target DNA enrichment in conjunction with the orthologous transcript sequences using the transcript sequence alignments (Section 2.3) onto which the gDNA sequences (Sections 2.4 and 2.5) had been mapped (Section 2.6). We generated a supermatrix mirroring the one inferred in Section 2.8 when analyzing the transcriptomic sequences alone. Thus, the supermatrix exhibited exactly the same number of sites and partitions. This was achieved by applying all previously acquired information about what sites in the transcript alignments to remove, mask, and combine (Section 2.8) onto the corresponding alignments containing the additional gDNA sequences. However, since the enriched sequences encompassed only a subset of the single-copy genes that were present (and analyzed) in the transcript sequences, we applied various filtering and modeling schemes to assess the impact of missing data and to improve the model fitting: (1) for a partition to be considered in the phylogenetic analysis, each partition previously inferred from analyzing only the transcript sequences (Section 2.8) had to contain the sequences of all 32 species, whose transcriptome we sequenced plus the sequences of the five reference species (same conditions we demanded when analyzing these species alone; Section 2.8), while no minimum number of sequences was specified for the DNA enrichment dataset (datasets based on amino acids [A1] and nucleotides [N1]). (2) Same conditions as in (1), except that sequence data obtained via DNA enrichment from at least one additional species had to be present in a given partition (datasets A2 and N2). (3) For a partition to be considered in the phylogenetic analysis, it had to contain the sequences of all 62 species, *i.e.*, those whose transcriptome we sequenced plus those of the reference species plus those which we sequenced via target DNA enrichment (datasets A3 and N3).

When analyzing the above three datasets at the amino acid level, we applied two substitution modeling schemes: (a) we applied the same substitution models as we did when analyzing the transcript sequences alone (Section 2.8) (analysis schemes A1/a, A2/a, and A3/a), and (b) we inferred the best fitting substitution model to each partition using PartitionFinder version 2.0.0 prerelease 10 (Lanfear et al., 2016) and testing the same substitution models as listed in Section 2.8 (analysis schemes A1/b, A2/b, and A3/b). We thus conducted a total of six phylogenetic inferences on the amino acid level using the maximum likelihood tree inference method implemented in ExaML (*i.e.*, A1/a, A1/b, A2/a, A2/b, A3/a, and A3/b; Supplementary Fig. 1). Since we consistently applied the GTR + G model when studying the dataset on the nucleotide level, the total number of phylogenetic inferences on the nucleotide level was three (*i.e.*, N1, N2, N3). A summary of the data processing workflow is given in Supplementary Fig. 1. Phylogenetic trees were inferred, branch support was assessed, and rogue taxa were identified in each of the analysis schemes as outlined in Section 2.8. As only species within subordinated lineages, whose relationships to each other we intended to infer, exhibited rogue behavior (*i.e.*, *E. papillarius* and *Ischnogasteroides leptogaster* within the tribe Eumenini; *A. rossii*, *E. incostans*, and *Orancistrocerus drewseni* within clade C of the tribe "Odynerini"; Supplementary Table 7), we refrained from excluding these species in any of our inferences. FcLM, as outlined in Section 2.8, was used to check for conflicting and/or confounding signal when exploring whether *Discoelius* spp. (Zethini and representatives of the former Zethinae) or *P. zeppelini* (Zethini and representative of the former Raphiglossinae) are the closest relatives of Polistinae and Vespinae, analyzing the amino acid datasets A1/a, A1/b, A2/a, A2/b, A3/a, and A3/b (Supplementary Table 8).

To assess the possible impact of the ML tree inference method on the inferred tree topology, we additionally conducted phylogenetic inferences in a Bayesian framework, using the software ExaBayes (Aberer

et al., 2014). We applied this approach exclusively to the datasets A1/a and N1, which contained all compiled sequence information (note that all other datasets represented subsets of the datasets A1/a and N1, and their analysis resulted in virtually identical tree topologies). ExaBayes was run as outlined by Peters et al. (2017), except that we generated Markov chain Monte Carlo chains (four coupled chains in three independent runs) for 1,000,000 generations each when analyzing the dataset A1/a, and for 3,000,000 generations each when analyzing dataset N1. Since one of the three runs got trapped in a local optimum when analyzing dataset N1 which prevented the three runs from converging (average standard deviation of split frequencies [ASDSF] = 12.25%), we additionally sampled trees from a fourth run, which converged with the two previously converging runs (ASDSF = 4.41%). We analyzed only the trees from the three runs that converged. The three runs from analyzing dataset A1/a also converged (ASDSF = 3.97%). While we applied the same data partitioning scheme that we used in ExaML, we enabled automatic substitution model detection when analyzing the amino acid dataset, since ExaBayes does not support the LG4X amino acid substitution model that PartitionFinder suggested to apply on several of the inferred data partitions. Trees were sampled every 500 generations and the first 25% of the sampled trees were discarded (burn-in phase). This resulted in a total of 4500 (dataset A1/a) and 13,500 (dataset N1) sampled trees based from which we calculated posterior probability values.

To assess the possible impact of species in our datasets, whose sequence evolution violated the assumption of global stationary, reversibility, and homogeneity (SRH conditions), on the tree topology (Jermiin et al., 2004; Ababneh et al., 2006), we conducted pairwise sequence comparisons using Bowker's matched-pairs tests of symmetry (Bowker, 1948) and generated heat maps based on the inferred *p*-values as implemented in SymTest version 2.0.47 (https://github.com/ottmi/symtest). We applied Bowker's test exclusively to the datasets A1/a and N1 (for the same reasons as given above in context of the Bayesian tree inference) and compared the results obtained from analyzing the two datasets with each other.

Branch support values were mapped onto the best corresponding phylogenetic tree. All phylogenetic trees were rooted with *N. vitripennis* as outgroup using FigTree version 1.4.3 (Rambaut, 2016). All bootstrap values in the rooted trees were visually checked (see above and Czech et al., 2017) before further editing the resulting vector graphics with Inkscape version 0.91 for publication.

3. Results

3.1. Transcriptome sequencing, assembly, contamination screening, and identification of single-copy protein-coding genes

We sequenced transcriptomes of 32 aculeate wasp species in context of the international 1KITE project (some of which had previously been released by Peters et al., 2017) comprising 24 representatives of the family Vespidae and eight outgroup species. All sequences have been submitted to the NCBI Transcriptome Shotgun Assembly (TSA) database (accession numbers are listed in Supplementary Table 2). Per species, we analyzed 5.9–19.0 M (median: 9.9 M) raw reads, which assembled after adapter clipping and quality trimming into 19,607–43,567 (median: 27,522) contigs. We removed between 56 and 2532 contigs identified as possible contaminants per assembly. The number of contigs in the cleaned assemblies consequently dropped to 19,309–43,415 (median: 27,150). We identified transcripts of 2766–3099 (median: 2990) of the 3260 protein-coding single-copy genes in the 32 transcriptomes. The number of different protein-coding single-copy genes identified in at least one of the 32 transcriptomes was 3251 which constitutes the number of gene alignments we obtained. These and additional assembly statistics are summarized in Supplementary Table 9.

3.2. Phylogenetic analysis of the transcript sequences

After identification of outlier sequences in the 3251 multiple sequence alignments at the amino acid level and subsequent alignment refinement, we removed 577 sequences referring to 217 single-copy genes. Search for protein domains in the refined amino acid sequence alignments assigned 30% of the alignment sites to Pfam-A domains and 6.1% of the alignment sites to Pfam-B domains. A total of 63.9% of the alignment sites consequently remained unannotated (voids). Based on the domain identification results, we split the 3251 multiple sequence alignments and rearranged their sites into 6066 different data blocks, with each block encoding a given protein domain or protein domain clan (comprising domains with a common evolutionary origin; Finn, 2006) or voids. Overall, 1669 data blocks referred to different Pfam-A domains (or domain clans), 1146 referred to different Pfam-B domains, and 3251 referred to voids of the 3251 analyzed genes. After removing ambiguously aligned sites identified by Aliscore (resulting in 5935 data blocks), removing data blocks that contained no phylogenetic information (resulting in 4939 data blocks), eliminating data blocks that did not encompass sequences of all 37 species (resulting in 2531 data blocks), and concatenating the supermatrix resulted in 1,004,596 amino acid and 3,013,788 nucleotide sites, respectively. Both supermatrices covered the sequences of 2531 data blocks and comprised 850 Pfam-A data blocks (incl. clan data blocks), 197 Pfam-B data blocks, and 1484 unannotated gene data blocks (voids). Finally, PartitionFinder suggested a best partitioning scheme integrating these data blocks into 511 partitions.

Phylogenetic analyses of the transcript sequences (1KITE) in combination with the corresponding sequences from five genome projects (OGS) on the amino acid (dataset A0) and on the nucleotide level (dataset N0) with ExaML resulted in two trees whose ingroup relationships were largely congruent (Supplementary Figs. 2 and 3). The results from analyzing the amino acid sequence data are identical to those illustrated in Fig. 2b. The only notable difference between the two obtained tree topologies was that eumenine clade C was inferred as sister to eumenine clade B when analyzing the nucleotide sequence data (Supplementary Fig. 3). Both analyses consistently inferred Stenogastrinae as sister lineage to all remaining Vespidae and confirmed Masarinae being the closest relatives of a clade ("Eumeninae" + (Polistinae + Vespinae)). Both analyses corroborate that the genus *Discoelius* (Eumeninae: Zethini; representative of the former Zethinae) is more closely related to Polistinae + Vespinae than to the remaining eumenine tribes. Finally, both analyses revealed that the tribe Odynerini (Eumeninae) is paraphyletic, with clade D of the tribe Odynerini being more closely related to Eumenini than to any of the remaining clades (A–C) of the tribe Odynerini. Our analyses also consistently inferred the genus *Alastor* (clade A) as the sister lineage to all remaining Eumeninae excluding Discoelius (Eumeninae: Zethini; representative of the former Zethinae).

When assessing the signal for the relationships of Eumeninae (16 species; excl. *Discoelius*), *Discoelius* (one species; Eumeninae: Zethini; representative of the former Zethinae), Polistinae and Vespinae (three species), and Masarinae and Stenogastrinae plus outgroup taxa (17 species in total) to each other via FcLM, we found the highest support (100% of the quartets in the analysis) for *Discoelius* and Polistinae + Vespinae being closest relatives (Supplementary Fig. 4). Permutation tests did not indicate that the FcLM results obtained when analyzing the original amino acid supermatrix were biased by confounding signal (*e.g.*, violation of SRH conditions, non-random distribution of [missing] data). We therefore consider the strong support for a possible sister group relationship of *Discoelius* to Polistinae + Vespinae in both the ML tree inference and in the FcLM results when analyzing the original supermatrix on the translational level (dataset A0) as reliable.

Table 1
Dataset characteristics and analysis schemes in eleven phylogenetic inferences outlined in Sections 2.8 and 2.9 (see also Supplementary Fig. 1). 1KITE: transcripts of 32 species whose transcript libraries were sequenced in the 1KITE project; OGS: genes from the official gene sets of five reference species with sequenced genome; enrichment: genomic sequences of target genes enriched in 25 species.

Dataset/analysis scheme	Data origin	Number of species	Character type	Size of dataset	Number of partitions	Minimum number of species in each partition	Partition-specific substitution models	Number of bootstrap replicates until convergence
A0	1KITE + OGS	37	Amino acids	1,004,596	511	37	Dataset-specific	50
N0	1KITE + OGS	37	Nucleotides	3,013,788	1533[a]	37	GTR + G	100
A1/a	1KITE + OGS + enrichment	62	Amino acids	1,004,596	511	All of A0	As in A0	150
A1/b	1KITE + OGS + enrichment	62	Amino acids	1,004,596	511	All of A0	Dataset-specific	200
A2/a	1KITE + OGS + enrichment	62	Amino acids	519,093	344	All of A0 + at least 1 and up to 25	As in A0	150
A2/b	1KITE + OGS + enrichment	62	Amino acids	519,093	344	All of A0 + at least 1 and up to 25	Dataset-specific	200
A3/a	1KITE + OGS + enrichment	62	Amino acids	335,029	199	62	As in A0	150
A3/b	1KITE + OGS + enrichment	62	Amino acids	335,029	199	62	Dataset-specific	150
N1	1KITE + OGS + enrichment	62	Nucleotides	3,013,788	1533[a]	All of N0	GTR + G	100
N2	1KITE + OGS + enrichment	62	Nucleotides	1,557,279	1032[a]	All of N0 + at least 1 and up to 25	GTR + G	300
N3	1KITE + OGS + enrichment	62	Nucleotides	1,005,087	597[a]	62	GTR + G	150

[a] Each of the three codon positions of a given partition in the inferred partitioning scheme was treated as separate partition.

3.3. Bait design

To extend the taxonomic sampling of vespid wasps by analyzing also ethanol-preserved samples, we inferred baits from the aligned transcripts of 23 vespid species for which transcript libraries were available. BaitFisher, using the parameters and specifications outlined in Section 2.4, suggested a set of 49,226 baits for enriching 2158 coding exons of a total of 913 genes. The 2158 coding exons were targeted using different tiling strategies: (i) seven baits tiled across 240 bp, with a new bait every 20 bp (663 exons referring to 506 genes); (ii) five baits tiled across 200 bp, with a new bait every 20 bp (390 exons referring to 366 genes); (iii) three baits tiled across 160 bp, with a new bait every 20 bp (468 exons referring to 458 genes); (iv) a single bait (637 exons referring to 320 genes).

3.4. Capture of target coding sequences

We applied the 49,226 designed baits to enrich coding exons of 913 single-copy genes in 25 vespid species. Per sample, we collected 2.6–7.7 M raw reads (median: 4.4 M). These assembled after adapter clipping and quality trimming into 7224–69,492 contigs (median: 24,884). After removing possible contaminated contigs (10–704 per assembly, median 160), the assemblies comprised 7186–69,361 contigs (median: 24,790). All sequences of the cleaned assemblies are available at Mendeley Data: http://dx.doi.org/10.17632/npht7b2426.2. The assembled transcripts of the species contained 895–911 (median: 904) target genes. After further data removal as outlined in Section 2.6, the number of target genes per species decreased to 671–733 (median: 709). Supplementary Table 10 provides an overview of the assembly statistics and target gene recovery rates.

The base-coverage depth of the enriched coding exons ranged between 355x and 1023x (median: 618x) in the 23 single-sequenced libraries and was 1577x and 1204x in the two double-sequenced libraries of *E. incostans* and *O. reniformis*, respectively. The base-coverage depth of the bait-binding sites ranged between 407x and 1245x (median: 730x) in the 23 single-sequenced libraries and was 1926x and 1468x in the two double-sequenced libraries (Supplementary Table 11).

Assuming congeneric species exhibiting a similar genome size and using the genome sizes of *O. spinipes* (197.1 Mbp) and *P. dominula* (246.3 Mbp) as references, we estimated enrichment coefficients (C_t/C_g) of 231, 260, and 298 when considering the average base-coverage depth of the bait-binding sites of *O. angustior* (C_t = 1049; 896.4 Mbp sequenced), *O. reniformis* (C_t = 1468; 1114.8 Mbp sequenced), and *P. biglumis* (C_t = 885; 730.7 Mbp sequenced) (all sequence volumes after adapter clipping and quality trimming).

Comparing the base-coverage depth of bait-binding sites between genes, for which we enriched a single exon each, we found a median increase across the 25 species of 82%, 14%, and 25% when rising the number of tiled baits from one to three, from three to five, and from five to seven, respectively (Supplementary Table 11).

3.5. Phylogenetic analysis of transcript and captured target coding sequences

After removal of data blocks without phylogenetic signal from the combined transcriptomic and gDNA alignments, our dataset covered 4939 data blocks. Using this dataset as basis, we applied various filtering and modeling schemes (Section 2.9) that reduced the number of considered data blocks. After (1) eliminating data blocks that did not encompass sequences of all 37 species, the supermatrix consisted of 1,004,596 amino acid and 3,013,788 nucleotide sites, respectively (A1; N1). Both supermatrices covered the sequences of 2531 data blocks and comprised 850 Pfam-A data blocks (incl. clans), 197 Pfam-B data blocks, and 1484 unannotated gene data blocks (merged void regions); (2) eliminating data blocks that did not encompass sequences of all 37 species and have at least one additional sequence obtained via DNA enrichment present in a given data partition, the supermatrix consisted of 519,093 amino acid and 1,557,279 nucleotide sites, respectively (A2; N2). Both supermatrices covered the sequences of 983 data blocks and comprised 452 Pfam-A data blocks (incl. clans), 79 Pfam-B data blocks, and 452 unannotated gene data blocks (voids); and (3) eliminating data blocks that did not encompass sequences of all analyzed 62 species, the supermatrix consisted of 335,029 amino acid and 1,005,087 nucleotide sites, respectively (A3; N3). Both supermatrices covered the sequences of 376 data blocks and comprised 220 Pfam-A data blocks (incl. clans), 16 Pfam-B data blocks, and 140 unannotated gene data blocks (voids).

Combined analysis of the transcript and enriched genomic nucleotide sequences and corresponding nucleotide sequences from five genome projects (OGS) on the translational (amino acid) and nucleotide level (see Table 1 for additional information on the datasets and analysis schemes) revealed, irrespective of the applied tree inference method, largely congruent topologies (Supplementary Figs. 5–14; see Section 2.9 for details on the various analyses schemes). Bowker's matched-pairs tests of symmetry revealed that the nucleotide dataset N1 strongly violates the SRH conditions (Supplementary Fig. 15). The amino acid dataset A1/a, by contrast, suffers much less from such violations (Supplementary Fig. 16). Most sequence comparisons violating the SRH conditions in the amino acid dataset A1/a include at least one outgroup taxon (*e.g.*, *Acromyrmex echinatior*, *Apis mellifera*, *Camponotus floridanus*, *Chrysis viridula*, *Dolichurus corniculus*, *Harpegnathos saltator*, *Pompilus cincereus*, *Sapyga quinquepunctata*, *Tiphia femorata*) and in particular the two ingroup taxa *Vespa crabro* and *Vespula germanica*.

Differences between the nine inferred topologies concern (i) the phylogenetic relationships of the genera *Discoelius* and *Psiliglossa* (both tribe Zethini and representing the former subfamilies Zethinae and Raphiglossinae) relative to Polistinae + Vespinae, (ii) the phylogenetic relationships of species within clade B and within clade C of the tribe Odynerini, (iii) the phylogenetic relationships of species within the tribe Eumenini, and (iv) the phylogenetic position of Scoliidae relative to Formicidae (both outgroup taxa). In context of the present study, only phylogenetic relationships of the genera *Discoelius* and *Psiliglossa* relative to Polistinae + Vespinae are of special interest (see below).

We inferred similar phylogenetic relationships of the major vespid wasp lineages to those obtained when analyzing the transcript sequence (plus the nucleotide sequences from five genome projects) alone (see Section 3.2; Fig. 2b): Stenogastrinae + (Masarinae + ("Eumeninae" + (Polistinae + Vespinae))). Within "Eumeninae", the tribe Odynerini is paraphyletic and comprises four major clades (A–D) of which clade D is sister to the Eumenini. The genus *Alastor* (clade A) was again inferred as sister lineage to all remaining "Eumeninae" (excl. Zethini). Finally, the obtained topologies strongly corroborate the hypothesis of Zethini being more closely related to Polistinae + Vespinae than to the remaining Eumeninae. However, in none of our analyses did the genera *Discoelius* and *Psiliglossa* cluster in a monophyletic clade Zethini. Instead, five of the inferred topologies suggest *Discoelius* (Zethini and representative of the former Zethinae) being more closely related to Polistinae + Vespinae than to *Psiliglossa* (Zethini and representative of the former Raphiglossinae), although with low bootstrap support (28–35%; Fig. 2; Supplementary Figs. 7–10). The remaining four topologies (including all three analyses on the nucleotide level) suggest *Psiliglossa* being more closely related to Polistinae + Vespinae than to *Discoelius*, but with weak (59%; Supplementary Fig. 6) to moderate (81–86 %; Supplementary Figs. 11, 13, 14) bootstrap support. Phylogenetic analysis in a Bayesian framework (datasets A1/a and N1) suggested *Discoelius* being more closely related to Polistinae + Vespinae with 100% posterior probability when analyzing dataset A1/a (amino acids; Supplementary Fig. 5) and suggest *Psiliglossa* more closely related to Polistinae + Vespinae with 100% posterior probability when analyzing dataset N1 (nucleotides; Supplementary Fig. 12).

We assessed the signal in the datasets A1/a, A1/b, A2/a, A2/b, A3/a, and A3b (Table 1; see also Section 2.9 for further details on the

datasets) for the possible phylogenetic relationships of *Discoelius* spp. (two species; Eumeninae: Zethini; representative of the former Zethinae), *Psiliglossa* (one species; Eumeninae: Zethini; representative of the former Raphiglossinae), Polistinae + Vespinae (four species), and all remaining species (55 species) via FcLM. We found few quartets supporting *Discoelius* and *Psiliglossa* being closely related (11–18% of the quartets in each of the six analyses) or *Discoelius* and Polistinae + Vespinae being closely related (10–22%). The majority of quartets support a closer relationship between *Psiliglossa* and Polistinae + Vespinae (60–72%; see Supplementary Figs. 17 and 18). The results from the FcLM permutation approaches I and II suggest that the support of a closer relationship between *Psiliglossa* and Polistinae + Vespinae when analyzing the original amino acid supermatrix cannot be explained by violation of SRH conditions or non-random distribution of (missing) data (or by a combination of both). Note that in the permutation approach I, which assessed violation of SHR conditions and non-random distribution of (missing) data, a sister group relationship *Discoelius* to Polistinae + Vespinae was supported by 30% of the quartets, indicating that the support for this relationship when analyzing the original supermatrix could be due to confounding signal in dataset A1/b (Supplementary Fig. 17-1b). Unexpected was the support of a close relationship of *Psiliglossa* to Polistinae + Vespinae by 39% of the quartets when applying permutation scheme III (Supplementary Fig. 17-1d). This result could be due to the low number of drawn quartets, which caused a random bias in the completely randomized dataset.

4. Discussion

We aimed to infer the phylogenetic relationships of the major vespid wasp lineages (*i.e.*, Eumeninae, Masarinae, Polistinae, Stenogastrinae, Vespinae; excl. Euparagiinae and Gayellini, which were not available to us). Specifically, we were interested in reassessing the hypothesis of eusociality having evolved twice in the family Vespidae and evaluating the monophyly of the subfamily Eumeninae as well as of its tribes (*i.e.*, Eumenini, Odynerini, Zethini). Our results are in line with previous molecular phylogenetic investigations which indicated that Stenogastrinae are likely to be the sister group of all remaining Vespidae (Schmitz and Moritz, 1998; Hines et al., 2007; Peters et al., 2017; see Figs. 1 and 2B), while earlier analyses of morphological and behavioral characters suggested a sister group relationship of Stenogastrinae to Polistinae + Vespidae (Carpenter, 1982, 2003; Pickett and Carpenter, 2010; Hermes et al., 2013). Each molecular phylogenetic study that included Stenogastrinae utilized largely different sets of molecular markers (the studies by Schmitz and Moritz, 1998 and Hines et al., 2007 had one gene in common), which contrast in substitution patterns and evolutionary constrains from each other. Yet, these studies obtained the same result in respect of the phylogenetic position of Stenogastrinae. While our current study builds on the same set of molecular markers as the one by Peters et al. (2017) (*i.e.*, single-copy protein-coding genes), our taxon sample is significantly denser (44 species vs. four species) in the lineage to which Stenogastrinae were previously thought to belong (*i.e.*, "Eumeninae" *sensu lato*, Polistinae, Vespinae; Carpenter, 2003; Pickett and Carpenter, 2010). Nevertheless, our analysis still lacks representatives of the subfamily Euparagiinae and of the tribe Gayellini of the subfamily Masarinae, which would be needed for an even more rigorous test of the relationships between the major vespid lineages. Given that Euparagiinae and Gayellini comprise exclusively solitary nesting species and assuming that the vespid wasp relationships inferred in our study have not been misled by long-branch attraction (Felsenstein, 1978), the specific phylogenetic positions of these two lineages have no impact on our conclusions on how often eusociality evolved within vespid wasps (see below).

The recent confirmation that Rhopalosomatidae likely represent the extant sister lineage of Vespidae (Branstetter et al., 2017; see also Pilgrim et al., 2008) opens up the possibility to even more accurately infer ancestral character states of the family Vespidae by including representatives of Rhopalosomatidae in phylogenetic studies. Having said that, we have currently no reason to assume that the rooting of Vespidae has been compromised by the omission of this outgroup taxon in our study: the number of substitutions that have to be hypothesized along the lineage leading to Vespidae has not been particularly high at the amino acid level. Furthermore, we obtained virtually the same tree topology when analyzing the nucleotide and amino acid datasets irrespective of the tree inference method. Finally, and despite of the fact that deviation from the assumptions of SRH conditions differs significantly between our most comprehensive dataset on the amino acid and on the nucleotide level, we inferred the same topology. This makes us presume that non-stationary processes across the analyzed taxa, which have been reported to have impacted phylogenetic inferences in other lineages of Hymenoptera (Romiguier et al., 2016; Bossert et al., 2017), likely had no major impact on our results. This assumption receives further support from the results of FcLM permutation tests which did not indicate that support for specific phylogenetic hypotheses was driven by compositional heterogeneity and/or non-random distribution of data.

We found *Discoelius* (representative of the former Zethinae) and *Psiliglossa* (representative of the former Raphiglossinae), currently united in the tribe Zethini within the subfamily Eumeninae (Hermes et al., 2013), to be more closely related to Polistinae + Vespinae than to the remaining Eumeninae. Hines et al. (2007) already suggested granting Zethini subfamily status, but our investigation indicates that *Discoelius* and *Psiliglossa* do not necessarily constitute a natural group, since we obtained such a relationship in none of our phylogenetic inferences. However, despite analyzing a significant amount of data, our results are unfortunately not fully conclusive in respect of whether *Discoelius* or *Psiliglossa* is closer related to Polistinae + Vespinae. In our ML tree inferences that suggested a sister group relationship of *Discoelius* to Polistinae + Vespinae, the bootstrap support for this relationship was negligible (28–35%). In those phylogenetic analyses obtained with ExaML that suggested a sister group relationship of *Psiliglossa* to Polistinae + Vespinae, the bootstrap support was 59–86%. The Bayesian phylogenetic inferences provided strong support (100% posterior probability) but contradictory results on whether *Discoelius* or *Psiliglossa* is more closely related to Polistinae + Vespinae. Future studies should improve the taxonomic sampling in this part of the phylogenetic tree (*e.g.*, via target DNA enrichment and exploitation of museum specimens; Mayer et al., 2016) in order to address the phylogenetic relationships between the representatives of the former Raphiglossinae, the representatives of the former Zethinae, and Polistinae + Vespinae. Given the distinct morphology of the former two lineages and their unclear phylogenetic relationship to each other, we propose granting both of them again subfamily status: Raphiglossinae and Zethinae.

The inferred close phylogenetic relationship between Raphiglossinae, Zethinae, and Polistinae + Vespinae substantiates the idea of two independent origins of eusociality within the family Vespidae (Hines et al., 2007): one in the Stenogastrinae and a second in the most recent common ancestor of Polistinae + Vespinae. As outlined by Hines et al. (2007), there are also morphological and behavioral differences between Stenogastrinae and Polistinae + Vespinae that would be consistent with two independent origins of eusociality (*e.g.*, differences in wing morphology, in the provisioning of the larvae, and in the eusocial behavior itself; Hunt, 1991, 2007; Strassmann et al., 1994; Turillazzi, 1991; Yoshikawa et al., 1969). The close phylogenetic relationship between Raphiglossinae, Zethinae, and Polistinae + Vespinae has also implications for the interpretation of the evolution of other traits, such as nest-building: Polistinae and Vespinae are well known for building nests from paper-like material (Evans and West-Eberhard, 1970). Intriguingly, Raphiglossinae and Zethinae apparently also exploit masticated and salivated plant material for constructing their nests (Ferton, 1920; Bischoff, 1927; Blüthgen, 1961 Bohart and

Stange, 1965; Krombein, 1991). Assuming that the use of moistened soil as nest building substrate represents the ancestral character state in Stenogastrinae, Euparagiinae, Masarinae, and Eumeninae *sensu stricto* (Hansell, 1985; Mauss, 2007), the use of plant material for nest-building could represent a synapomorphy of Zethinae, Raphiglossinae, Polistinae, and Vespinae, a hypothesis already discussed by Evans and West-Eberhard (1970). Utilizing plant material enables the eusocial Polistinae and Vespinae to overcome nest size constrains enforced by the limited availability of naturally occurring structures with individual chambers suitable for raising colonies. In this respect, the evolutionary success of Polistinae and Vespinae likely only became possible after their solitary ancestors evolved the ability to exploit masticated and salivated plant material for constructing nests. A similar reasoning has been put forth by Litman et al. (2011) for explaining the evolutionary success of bee lineages that include foreign material in their nest construction. Knowledge of the sister lineage of Polistinae + Vespinae furthermore provides the basis for testing hypotheses on the evolution of eusociality *per se* in vespid wasps (*e.g.*, Hunt and Amdam, 2005).

The phylogenetic relationships within the subfamily Eumeninae (excl. Raphiglossinae and Zethinae) do not support the idea of a monophyletic tribe Odynerini. Given that a tribe Zethini within the subfamily Eumeninae can no longer be justified (see above), the only monophyletic tribe within the subfamily Eumeninae is the Eumenini. We therefore suggest relinquishing a tribal subdivision of the subfamily Eumeninae until the phylogenetic relationships of all major lineages of Eumeninae have been satisfactorily inferred. The present study provides a strong basis for such efforts by delivering both a robust basic phylogenetic framework that can help guide future taxon sampling and designed target DNA enrichment baits.

One aim of our study was to develop and test a dedicated set of target DNA enrichment baits for studying single-copy protein-coding genes in vespid wasps. Target DNA enrichment requires prior knowledge of the target nucleotide sequence in order to design baits for enriching target sites. Given that ingroup nucleotide sequence information for the design of enrichment baits is still often limited, one popular strategy has been to enrich ultra-conserved elements (UCEs) whose nucleotide sequences do not differ even among distantly related reference species for which (typically) sequenced genomes are available (Faircloth et al., 2014; Faircloth, 2017). A second strategy has been termed anchored hybrid-enrichment. It also targets conserved regions of the genome for enrichment, but it copes with known target locus nucleotide sequence variation by using a more diverse set of (reference species-specific) baits per locus (Lemmon et al., 2012). What both strategies have in common is that they primarily exploit the phylogenetic signal of the flanking regions of target loci. The main drawbacks of the two strategies are consequently (a) that it is unavoidable that phylogenetically uninformative sequence sections are enriched and sequenced (due to the fact that these sections serve as anchors for enrichment), (b) that it remains *a priori* uncertain whether or not the obtained flanking sequence sections are orthologous and phylogenetically informative among the analyzed species, and (c) that there is a low probability (primarily when enriching UCEs) that the flanking sequence sections can be analyzed on both the nucleotide and the amino acid level. Being able to study DNA sequences on the amino acid level typically allows to more reliably align the corresponding nucleotide sequences, to phylogenetically analyze more strongly diverged lineages, and to potentially circumvent problems associated with compositional heterogeneity on the nucleotide level (*e.g.*, Misof et al., 2014; present study). For these reasons, we followed a different approach proposed by Mayer et al. (2016). Thus, we first sequenced transcriptomes of representative ingroup species to obtain reliable nucleotide sequence information on potential protein-coding target loci as well as their variation among species. We then designed a set of baits to capture these protein-coding loci in additional species by exploiting all available nucleotide sequence information and optimizing bait design using the software BaitFisher (Mayer et al., 2016). Since the enriched and sequenced protein-coding loci represent a subset of the loci in the sequenced transcriptomes, the enriched protein-coding nucleotide sequences can seamlessly be aligned to the transcriptome sequence data. We consider this a major advantage of the applied approach. The main disadvantage of our approach is the necessity to first have to invest in obtaining ingroup sequence information.

Our sets of baits proved to be highly efficient (~231x to 298x), with a DNA sequence recovery of 98–99.8% of the 913 target genes being captured. Note that we conservatively discarded parts of the enriched sequences in downstream analyses due to the fact that the applied software for identifying and concatenating coding target DNA sequences (Orthograph; Petersen et al., 2017) is optimized for analyzing transcript sequences (cDNA) rather than genomic DNA (gDNA) (outlined in Section 2.6). Since we relied on gene models of the honeybee (Elsik et al., 2014) to identify and remove any possibly erroneously annotated coding sequence section that is not necessarily identical to those of the investigated vespid wasps, we focused the phylogenetic analyses on those exons that largely corresponded in length between the honeybee and vespid wasp. The recently published gene models of the European paper wasp, *Polistes dominula* (Standage et al., 2016), had unfortunately not been available for our study, but will allow future studies to use gene models for an ingroup lineage and will likely reduce the amount of discarded data.

The comprehensive set of baits for enriching single-copy protein-coding genes in vespid wasps will facilitate extending the taxonomic sampling considerably, because it allows for exploiting genomic information from ethanol-preserved samples and possibly also from older museum specimens (Mayer et al., 2016). Enrichment of hundreds of exons in closely related species could also enable coping with phylogenetic uncertainties that result from incomplete lineage sorting by applying shortcut coalescence approaches (Liu et al., 2009a,b; Liu et al., 2010; but see also Springer and Gatesy, 2016). At the same time, genome and transcriptome sequencing data will continue to accumulate (*e.g.*, Lopez-Osorio et al., 2017) and rapidly increase our knowledge of the evolutionary history of the family Vespidae.

Acknowledgments

This study has been enabled by the 1KITE consortium (www.1kite.org). Parts of it were supported by the Germany Research Foundation (DFG; NI 1387/1-1; NI 1387/2-1; NI 1387/4-1; SCHM 2645/1-1; SCHM 2645/2-1). Funding for transcriptome sequencing and assembly was supported by BGI-Shenzhen. X.Z. is also supported by the Chinese Universities Scientific Fund (2017QC114) through China Agricultural University. We thank A. Berg, S. Hopfenmüller, and M. Staab for providing samples, Niklas Noll for bioinformatic input in the development of scripts used to conduct the permutation test, and V. Mauss and two anonymous reviewers for valuable comments on the manuscript. O.N. acknowledges Hessen Forst, the Israeli Nature and National Parks Protection Authority, the Mercantour National Park Service, and the Struktur- und Genehmigungsbehörde Süd for granting permissions to collect samples. All analyzed species were collected before October 2014. O.N. is indebted to J. Gusenleitner for help identifying vespid wasps. O.N. thanks N. Dorchin and M. Niehuis for supporting his field trips to Israel. B.M. and O.N. acknowledge C. Etzbauer and S. Kukowka for technical assistance. K.M., O.N., and S.B. acknowledge V. Achter and the Cologne High Efficient Operating Platform for Science (CHEOPS) at the Regionales Rechenzentrum Köln (RRZ) for bioinformatic and computational support. We thank Ondrej Hlinka, CSIRO, Australia for help when using the CSIRO HPC Cluster Pearcey. We furthermore acknowledge the Gauss Centre for Supercomputing e. V. for funding computing time on the GCS Supercomputer SuperMUC at the Leibniz Supercomputing Centre (LRZ). A.D., B.M., C.M., M.P., O.N., and R.S.P. acknowledge the Leibniz association for installing the graduate school Genomic Biodiversity Research, in the context of which the present study arose.

Author contributions

B.M., O.N., R.S.P. conceived the study. L.K., M.W., O.N., P.R., R.M., T.S. collected samples. A.D., K.M., L.P., O.N., R.S.P., S.L., X.Z. sequenced, assembled, and processed the transcriptomes. C.M., M.S., O.N. conducted the target DNA enrichment and sequencing experiments. A.K., B.M., C.M., K.M., M.P., M.S., O.N., R.S.P., S.B. phylogenetically analyzed the sequence data. All authors contributed to the writing of the manuscript, with M.S., O.N., R.S.P., S.B. taking the lead.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2017.08.020.

References

Ababneh, F., Jermiin, L.S., Ma, C., Robinson, J., 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22, 1225–1231.

Aberer, A.J., Kobert, K., Stamatakis, A., 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31, 2553–2556.

Aberer, A.J., Krompass, D., Stamatakis, A., 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62, 162–166.

Archer, M.E., 2012. Vespine Wasps of the World. Siri Scientific Press, Manchester, UK.

Arévalo, E., Zhu, Y., Carpenter, J.M., Strassmann, J.E., 2004. The phylogeny of the social wasp subfamily Polistinae: evidence from microsatellite flanking sequences, mitochondrial COI sequence, and morphological characters. BMC Evol. Biol. 4, 8.

Bischoff, H., 1927. Biologie der Hymenopteren. Springer, Berlin.

Blüthgen, P., 1961. Die Faltenwespen Mitteleuropas (Hymenoptera, Diploptera). Abh. Dt. Akad. Wiss. Berlin 2, 1–249.

Bohart, R.M., Stange, L.A., 1965. A revision of the genus *Zethus* in the Western Hemisphere (Hymenoptera, Eumenidae). Univ. Calif. Publ. Entomol. 40, 1–208.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., Zhang, P., Huang, Z., Berger, S.L., Reinberg, D., Wang, J., Liebig, J., 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329, 1068–1071.

Bossert, S., Murraya, E.A., Blaimer, B.B., Danforth, B.N., 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol. Phylogenet. Evol. 111, 149–157.

Bowker, A.H., 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43, 572–574.

Branstetter, M.G., Danforth, B.N., Pitts, J.P., Faircloth, B.C., Ward, P.S., Buffington, M.L., Gates, M.W., Kula, R.R., Brady, S.G., 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27, 1019–1025.

Brothers, D.J., Carpenter, J.M., 1993. Phylogeny of Aculeata: Chrysidoidea and Vespoidea. J. Hym. Res. 2, 227–302.

Budriene, A., 2003. Prey of *Symmorphus* wasps (Hymenoptera: Eumeninae) in Lithuania. Acta Zool. Lituanica 13, 306–310.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinform. 10, 421.

Carpenter, J.M., 1982. The phylogenetic relationships and natural classification of the Vespoidea (Hymenoptera). Syst. Entomol. 7, 11–38.

Carpenter, J.M., 1987. Phylogenetic relationships and classification of the Vespinae (Hymenoptera: Vespidae). Syst. Entomol. 12, 413–431.

Carpenter, J.M., 1988a. The phylogenetic systems of the Gayellini (Hymenoptera: Vespidae, Masarinae). Psyche 95, 211–241.

Carpenter, J.M., 1988b. The phylogenetic system of the Stenogastrinae (Hymenoptera, Vespidae). J. New York Ent. Soc. 96, 140–175.

Carpenter, J.M., 1991. Phylogenetic relationships and the origin of social behaviour in the Vespidae. In: Ross, K.G., Matthews, R.W. (Eds.), The Social Biology of Wasps. Cornell University Press, Ithaca, New York, USA, pp. 7–32.

Carpenter, J.M., 1993. Biogeographic Patterns in the Vespidae (Hymenoptera): Two Views of Africa and South America. In: Goldblatt, P. (Ed.), Biological Relationships Between Africa and South America Proceedings of the 37th Annual Systematics Symposium, Held at Missouri Botanical Gardens, 4–6 October 1990. Yale Univ. Press, New Haven, London, pp. 139–155.

Carpenter, J.M., 1996. Generic classification of the Australian pollen wasps (Hymenoptera: Vespidae; Masarinae). J. Kans. Entomol. Soc. 69, 384–400.

Carpenter, J.M., 2003. On "Molecular Phylogeny of Vespidae (Hymenoptera) and the Evolution of Sociality in Wasps". Am. Mus. Novit. 3389, 1–20.

Carpenter, J.M., Cumming, J.M., 1985. A character analysis of the North American potter wasps (Hymenoptera: Vespidae; Eumeninae). J. Nat. Hist. 19, 877–916.

Carpenter, J.M., Perera, E.P., 2006. Phylogenetic relationships among yellowjackets and the evolution of social parasitism (Hymenoptera: Vespidae, Vespinae). Am. Mus. Novit. 3507, 1–19.

Carpenter, J.M., Rasnitsyn, A.P., 1990. Mesozoic Vespidae. Psyche 97, 1–20.

Crespi, B.J., Yanega, D., 1995. The definition of eusociality. Behav. Ecol. 6, 109–115.

Czech, L., Huerta-Cepas, J., Stamatakis, A., 2017. A critical review on the use of support values in tree viewers and bioinformatics toolkits. Mol. Biol. Evol. 34, 1535–1542.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., Minh, B.Q., von Haeseler, A., Ebersberger, I., Pass, G., Misof, B., 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. 31, 239–249.

Eddy, S.R., 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195.

Elsik, C.G., Worley, K.C., Bennett, A.K., Beye, M., Camara, F., Childers, C.P., de Graaf, D.C., Debyser, G., Deng, J., Devreese, B., Elhaik, E., Evans, J.D., Foster, L.J., Graur, D., Guigo, R., HGSC production teams, Hoff, K.J., Holder, M.E., Hudson, M.E., Hunt, G.J., Jiang, H., Joshi, V., Khetani, R.S., Kosarev, P., Kovar, C.L., Ma, J., Maleszka, R., Moritz, R.F., Muñoz-Torres, M.C., Murphy, T.D., Muzny, D.M., Newsham, I.F., Reese, J.T., Robertson, H.M., Robinson, G.E., Rueppell, O., Solovyev, V., Stanke, M., Stolle, E., Tsuruda, J.M., Vaerenbergh, M.V., Waterhouse, R.M., Weaver, D.B., Whitfield, C.W., Wu, Y., Zdobnov, E.M., Zhang, L., Zhu, D., Gibbs, R.A., 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genom. 15, 86.

Evans, H.E., West-Eberhard, M.J., 1970. The Wasps. University of Michigan Press, Ann Arbor, Michigan.

Faircloth, B.C., 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods Ecol. Evol. (early access). http://dx.doi.org/10.1111/2041-210X.12754.

Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G., 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Mol. Ecol. Res. 15, 489–501.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27, 401–410.

Ferton, C., 1920. Notes détachées sur l'instinct des Hyménoptères mellifères et ravisseurs avec la description deux espèces nouvelles. (9e Série). Ann. Soc. Ent. Fr. 89, 329–375.

Finn, R.D., 2006. Pfam: clans, web tools and services. Nucl. Acids Res. 34, D247–D251.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucl. Acids Res. 42, D222–D230.

Gess, F.W., 1998. *Priscomasaris namibiensis* Gess, a new genus and species of Masarinae (Hymenoptera: Vespidae) from Namibia, southern Africa, with a discussion of its position within the subfamily. J. Hym. Res. 7, 296–304.

Gess, S.K., 1996. The Pollen Wasps – Ecology and Natural History of the Masarinae. Harvard University Press, Cambridge, Massachusetts, pp. 1–340.

Hansell, M.H., 1985. The nest material of Stenogastrinae (Hymenoptera, Vespidae) and its effect on the evolution of social behaviour and nest design. Actes Coll. Insectes Soc. 2, 57–63.

Hermes, M.G., Melo, G.A.R., Carpenter, J.M., 2013. The higher-level phylogenetic relationships of the Eumeninae (Insecta, Hymenoptera, Vespidae), with emphasis on *Eumenes* sensu lato. Cladistics 30, 1–32.

Hines, H.M., Hunt, J.H., O'Connor, T.K., Gillespie, J.J., Cameron, S.A., 2007. Multigene phylogeny reveals eusociality evolved twice in vespid wasps. Proc. Natl. Acad. Sci. USA 104, 3295–3299.

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L., Teupser, D., Hackermueller, J., Stadler, P.F., 2014. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection. Genome Biol. 15, R34.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C.M., Khaitovich, P., Vogel, J., Stadler, P.F., Hackermueller, J., 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput. Biol. 5, e1000502.

Honeybee Genome Sequencing Consortium, 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443, 931–949.

Hunt, J.H., 1991. Nourishment and the evolution of the social vespidae. In: Ross, K.G., Matthews, R.W. (Eds.), The Social Biology of Wasps. Cornell University Press, Ithaca, pp. 426–450.

Hunt, J.H., 2007. The Evolution of Social Wasps. Oxford University Press, New York, USA.

Hunt, J.H., Amdam, G.V., 2005. Bivoltinism as an antecedent to eusociality in the paper wasp genus *Polistes*. Science 308, 264–267.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Iwata, K., 1976. Evolution of Instinct – Comparative Ethology of Hymenoptera. Amerind Publishing Co., New Dehli.

Jermiin, L., Ho, S.Y.W., Ababneh, F., Robinson, J., Larkum, A.W.D., 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53, 638–643.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Katoh, K., Standley, D.M., 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics 32, 1933–1942.

Kimsey, L.S., Bohart, R.M., 1991. [1990]): The Chrysidid Wasps of the World. Oxford University Press, Oxford, New York, Toronto.

Kozlov, A., Aberer, A.J., Stamatakis, A., 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputer. Bioinformatics 31, 2577–2579.

Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B.M., Wägele, J.W., Misof, B., 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front. Zool. 7, 10.

Krenn, H.W., Mauss, V., Plant, J., 2002. Evolution of the suctorial proboscis in pollen wasps (Masarinae, Vespidae). Arthropod Struct. Develop. 31, 103–120.

Krombein, K.V., 1979. Vespoidea. In: Krombein, K.V., Hurd, P.D., Smith, D.R., Burks, B.D. (Eds.), Catalog of Hymenoptera in America North of Mexico 2. Smithsonian Institution Press, Washington, pp. 1469–1522.

Krombein, K.V., 1991. Biosystematic studies of Ceylonese wasps XIX: Natural history notes in several families (Hymenoptera: Eumenidae, Vespidae, Pompilidae and Crabronidae). Smithson. Contrib. Zool. 515, 1–41.
Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A., 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14, 82.
Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34, 772–773.
Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320.
Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.
Litman, J.R., Danforth, B.N., Eardley, C.D., Praz, C.J., 2011. Why do leafcutter bees cut leaves? New insights into the early evolution of bees. Proc. R. Soc. B 278, 3593–3600.
Liu, L., Yu, L., Kubatko, L.S., Pearl, D.K., Edwards, S.V., 2009a. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53, 320–328.
Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.
Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.
Lopez-Osorio, F., Pickett, K.M., Carpenter, J.M., Ballif, B.A., Agnarsson, I., 2017. Phylogenomic analysis of yellowjackets and hornets (Hymenoptera: Vespidae, Vespinae). Mol. Phylogenet. Evol. 107, 10–15.
Mauss, V., 2007. Evolution verschiedener Lebensformtypen innerhalb basaler Teilgruppen der Faltenwespen (Hymenoptera, Vespidae). Denisia 20, 701–722.
Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R.S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.W., Misof, B., Bleidorn, C., Ohl, M., Niehuis, O., 2016. BaitFisher: a software package for multispecies target DNA enrichment probe design. Mol. Biol. Evol. 33, 1875–1886.
Meusemann, K., von Reumont, B.M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J.W., Misof, B., 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27, 2451–2464.
Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B.M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M., Zhou, X., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767.
Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K., Meusemann, K., 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinform. 14, 348.
Misof, B., Misof, K., 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol. 58, 21–34.
Nygaard, S., Zhang, G., Schiøtt, M., Li, C., Wurm, Y., Hu, H., Zhou, J., Ji, L., Qiu, F., Rasmussen, M., Pan, H., Hauser, F., Krogh, A., Grimmelikhuijzen, C.J., Wang, J., Boomsma, J.J., 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. Genome Res. 21, 1339–1348.
Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., Stamatakis, A., 2010. How many bootstrap replicates are necessary? J. Comput. Biol. 17, 337–354.
Peng, Y., Leung, H.C., Yiu, S.M., Chin, F.Y., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428.
Perrard, A., Grimaldi, D., Carpenter, J.M., 2017. Early lineages of Vespidae (Hymenoptera) in Cretaceous amber. Syst. Entomol. 42, 379–386.
Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P.A., Heraty, J., Kjer, K.M., Klopfstein, S., Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B., Niehuis, O., 2017. Evolutionary history of the Hymenoptera. Curr. Biol. 27, 1–6.
Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., Niehuis, O., 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinform. 18, 111.
Pickett, K.M., Carpenter, J.M., 2010. Simultaneous analysis and the origin of eusociality in the Vespidae (Insecta: Hymenoptera). Arthropod Syst. Phylo. 68, 3–33.
Pilgrim, E.M., von Dohlen, C.D., Pitts, J.P., 2008. Molecular phylogenetics of Vespoidea indicate paraphyly of the superfamily and novel relationships of its component families and subfamilies. Zool. Scr. 37, 539–560.
Rambaut, A., 2016. FigTree version 1.4.3 for Mac OS X. http://tree.bio.ed.ac.uk/software/figtree/.
Richards, O.W., 1962. A Revisional Study of the Masarid Wasps. British Museum (Natural History), London, UK.
Romiguier, J., Cameron, S.A., Woodard, S.H., Fischman, B.J., Keller, L., Praz, C.J., 2016. Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. Mol. Biol. Evol. 33, 670–678.
Schmitz, J., Moritz, R.F.A., 1998. Molecular phylogeny of Vespidae (Hymenoptera) and the evolution of sociality in wasps. Mol. Phylogenet. Evol. 9, 183–191.
Springer, M.S., Gatesy, J., 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94, 1–33.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Standage, D.S., Berens, A.J., Glastad, K.M., Severin, A.J., Brendel, V.P., Toth, A.L., 2016. Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. Mol. Ecol. 25, 1769–1784.
Strassmann, J.E., Hughes, C.R., Turillazzi, S., Solís, C.R., Queller, D.C., 1994. Genetic relatedness and incipient eusociality in stenogastrine wasps. Anim. Behav. 48, 813–821.
Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 94, 6815–6819.
Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucl. Acids Res. 34, W609–W612.
Tribolium Genome Sequencing Consortium, 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452, 949–955.
Turillazzi, S., 1991. The Stenogastrinae. In: Ross, K.G., Matthews, R.W. (Eds.), The Social Biology of Wasps. Cornell University Press, Ithaca, pp. 74–98.
Vernier, R., 1997. Essai d'analyse cladistique des genres d'Eumeninae (Vespidae, Hymenoptera) représentés en Europe septentrionale, occidentale et centrale. B. Soc. Neuchâteloise Sci. Nat. 120, 87–98.
Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M., Kriventseva, E.V., 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucl. Acids Res. 41, D358–D365.
Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Beukeboom, L.W., Desplan, C., Elsik, C.G., Grimmelikhuijzen, C.J., Kitts, P., Lynch, J.A., Murphy, T., Oliveira, D.C., Smith, C.D., van de Zande, L., Worley, K.C., Zdobnov, E.M., Aerts, M., Albert, S., Anaya, V.H., Anzola, J.M., Barchuk, A.R., Behura, S.K., Bera, A.N., Berenbaum, M.R., Bertossa, R.C., Bitondi, M.M., Bordenstein, S.R., Bork, P., Bornberg-Bauer, E., Brunain, M., Cazzamali, G., Chaboub, L., Chacko, J., Chavez, D., Childers, C.P., Choi, J.H., Clark, M.E., Claudianos, C., Clinton, R.A., Cree, A.G., Cristino, A.S., Dang, P.M., Darby, A.C., de Graaf, D.C., Devreese, B., Dinh, H.H., Edwards, R., Elango, N., Elhaik, E., Ermolaeva, O., Evans, J.D., Foret, S., Fowler, G.R., Gerlach, D., Gibson, J.D., Gilbert, D.G., Graur, D., Gründer, S., Hagen, D.E., Han, Y., Hauser, F., Hultmark, D., Hunter 4th, H.C., Hurst, G.D., Jhangian, S.N., Jiang, H., Johnson, R.M., Jones, A.K., Junier, T., Kadowaki, T., Kamping, A., Kapustin, Y., Kechavarzi, B., Kim, J., Kim, J., Kiryutin, B., Koevoets, T., Kovar, C.L., Kriventseva, E.V., Kucharski, R., Lee, H., Lee, S.L., Lees, K., Lewis, L.R., Loehlin, D.W., Logsdon Jr, J.M., Lopez, J.A., Lozado, R.J., Maglott, D., Maleszka, R., Mayampurath, A., Mazur, D.J., McClure, M.A., Moore, A.D., Morgan, M.B., Muller, J., Munoz-Torres, M.C., Muzny, D.M., Nazareth, L.V., Neupert, S., Nguyen, N.B., Nunes, F.M., Oakeshott, J.G., Okwuonu, G.O., Pannebakker, B.A., Pejaver, V.R., Peng, Z., Pratt, S.C., Predel, R., Pu, L.L., Ranson, H., Raychoudhury, R., Rechtsteiner, A., Reese, J.T., Reid, J.G., Riddle, M., Robertson, H.M., Romero-Severson, J., Rosenberg, M., Sackton, T.B., Sattelle, D.B., Schlüns, H., Schmitt, T., Schneider, M., Schüler, A., Schurko, A.M., Shuker, D.M., Simões, Z.L., Sinha, S., Smith, Z., Solovyev, V., Souvorov, A., Springauf, A., Stafflinger, E., Stage, D.E., Stanke, M., Tanaka, Y., Telschow, A., Trent, C., Vattathil, S., Verhulst, E.C., Viljakainen, L., Wanner, K.W., Waterhouse, R.M., Whitfield, J.B., Wilkes, T.E., Williamson, M., Willis, J.H., Wolschin, F., Wyder, S., Yamada, T., Yi, S.V., Zecher, C.N., Zhang, L., Gibbs, R.A., 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327, 343–348.
Wurdack, M., Herbertz, S., Dowling, D., Kroiss, J., Strohm, E., Baur, H., Niehuis, O., Schmitt, T., 2015. Striking cuticular hydrocarbon dimorphism in the mason wasp *Odynerus spinipes* and its possible evolutionary cause (Hymenoptera: Chrysididae, Vespidae). Proc. Royal Soc. B 282, 20151777.
Yoshikawa, K., Ohgushi, R., Sakagami, S.F., 1969. Preliminary report on entomology of the Osaka City University 5th Scientific Expedition to Southeast Asia 1966 – with descriptions of two new genera of stenogastrine wasps by J. van der Vecht. Nat. Life Southeast Asia 6, 153–200.

Pauli et al. BMC Genomics (2016) 17:861
DOI 10.1186/s12864-016-3205-1

BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects

Insect insulator proteins

Thomas Pauli[1*], Lucia Vedder[2], Daniel Dowling[3], Malte Petersen[1], Karen Meusemann[1,4,5], Alexander Donath[1], Ralph S. Peters[6], Lars Podsiadlowski[7], Christoph Mayer[1], Shanlin Liu[8,9], Xin Zhou[10,11], Peter Heger[12], Thomas Wiehe[12], Lars Hering[13], Georg Mayer[13], Bernhard Misof[1] and Oliver Niehuis[1*]

Abstract

Background: Body plan development in multi-cellular organisms is largely determined by homeotic genes. Expression of homeotic genes, in turn, is partially regulated by insulator binding proteins (IBPs). While only a few enhancer blocking IBPs have been identified in vertebrates, the common fruit fly *Drosophila melanogaster* harbors at least twelve different enhancer blocking IBPs. We screened recently compiled insect transcriptomes from the 1KITE project and genomic and transcriptomic data from public databases, aiming to trace the origin of IBPs in insects and other arthropods.

Results: Our study shows that the last common ancestor of insects (Hexapoda) already possessed a substantial number of IBPs. Specifically, of the known twelve insect IBPs, at least three (*i.e.*, CP190, Su(Hw), and CTCF) already existed prior to the evolution of insects. Furthermore we found GAF orthologs in early branching insect orders, including Zygentoma (silverfish and firebrats) and Diplura (two-pronged bristletails). Mod(mdg4) is most likely a derived feature of Neoptera, while Pita is likely an evolutionary novelty of holometabolous insects. Zw5 appears to be restricted to schizophoran flies, whereas BEAF-32, ZIPIC and the Elba complex, are probably unique to the genus *Drosophila*. Selection models indicate that insect IBPs evolved under neutral or purifying selection.

Conclusions: Our results suggest that a substantial number of IBPs either pre-date the evolution of insects or evolved early during insect evolution. This suggests an evolutionary history of insulator binding proteins in insects different to that previously thought. Moreover, our study demonstrates the versatility of the 1KITE transcriptomic data for comparative analyses in insects and other arthropods.

Keywords: Insulator binding proteins, Comparative transcriptomic analyses, Gene evolution, Arthropod evolution

Background

Chromatin insulation accounts for the formation of independent transcriptional units on eukaryote chromosomes [1–3]. Chromatin insulation is mediated by insulator binding proteins (IBPs), which insulate transcriptional units either by acting as chromatin barriers (preventing the formation of heterochromatin and thus the silencing of active genes) or as enhancer blockers (preventing enhancers from binding to off-target promoters). Due to their large-scale effects on transcription and on the regulation of fundamental developmental processes, IBPs can significantly impact body plan formation [4–6]. Consequently, IBPs may play an important role in the evolution of body plans and biological diversity. Following this line of reasoning, studying the evolution of IBPs in insects[1] appears rewarding. In the common fruit fly, *Drosophila melanogaster*, twelve different IBPs have been identified (Table 1). However,

* Correspondence: s6thpaul@uni-bonn.de; oliver.niehuis@gmail.com
[1]Center of Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 51113 Bonn, Germany
Full list of author information is available at the end of the article

Table 1 Summary of all currently known insulator binding proteins (IBPs) in *Drosophila melanogaster*, with information on the Pfam symbol of the conserved protein domain families found in the respective proteins with the corresponding references

Insulator binding protein	Conserved domains	Reference
CTCF	zf-C2H2 [11]	[24]
Su(Hw)	zf-C2H2 [12]	[22, 23]
Pita	zf-AD [1], zf-C2H2 [10]	[43]
ZIPIC	zf-C2H2 [7]	[43]
Zw5	zf-C2H2 [8]	[67]
CP190	BTB [1], zf-C2H2 [4]	[32, 68]
GAF	BTB [1], GAGA [1]	[69, 70]
Mod (mdg4)	BTB [1], FLYWCH [1]	[71, 72]
BEAF-32	zf-BED [1], BESS [1]	[34]
Ibf1	zf-BED [1]	[44]
Ibf2	zf-BED [1]	[44]
Elba-complex (Elba 1,2,3)	BEN [1] (Elba 1,2), none (Elba 3)	[30]

The number of repeats of each conserved domain in the respective protein is given in square brackets

the taxonomic distribution of IBPs in insects and the IBPs' possible correlation with biological diversity has only been studied in a small number of species [7, 8]. In the present investigation, we therefore exploit information in recently published transcriptome and genome sequence data to trace the evolution of IBPs in insects and show that the evolution of IBPs in 100 insect species is more complex than previously anticipated.

Transcriptional units comprise groups of genes and associated regulatory elements, such as enhancers, silencers, and promoters, that can be brought into close spatial proximity to each other by folding of chromatin fibers [9]. It has been shown that transcriptionally active units can be immediately adjacent to inactive genomic regions [10]. Such a spatial arrangement can result in inadvertent genic interactions. Experiments show that IBPs are capable of effectively impeding such interactions [11, 12]. In *D. melanogaster*, the protein Cut acts as a chromatin barrier insulator, like the homologous protein CDP of humans that binds to a similar target region [13]. As chromatin barriers, Cut and CDP inhibit interactions between heterochromatin and actively transcribed euchromatin [14]. In general, when heterochromatin comes into spatial proximity of transcribed euchromatin, it can spread along the chromatin fiber into adjacent euchromatin regions and repress transcription. Chromatin barrier IBPs seem to be ancient proteins in eukaryotes since it has also been demonstrated by the interaction between TFIIIC and tRNA genes found in yeast and humans [15–18]. The taxonomically wide distribution of chromatin-barring IBPs (*e.g.,* Cut in *D. melanogaster* and CDP and TFIIIC in humans and yeast) implies that chromatin barring is essential for chromosomal organization in eukaryotes [19].

Enhancer blocking IBPs apparently evolved later than chromatin barrier IBPs and are possibly restricted to bilaterians [20]. Enhancers are regulatory elements that can bind to a promoter and thereby enhance transcription of the associated gene. The switch between a euchromatic and a heterochromatic state of adjacent chromosome regions can result in unfavorable alignments of enhancers in spatial proximity of otherwise distant promoters. Consequently, enhancers could interact with off-target promoters. Such interactions can be prevented by enhancer-blocking IBPs [21]. Su(Hw) (suppressor of hairy wing) was the first enhancer blocker to be functionally characterized in *D. melanogaster*. Su(Hw) was discovered due to its ability to protect DNA of transgenic flies from the phenotypic effect of the transposable element *gypsy*, which induces mutations affecting transcription by inserting itself into splice sites and sequences necessary for initiating transcription [22, 23]. Su(Hw) seems to be restricted to arthropods [7, 8]. Bell and colleagues [24] described a second enhancer blocker, called CTCF (CCCTC binding factor), in birds and mammals. In contrast to Su(Hw), CTCF was shown to be taxonomically widespread and has been found in all bilaterian lineages studied [7, 20].

As of yet CTCF is the only enhancer-blocking IBP known in vertebrates. However, B1 and B2 type *SINEs* (Short Interspersed Nuclear Elements), which are transposable elements, can also encode for enhancer blocking peptides [25, 26]. Additionally, tRNA genes have been shown to exhibit enhancer-blocking or chromatin barring properties [18, 27]. Furthermore, a homolog of the GAGA factor (GAF) has been identified in vertebrates, where it might function as an enhancer blocking IBP [28]. So far, twelve IBPs with enhancer-blocking properties have been identified in *D. melanogaster,* including CTCF and Su(Hw) (Table 1). All IBPs contain DNA-binding domains. The most common are zinc-finger domains, or domains with a zinc-finger core, such as zf-C2H2, zf-BED, GAGA and FLYWCH. The Elba (Early boundary activity) protein complex and a specific isoform of Mod(mdg4) (modifier of mdg4) use BEN domains to bind DNA instead [29, 30]. Three IBPs, CP190 (Centrosomal protein 190 kD), GAF, and Mod(mdg4), additionally have a BTB domain (bric-a-brac, ttk and broad complex), which is assumed to mediate DNA binding and protein binding [31]. Mod(mdg4) and CP190 often interact with CTCF [5] and Su(Hw) [32] and are shown to form complexes in *D. melanogaster*. These interactions might possibly be mediated through the BTB domain. Other domains are a zf-AD (zinc-finger associated domain) found in Pita and a BESS domain (named after the three proteins in which it was found: BEAF-32 (Boundary element associated factor of 32 kD), Suvar(3)7, and Stonewall [33–35]) found in BEAF-32.

In *D. melanogaster*, IBPs exhibiting enhancer-blocking function actively regulate larval development. For example, individual deletion of *CTCF*, *CP190*, *BEAF-32*, and *GAF* alters the expression of hox genes, resulting in lethal homeotic transformations [4–6]. Deletion of *Su(Hw)* induces sterility in female *D. melanogaster* due to changes in the expression of oogenesis-related genes [36]. These experiments demonstrate the importance of IBP-mediated transcriptional regulation for proper larval development and oogenesis in *D. melanogaster* and raise the intriguing question of when and how these important IBPs evolved in arthropods.

Schoborg and Labrador [7] as well as Heger and colleagues [8, 20] screened publicly available transcriptomes as well as draft genomes of insects for genes orthologous to *D. melanogaster* IBPs. They inferred that *CTCF* likely evolved in the stem lineage of Bilateria. *Su(Hw)* possibly evolved in the stem lineage of arthropods and *CP190* possibly evolved in the stem lineage of the Pancrustacea (insects plus crustaceans). The IBP *GAF* likely evolved in the last common ancestor of Holometabola and Hemiptera, and Mod(mdg4) likely emerged in the last common ancestor of Aparaglossata (all holometabolan insects except Hymenoptera, see [37]). Finally, *Zw5* and *BEAF-32* are possibly unique to the dipteran family Drosophilidae. Because *GAF* and *Mod(mdg4)* apparently emerged during the diversification of Holometabola, we suggest that IBPs may have played a key role for the tremendous diversification of holometabolous insects.

We therefore analyzed whole-body transcriptomes sampled across all described insect orders, which were compiled in the international 1KITE project [38]. We additionally considered sequence data of other panarthropod lineages, including RNAseq data of onychophorans and a tardigrade. Additionally, we screened the genome of a nematode (*Trichinella spiralis*). We screened for all twelve enhancer-blocking IBPs that have previously been identified in insects (Hexapoda). We assessed the orthology of all identified candidate transcripts of IBPs by using the best reciprocal hit criterion, inferred the phylogeny of each gene from the assembled transcripts and studied selective forces that might have acted on these genes. Our data and results furthermore set the stage for future comparative and experimental studies on this intriguing group of proteins.

Results

We used profile Hidden Markov Models (pHMMs) in order to search for orthologous sequences of twelve enhancer-blocking IBPs known from *D. melanogaster* in transcriptome data sets from 100 insect species and in transcriptomes and genomes of ten outgroup species, including crustaceans, chelicerates, myriapods, onychophorans (velvet worms), a tardigrade, and a nematode (Fig. 1). We found that three IBPs are particularly widespread across insect orders and outgroups: (i) *CTCF* was found in the transcript libraries of 105 species, including the nematode, *Trichinella spiralis*; (ii) *Su(Hw)* occurs in the transcript libraries of 86 species, including crustaceans, chelicerates, and myriapods (iii) *CP190* was found in the transcript libraries of 81 species, including crustaceans. Ancestral state reconstruction corroborates the idea that *CTCF* was already present in the last common ancestor of Panarthropoda (Onychophora + Tardigrada + Arthropoda; Additional file 1: Figure S1), Su(Hw) was already present in the last common ancestor of Arthropoda (Additional file 1: Figure S2), and *CP190* in the last common ancestor of Pancrustacea (Additional file 1: Figure S3).

In contrast, we detected *GAF* exclusively in insects, including coneheads (Protura), but not in all species studied. In fact, only 38 screened insect transcriptome assemblies included putative transcripts of *GAF*. We did not find any *GAF* transcripts in the screened transcriptomes of butterflies and moths (Lepidoptera), caddisflies (Trichoptera), scorpionflies (Mecoptera), fleas (Siphonaptera), and springtails (Collembola). In addition, we did not find *GAF* in the draft genomes of *Bombyx mori* (Lepidoptera), *Limnephilus lunatus* (Trichoptera), *Machilis hrabei* (Archaeognatha), and *Catajapyx aquilonaris* (Diplura). Ancestral state reconstruction for GAF reveals multiple losses of this protein (Additional file 1: Figure S4). A search for the vertebrate *GAF* homolog in the insect transcriptomes yielded several positive hits, which however did not fulfill the best reciprocal hit criterion.

Transcripts of *Mod(mdg4)* were exclusively detected in species of neopteran insects (*i.e.*, insects with the ability to flex their wings above their abdomen; 57 species of all extant neopteran insect orders, except for ground lice, Zoraptera, and earwigs, Dermaptera). We also searched an early draft genome of a bristletail (*Machilis hrabei*; Archaeognatha), a mayfly (*Ephemera danica*; Ephemeroptera), and a dragonfly (*Ladona fulva*; Odonata) for possible orthologs of *Mod(mdg4)*. We identified a FLYWCH zinc finger domain (domain orthology was confirmed by the best reciprocal hit criterion; see the Methods section) when searching the *M. hrabei* genome. However, since other proteins, such as Su(Kpn) (Suppressor of Killer of prune) [39], are known to also contain FLYWCH domains, we deem these hits as insufficient evidence for the occurrence of *Mod(mdg4)* in bristletails.

We found orthologs of *Pita* only in transcript assemblies of holometabolous insects (30 species, covering 11 orders), and ancestral state reconstruction of *Pita* suggests that this IBP was present in the last common ancestor of Holometabola (Additional file 1: Figure S5).

We identified transcripts encoding the IBP *Zw5* only in two species of Diptera (*i.e.*, *Lipara lucens* and *Triarthria setipennis*).

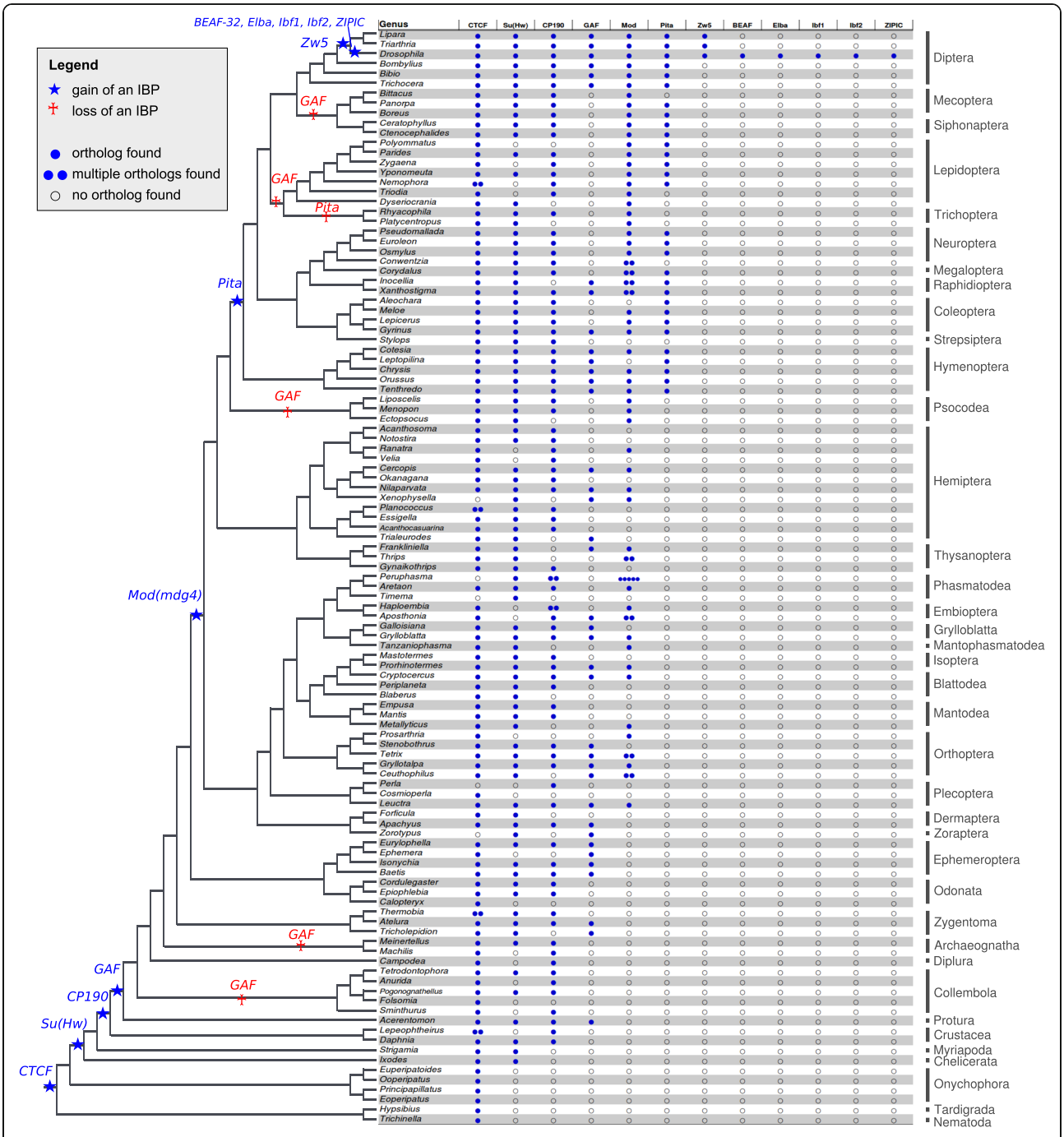

Fig. 1 Evolution of enhancer-blocking insulator proteins (IBP) in arthropods. Gains of IBPs are indicated by blue stars, potential losses by red crosses. Blue circles in the table indicate the presence of transcripts of a given IBP in transcriptomes. Multiple blue circles indicate the number of copies found. Transparent circles indicate the putative absence of transcripts of an IBP. The backbone tree topology is adopted from [38]. The phylogenetic relationships among outgroup taxa (*i.e.*, Onychophora, Tardigrada, and Nematoda) are according to [62]

We could not find evidence for the presence of orthologs of *ZIPIC* (zinc-finger protein interacting with CP190), *BEAF-32*, *Ibf1* (Insulator binding factor 1), *Ibf2*, (Insulator binding factor 1) and the genes encoding the Elba complex in any of the investigated species when searching all available transcriptomes. We did find such evidence, however, in the genome of *D. willistoni* (Drosophilidae). Note that *Ibf1*, *Ibf2*, *ZIPIC*, *BEAF-32*, and the proteins of the Elba complex have only been identified in Drosophila to date.

Finally, we conducted a branch-specific analysis of d_N/d_S-ratios to test for positive selective pressure (Table 2).

Table 2 Results from analyzing d_N/d_S ratios in genes encoding insulator proteins in insects

Gene	Branch	lnL0	lnL1	LRT		*p*-value
CP190	Crustacea	−5495.527	−5495.388	0.278		0.598
CP190	Holometabola	−5495.284	−5495.260	0.047		0.828
CTCF	Onychophora	−1314.281	−1310.626	7.308		0.007
CTCF	Holometabola	−1311.997	−1312.166	0.338		0.561
GAF	Acerentomon	−2006.810	−2006.810	0.0		1.000
GAF	Holometabola	−2006.810	−2006.810	0.0		1.000
Mod (mdg4)	Polyneoptera	−15377.060	−15377.060	4.000	10–6	0.998
Mod (mdg4)	Holometabola	−15374.903	−15373.888	2.032		0.154
Pita	Hymenoptera	−1054.403	−1046.840	15.13		<0.001*
Su (Hw)	Holometabola	−11052.401	−11052.210	0.383		0.536

Shown are the gene name and the branch, along which the respective selection model was tested, the log-likelihood for the neutral model (lnL0) and for positive selection (lnL1), the likelihood ratio test statistic (LRT), and the associated *p*-value. Branches on which the positive selection model fits significantly better than the neutral selection model are indicated by *. Bonferroni corrected significance threshold was $\alpha = 0.005$. The degree of freedom (df) was 1 for all tests

We found no statistically significant evidence for positive selection in *CTCF* in Onychophora ($p = 0.007$; Bonferroni corrected $\alpha = 0.005$). *Pita* showed evidence for positive selection in Hymenoptera ($p < 0.001$; Bonferroni corrected $\alpha = 0.005$).

Completeness of the transcriptomes was assessed by using the BUSCO (Benchmarking Universal Single-Copy Orthologs) pipeline [40]. The transcriptome completeness ranges from 15.2 % (*Bittacus pilicornis,* Mecoptera) to 81.2 % (*Lipara lucens*, Diptera). Results of the analysis are summarised in Table 3, absolute values for all used 1KITE transcriptomes can be found in Additional file 2: Table S1.

None of the phylogenetic analyses of the transcripts of the above genes and proteins provided evidence for gene duplication events (Additional file 1: Figures S8–S14).

Discussion

We traced the evolutionary origin of all twelve enhancer-blocking insulator proteins (IBPs) known from *D. melanogaster*. We searched for transcripts of these IBPs in 110 different species of panarthropods by applying profile hidden Markov models (pHMMs) and the best reciprocal hit criterion. This procedure proved necessary to account for the fact that some IBPs are comprised of multiple zinc finger domains. These domains are found in various chromatin binding proteins [41, 42] and are not specific to IBPs.

Table 3 BUSCO assessment for completeness of the 100 1KITE transcriptomes

	Complete [%]	Fragmented [%]	Missing [%]
Min	15.3	3.8	14.7
1st Qu.	49.0	9.3	22.4
Median	57.9	11.0	30.7
Mean	57.3	11.0	31.8
3rd Qu.	68.6	12.5	37.9
Max	81.2	19.0	72.5

Given are the proportions of complete, fragmented and missing BUSCO genes

Since our pHMMs were constructed from IBP amino acid sequences of primarily dipteran species, we can expect a taxonomic bias in the analysis. However, this caveat was unavoidable, since many of these proteins have not been detected in other insect species yet.

Since the IBP CTCF is expected to occur in all Bilateria, we used it to assess the sensitivity of our search strategy and the quality of the analyzed transcript libraries. As expected, we identified transcripts of *CTCF* in almost all analyzed transcript assemblies, confirming the ubiquitous occurrence of this IBP in arthropods. We also found the zinc finger protein Su(Hw) in all major investigated arthropod lineages. Ancestral state reconstruction suggests that Su(Hw) evolved in the last common ancestor of Euarthropoda. We further inferred that the BTB domain protein CP190 evolved either in the last common ancestor, or during the early radiation of Pancrustacea. Consequently, the sequences encoding for CTCF, Su(Hw), and CP190 must have been part of the ancestral gene repertoire of insects, which is in accordance with the current knowledge on the evolution of IBPs [8].

The BTB domain protein GAF was assumed to be unique to holometabolous insects and Hemiptera and was lost secondarily in moths and butterflies [8]. In contrast, we recovered GAF orthologs in nearly all insect orders, except for moths and butterflies (Lepidoptera), caddisflies (Trichoptera), scorpionflies (Mecoptera), fleas (Siphonaptera), twisted wing parasites (Strepsiptera), bark lice and true lice (Psocodea), two-pronged bristletails (Diplura), jumping bristletails (Archaeognatha) and springtails (Collembola). Thus, this pattern suggests that GAF most likely evolved in the last common ancestor of insects and was secondarily lost in some insect lineages. Since GAF was found to play an important role in early embryonic development of *D. melanogaster* [4], it is possible that its expression is

down-regulated in adult individuals of the above lineages (*i.e.,* Lepidoptera, Trichoptera, Mecoptera, Siphonaptera, and Collembola). However, we confirmed the absence of *GAF* in the publicly available draft genome assemblies of *B. mori (Lepidoptera), L. lunatus (Trichoptera), M. hrabei (Archaeognatha),* and *C. aquilonaris (Diplura)* (see Fig. 1). Therefore the absence of *GAF* in the transcriptomes of the aforementioned insect orders corroborates the likely secondary loss of *GAF* in these insect orders. The IBP GAF must have evolved during the Ordovician (509–452 million years ago (mya); [38]), between 106–220 million years earlier than previously thought [8]. While ancestral state reconstruction inferred separate gains of GAF within insects, we deem this scenario highly unlikely. We furthermore investigated the transcriptomes for the vertebrate GAF sequence, but were unable to infer an orthologous relationship between the best hits in insects and the vertebrate sequences.

The occurrence of the zinc finger protein Pita in holometabolous insects, previously only known from *D. melanogaster*, suggests that it was already present in the last common ancestor of Holometabola. Since Pita has previously been investigated only in Diptera [43], our data represent the first evidence for a much older evolutionary origin (Carboniferous, 372–317 mya) and a wider taxonomic distribution of this gene in insects.

Mod(mdg4) is another example of an IBP that shows a much wider taxonomic distribution than previously thought. The data available to Heger and colleagues [8] led the authors to the conclusion that *Mod(mdg4)* likely evolved in the last common ancestor of Aparaglossata (all Holometabola, excluding Hymenoptera). The presence of *Mod(mdg4)* transcripts in various polyneopteran insect lineages suggests, however, that *Mod(mdg4)* must have evolved in the stem lineage of Neoptera (see Fig. 1), whose origin was in the Devonian (413–360 mya) [38]. The occurrence of the FLYWCH domain in sections of coding sequences in the early draft genome of the bristletail *M. hrabei* (Archaeognatha) suggests that *Mod(mdg4)* might have evolved even earlier, within primarily apterygote insects. However, the presence of the FLYWCH domain alone is insufficient to draw solid conclusions, as the domain has also been found in other proteins, such as Su(Kpn) [39].

While most previously discussed IBPs, except for Pita, have already been found in species other than *D. melanogaster*, Zw5 and the proteins discussed in the following section are only known from *D. melanogaster* [7, 8, 43, 44]. Our search for Zw5 in the 1KITE data revealed orthologous transcripts in two additional species of Diptera, *Lipara lucens* (Chloropidae) and *Triarthria setipennis* (Tachinidae). Both belong to the lineage Schizophora, which uses an eversible front pouch to escape from their puparium. This lineage comprises one-third of all extant dipteran species, including those of the genus *Drosophila.* Schizophora diverged from the remaining Diptera in the early Tertiary (65–40 mya; [45]). This distribution is in accordance with the results obtained by Heger and colleagues [8], who found *Zw5* already in another schizophoran fly, *Glossina morsitans*. When searching for *Zw5* transcripts in the 1KITE transcriptome assemblies, we consistently received also transcripts of the protein "meiotic central spindle" (Meics) as promising hits. Both proteins share a similar domain configuration, with Zw5 differing from Meics by having one fewer zinc finger domain. This led us to speculate that *Zw5* could be a paralog of the *meics* gene that evolved within Diptera. We tested this hypothesis by inferring a gene tree from amino acid sequences of Zw5 and Meics, including representatives of Diptera and holometabolous insects. However, in the inferred gene tree (see Additional file 1: Figure S15), Zw5 does not group with the Meics protein subtree. We therefore conclude that *Zw5* is unlikely to be the result of a duplication of *meics* in Diptera.

The IBPs BEAF-32, ZIPIC, Ibf1, Ibf2 as well as the proteins of the Elba protein complex are known only from *D. melanogaster.* We were unable to identify transcripts of these IBPs in any of the analyzed transcriptomes. Since BEAF-32 contains the BESS domain only known from *Drosophila* [33–35], chances of finding the gene in non-dipterans seem to be low, and previous reports already concluded that BEAF-32 is likely being restricted to species of the genus Drosophila [7, 8]. Elba1 and Elba2 of the tripartite protein complex Elba, each contain a chromatin-binding BEN domain, which is known to occur in invertebrates, vertebrates, and viral proteins [29]. In *D. melanogaster,* expression of genes of the Elba complex is restricted to embryonic development [30]. Thus, the transcriptomes from the 1KITE project, which primarily represent tissue samples from adult insects, may be unsuitable to trace back the evolution of this gene, since they do not cover the appropriate developmental stages. The same might hold true for the zinc finger IBPs ZIPIC, Ibf1, and Ibf2, since our searches for the corresponding coding sequences in the draft genomes of *D. willistoni, Aedes aegypti* and *Anopheles gambiae* (Diptera) only revealed significant hits in *D. wilistoni.* This finding corroborates the idea that the absence of transcripts of these IBPs in the screened 1KITE transcriptomes indeed reflects the actual distribution of these proteins in insect transcriptomes.

We found possible evidence for positive selection in the genes encoding for CTCF and Pita. *CTCF* was seemingly underlying positive selection in the onychophoran branch. This might be an artifact of the d_N/d_S.ratio test however. Long divergence times lead to a saturation of d_S [46, 47]. This results in an increase of ω (*i.e.* the ratio of the nonsynonymous substitution rate and the synonymous substitution rate), which means that positive selection is more likely to be erroneously detected, as could be the case for *CTCF*, for which we analyzed sequence data spanning the entire range of Arthropoda. Evidence for positive selection in Pita

corresponds with the branch lengths in the Pita gene tree (Additional file 1: Figure S5) and suggests that the gene is rapidly evolving. Identification of Pita orthologs consequently proved to be difficult. This opens the possibility that the gene could have evolved even earlier and occurs also in hemimetabolous insects. We might have been unable to identify it properly due to its high amino acid sequence divergence.

The occurrence of IBPs in a wide range of species, or restricted to particular taxa, may provide clues about evolutionarily conserved and evolutionarily labile autonomous transcriptional units. Both phylogenetically older and younger IBPs have been shown to actively insulate regions of the same gene complex. The bithorax complex in *D. melanogaster,* for example, contains binding sites of CTCF, GAF and also of Elba [30, 48]. It is possible that the presence of CTCF, Su(Hw), CP190, and GAF across insects most likely ensures proper transcription of genes in rather conserved units and regions (*e.g.*, genes that share an evolutionary conserved gene neighborhood and/or that are in close spatial proximity to, at least temporarily, heterochromatic regions). Likewise, we hypothesize that the restricted occurrence of Mod(mdg4), Pita and, in particular, of Zw5, BEAF-32, ZIPIC and the Elba complex may be the result of recent evolutionary changes in the architecture or transcription of genomic regions in the respective insect lineages.

Conclusions

The exceptionally broad taxonomic sampling of whole-body transcriptomes and the sequencing depth of the analyzed transcriptomes of insects from the 1KITE project proved to be useful for screening and delineating the occurrence of IBPs in arthropods. Our search for and identification of IBPs in all currently recognized extant insect orders implies that the enhancer-blocking IBPs CTCF, Su(Hw), CP190, and GAF were already present in the last common ancestor of insects. The evolution of two insect-specific IBPs is associated with the origin of two major insect lineages: Mod(mdg4) with evolution of Neoptera (413-360 mya) and Pita with the evolution of Holometabola (372-317 mya). Finally, the IBPs Zw5, BEAF-32, and ZIPIC as well as the IBPs of the Elba complex are apparently restricted to Diptera, with BEAF-32, ZIPIC, and Elba possibly being unique to drosophilids. Considering the likely fundamental importance of IBPs for maintaining proper transcription of genes in a frequently altering genomic environment, the currently known diversity of IBPs in *D. melanogaster* likely still represents a lower estimate of the actual diversity of IBPs in flies. The large number of IBPs that are seemingly unique to drosophilids furthermore implies that, if IBP diversity in drosophilids is representative for a given insect lineage with a given age, a plethora of IBPs is yet to be discovered in other insect lineages.

Methods

Transcript libraries and draft genomes

We screened the transcriptomic assemblies of 100 insect (Hexapoda) species sequenced by Misof and colleagues [38] in the 1KITE project for potential transcripts orthologous to IBP genes known from *D. melanogaster* (accession and version numbers are provided in Additional file 3: Table S2). The 100 analyzed species comprise all currently recognized insect orders. We also studied sequence data of species previously analyzed by Heger and colleagues [8]: two crustaceans (*Daphnia pulex* and *Lepeophtheirus salmonis*), one myriapod (*Strigamia maritima*), one chelicerate (*Ixodes scapularis*), and one nematode (*Trichinella spiralis*). We furthermore analyzed the transcript sequences of one tardigrade (*Hypsibius dujardini*) [49], and four species of onychophorans (*Euperipatoides rowelli*, *Ooperipatus hispidus*, *Principapillatus hitoyensis*, and *Eoperipatus* sp.) [50]. We additionally screened genomes of the following species for IBP-coding genes (see Additional file 2: Table S1 for accession numbers): *Drosophila wilistoni* [51], *Aedes aegypti* [52], *Anopheles gambiae* (Diptera) [53], *Bombyx mori* (Lepidoptera) [54], *Limnephilus lunatus* (Trichoptera), *Machilis hrabei* (Archaeognatha), *Catajapyx aquilonaris* (Diplura), *Ephemera danica* (Ephemeroptera), and *Ladona fulva* (Odonata) [55].

Identification of insulator proteins (IBPs)

We searched the transcriptome assemblies for IBP candidate transcripts using profile hidden Markov models (pHMMs) specific to each IBP. The pHMMs were obtained by first aligning all published amino acid sequences that are orthologous to a given *D. melanogaster* IBP with the program MAFFT using the L-INS-i algorithm (v7.164b) [56]. Specifically, we used the IBP amino acid sequences identified and published by Heger and colleagues [8] for building multiple sequence alignments of CTCF, Su(Hw), Mod(mdg4), GAF, CP190, and Zw5. We additionally retrieved the amino acid sequences of all remaining IBPs from NCBI: BEAF-32 (AFH08082.1), Elba1 (AAF50991.2), Elba2 (AAF51239.1), Elba3 (AAF50989.1), Pita (AAF47025.2), ZIPIC/CG7928 (AAF56994.1), Ibf1 (NP_649875), Ibf2 (NP_649874.1). We subsequently built pHMMs from each multiple sequence alignment with the program hmmbuild of the HMMER software package (version 3.1b) [57]. We then screened each transcriptome assembly with the program hmmsearch (also part of the HMMER package) after translating the transcripts into all six possible reading frames with the program fastatranslate (part of the Exonerate software package version 2.2.0) [58]. Only hits with a global *e*-value $\leq 10^{-14}$ were considered as promising IBP transcript candidates. All IBP candidate transcripts were then reciprocally searched against the non-redundant protein (nr) databases entries of *D. melanogaster* (Diptera), *Bombyx mori* (Lepidoptera), *Camponotus*

floridanus (Hymenoptera), and *Zootermopsis nevadensis* (Isoptera) available at NCBI between January and March 2016 using BLASTP [59] in order to identify best reciprocal genome/transcriptome-wide hits. We considered those identified transcripts orthologous to a specific IBP for which the reciprocal search found the same IBP as best reciprocal database-wide hit. The identified IBP transcripts were subsequently aligned at the transcriptional level with the MAFFT L-INS-i algorithm. If the absence of transcripts suggested a possible IBP-coding gene loss, we searched (draft) genomes with TBLASTN (part of the BLAST+ program suite version 2.2.31) for possible coding sequences of the target proteins.

Domain identification

To annotate the domains within amino acid sequences, we used pHMMs of protein family domains compiled in the Pfam-A database (Release 29.0) [60]. All candidate transcripts of IBPs were searched for protein domains with the program hmmscan (part of the HMMER package) [57] employing the above pHMMs.

Transcriptome completeness assessment

To assess transcriptome assembly completeness, we used BUSCO [40] to search for a set of 2675 conserved genes that are near-universal single copy orthologs in arthropods. These genes are present in single-copy in 95 % of the arthropod species in the OrthoDB database and serve as a benchmark for genome or transcriptome completeness. BUSCO uses a combination of BLAST, pHMMs and a gene model refinement procedure to identify and discriminate present, duplicated, fragmented and missing genes in the searched nucleotide sequence database.

Ancestral state reconstruction

Ancestral state reconstruction was applied in order to infer a hypothesis about the evolutionary gains, or losses, of all IBPs. We compiled a matrix, in which we coded the presence and absence of transcripts of each IBP in each species studied. We used Mesquite (version 3.03; http://mesquiteproject.org) [61] to map the gains and losses of insulator proteins on the phylogenetic tree of insects and added the phylogenetic relationships among outgroup taxa (*i.e.*, Onychophora, Tardigrada, and Nematoda) according to Meusemann and colleagues [38, 62] under the Maximum Parsimony optimality criterion. Note that Mesquite does not allow Ancestral state reconstruction under Dollo's parsimony criterion.

Phylogenetic analyses

To better assess the possible occurrence of gene duplication events, we inferred gene trees from the identified putative transcripts of each IBP. For this purpose, we inferred for each IBP a Maximum Likelihood phylogenetic tree based on the corresponding multiple sequence alignment with the program PhyML (version 3.0) [63], using the WAG + Γ substitution model with default settings. Tree robustness was assessed from 1000 bootstrap replicates. We applied the same method when testing whether or not *Zw5* could be a Diptera-specific paralog of the gene *meics*. Specifically, we aligned all available amino acid sequences of Zw5 to the amino acid sequences of Meics of holometabolous insects. We retrieved the latter sequences from OrthoDB (version 8) [64]. Phylogenetic analysis was done as described in the preceding paragraph.

Modes of selection

To search for evidence of positive or negative selection on insulator protein genes, we used the program codeML of the PAML package (version 4.8) [65] to measure the ratio of non-synonymous (amino acid replacing) to synonymous (silent) substitutions (ω). For this purpose, we compiled corresponding nucleotide multiple sequence alignments of the identified transcripts for each IBP separately with Pal2Nal (version 14) [66] by using the multiple sequence alignments of the translated transcripts as blueprints. We used a branch site model, in which ω is allowed to vary along specific branches of the phylogenetic tree, to test for positive selection along these branches. We specifically tested for changes of ω along branches that immediately followed nodes at which we inferred the evolutionary origin of a specific IBP. We used a likelihood ratio test with one degree of freedom to test models, in which ω was allowed to vary along a specific branch, against the null model, in which ω was kept at 1 in all branches of the phylogenetic tree. For each gene, we used the same tree topology as in Fig. 1. Species in which we did not find orthologs of the respective gene were pruned from the tree.

Endnotes

[1]We are using the term insects in a broad sense, including all Hexapoda, equivalent to the nomenclature used in [46].

Additional files

Additional file 1: Figure S1. Tracing the evolutionary origin of CTCF with ancestral state reconstruction. Figure S2. Tracing the evolutionary origin of Su(Hw) with ancestral state reconstruction. Figure S3. Tracing the evolutionary origin of CP190 with ancestral state reconstruction. Figure S4. Tracing the evolutionary origin of GAF with ancestral state reconstruction. Figure S5. Tracing the evolutionary origin of Pita with ancestral state reconstruction. Figure S6. Tracing the evolutionary origin of Mod(mdg4) with ancestral state reconstruction. Figure S7. Tracing the evolutionary origin of Zw5 with ancestral state reconstruction. Figure S8. Phylogenetic gene tree of *CTCF* orthologs. Figure S9. Phylogenetic gene tree of *Su(Hw)* orthologs. Figure S10. Phylogenetic gene tree of *CP190* orthologs. Figure S11. Phylogenetic gene tree of *GAF* orthologs. Figure S12. Phylogenetic gene tree of *Pita* orthologs. Figure S13. Phylogenetic gene tree of *Mod(mdg4)* orthologs. Figure S14. Phylogenetic gene tree of *Zw5* orthologs. Figure S15. Phylogenetic analysis of Zw5 and *meiotic central spindle* (Meics). (PDF 474 kb)

Additional file 2: Table S1. BUSCO assessment of the 1KITE transcriptomes. (XLS 21 kb)

Additional file 3: Table S2. NCBI accession numbers of transcriptome and genome data. (XLS 35 kb)

Abbreviations

BEAF-32: Boundary element associated factor of 32 kD;
BUSCO: Benchmarking universal single-copy orthologs; CP190: Centrosomal protein 190 kD; CTCF: CCCTC binding factor; Elba: Early boundary activity; GAF: GAGA-Factor; Ibf1: Insulator binding factor 1; Ibf2: Insulator binding factor 2; IBP: Insulator binding protein; Mod(mdg4): Modifier of mdg4; pHMM: Profile Hidden Markov Model; Su(Hw): Suppressor of hairy wing; ZIPIC: Zinc-finger protein interacting with CP190; Zw5: Zeste white 5

Acknowledgments

This manuscript has been enabled by the 1KITE consortium and the i5K initiative. Alexander Donath, Christoph Mayer, Bernhard Misof, Oliver Niehuis, and Ralph S. Peters furthermore acknowledge the Leibniz association for installing the graduate school for Genomic Biodiversity Research, in which the present study arose. We especially thank Stephen Richards and Richard Gibbs of the Baylor College of Medicine Human Genome Sequencing Center for granting access to i5K pilot data prior to their official publication.

Funding

The sequencing and assembly of the 1KITE transcriptomes were funded by BGI through support to the China National GeneBank. TW and PH were supported by grants from the German Research Foundation (Sonderforschungsbereich 680).

Availability of data and materials

The accession numbers of sequence data used in this study is given in Additional file 3: Table S2. In the rare cases in which no accession number is given, a download link or contact information of a responsible person is given instead. Additionally we made the following data publicly available: (1) The amino acid alignments we based the inference of our gene trees on, (2) the gene trees, (3) the nucleotide alignments we based our inference of d_N/d_S-ratios on. The data is available from the Dryad digital repository: http://dx.doi.org/10.5061/dryad.f4r38.

Authors' contributions

Participated in the design of the study: BM, ON, TP. Contributed data: AD, BM, CM, GM, KM, ON LH, MP, LP, PH, RSP, SL, TW, XZ. Performed data analysis: AD, CM, DD, KM, MP, LP, LV, TP. Manuscript preparation: all authors contributed to the writing of the manuscript, with BM, ON, and TP taking the lead. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

This section is not applicable to the present study.

Ethics approval and consent to participate

This section is not applicable to the present study.

Author details

[1]Center of Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 51113 Bonn, Germany. [2]University of Tübingen, Geschwister-Scholl-Platz, 72074 Tübingen, Germany. [3]Johannes Gutenberg University Mainz, Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [4]Department for Evolutionary Biology and Ecology (Institut for Biology I, Zoology), University of Freiburg, Hauptstr. 1, 79104 Freiburg, Germany. [5]Australian National Insect Collection, CSIRO National Research Collections Australia, Clunies Ross Street, Acton, ACT 2601, Australia. [6]Zoological Research Museum Alexander Koenig, Arthropod Department, Adenauerallee 160, 53113 Bonn, Germany. [7]University of Bonn, Institute of Evolutionary Biology and Ecology, An der Immenburg 1, 53121 Bonn, Germany. [8]China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen, Guangdong Province 518083, China. [9]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark. [10]Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, China. [11]College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, China. [12]University of Cologne, Cologne Biocenter, Institute for Genetics, Zülpicher Straße 47a, 50674 Köln, Germany. [13]Department of Zoology, University of Kassel, Heinrich-Plett-Str. 40, 34132 Kassel, Germany.

Received: 7 June 2016 Accepted: 25 October 2016
Published online: 03 November 2016

References

1. Wallace JA, Felsenfeld G. We gather together: insulators and genome organization. Curr Opin Genetics Dev. 2007;17(5):400–7.
2. Hou C, Li L, Qin ZS, Corces VG. Gene Density, Transcription and Insulators contribute to the partition of the *Drosophila* genome into physical domains. Mol Cell. 2012;3:471–84.
3. Yang J, Corces VG. Insulators, long-range interactions, and genome function. Curr Opin Genet Dev. 2012;22(2):86–92.
4. Bhat KM, Farkas G, Karch F, Gyurkovics H, Gausz J, et al. The GAGA factor is required in the early *Drosophila* embryo not only for transcriptional regulation but also for nuclear division. Development. 1996;122:1113–24.
5. Mohan M, Bartkuhn M, Herold M, Philippen A, Heinl N, et al. The *Drosophila* insulator proteins CTCF and CP190 link enhancer blocking to body patterning. EMBO J. 2007;26:4203–14.
6. Roy S, Jiang N, Hart CM. Lack of the *Drosophila BEAF* insulator proteins alters regulation of genes in the antennapedia complex. Mol Genet Genomics. 2011;285:113–23.
7. Schoborg TA, Labrador M. The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. J Mol Evol. 2010;70:74–84.
8. Heger P, George R, Wiehe T. Successive gain of insulator proteins in arthropod evolution. Evolution(N Y). 2013;67:2945–56.
9. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998;291:60–3.
10. Kellum R, Schedl P. A position-effect assay for boundaries of higher order chromosomal domains. Cell. 1991;64(5):941–50.
11. Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet. 2006;7:703–13.
12. Burgesse-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, et al. The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci USA. 2002;99:16433–7.
13. Neufeld EJ, Skalnik DG, Lievens PMJ, Orkin SH. Human CCAAT displacement protein is homologous to the Drosophila homeoprotein, *cut*. Nature Genetics. 1992;1:50–5.
14. Lin N, Li X, Cui K, Chepelev I, Tie F, et al. A barrier-only boundary element delimits the formation of facultative heterochromatin in Drosophila melanogaster and vertebrates. Mol Cell Biol. 2011;31:2729–41.
15. Donze D, Adams CR, Rine J, Kamakaka RT. The boundaries of the silenced HMR domain in *Saccharomyces cerevisiae*. Genes Dev. 1999;13:698–708.
16. Donze D, Kamakaka RT. RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in *Saccharomyces cerevisiae*. EMBO J. 2001;20:520–31.
17. Noma KI, Cam HP, Maraia RJ, Grewal SIS. A role for *TFIIIC* transcription factor complex in genome organization. Cell. 2006;125:859–72.
18. Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, et al. Human tRNA genes function as chromatin insulators. EMBO J. 2012;31:330–50.
19. Heger P, Wiehe T. New tools in the box: An evolutionary synopsis of chromatin insulators. Trends Genet. 2014;30:161–70.
20. Heger P, Marin B, Bartkuhn M, Schierenberg E, Wiehe T. The chromatin insulator CTCF and the emergence of metazoan diversity. Proc Natl Acad Sci. 2012;109:17507–12.
21. Van Bortle K, Corces VG. The role of chromatin insulators in nuclear architecture and genome function. Curr Opin Genet Dev. 2013;23:212–8.
22. Parkhurst SM, Harrison DA, Remington MP, Spana C, Kelley RL, et al. The *Drosophila* su(Hw) gene, which controls the phenotypic effect of the gypsy transposable element, encodes a putative DNA-binding protein. Genes Dev. 1988;2:1205–15.
23. Spana C, Harrison DA, Corces VG. The *Drosophila melanogaster* suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. Genes Dev. 1999;2:1414–23.

24. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell. 1999;98(3):387–96.
25. Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju BG. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. Science. 2007;317:248–51.
26. Román AC, Gonzáles-Rico FJ, Moltó E, Hernando H, Neto A. Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. Genome Res. 2011;21:422–32.
27. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. 2014;15(6):R82.
28. Matharu NK, Hussain T, Sankaranarayanan R, Mishra RK. Vertebrate homologue of *Drosophila* GAGA factor. J Mol Biol. 2010;400(3):434–47.
29. Abhiman S, Iyer LM, Aravind L. BEN: a novel domain in chromatin factors and DNA viral proteins. Bioinformatics. 2008;24:458–61.
30. Aoki T, Sarkeshik A, Yates J, Schedl P. Elba, a novel developmentally regulated chromatin boundary factor is a hetero-tripartite DNA binding complex. Elife. 2012;2012:1–24.
31. Zollman S, Godt D, Privé GG, Couderc JL, Laski FA. The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in *Drosophila*. P Natl Acad Sci USA. 1994;91:10717–21.
32. Pai CY, Lei EP, Ghosh D, Corces VG. The Centrosomal Protein CP190 is a component of the *gypsy* chromatin insulator. Mol Cell. 2004;16(5):737–48.
33. Reuter G, Giarre M, Farah J, Gausz J, Spierer A, et al. Dependence of position-effect variegation in *Drosophila* on dose of a gene encoding an unusual zinc-finger protein. Nature. 1990;344:219–23.
34. Zhao K, Hart CM, Laemmli UK. Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell. 1995;81:879–89.
35. Clark KA, McKearin DM. The *Drosophila* stonewall gene encodes a putative transcription factor essential for germ cell development. Development. 1996;122:937–50.
36. Hsu S-J, Plata MP, Ernest B, Asgarifar S, Labrador M. The insulator protein *Suppressor of Hairy wing* is required for proper ring canal development during oogenesis in *Drosophila*. Dev Biol. 2015;403(1):57–68.
37. Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, et al. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol Biol. 2014;14:52.
38. Misof B, Liu S, Meusemann K, Peters RS, Donath A, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346:763–7.
39. Provost E, Shearn A. The suppressor of killer of prune, a unique glutathione S-transferase. J Bioenerg Biomembr. 2006;38:189–95.
40. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
41. Klug A, Rhodes D. Zinc fingers: a novel protein fold for nucleic acid recognition. Cold Spring Harb Symp Quant Biol. 1987;52:473–82.
42. Klug A. The discovery of zinc fingers and their applications in genome manipulation. Annu Rev Biochem. 2010;79:213–31.
43. Maksimenko O, Bartkuhn M, Stakhov V, Herold M, Zolotarev N, et al. Two new insulator proteins, Pita and *ZIPIC*, target CP190 to chromatin. Genome Res. 2015;25:89–99.
44. Cuartero S, Fresán A, Reina O, Planet E, Espinàs ML. Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. EMBO J. 2014; 33(6):637–47.
45. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J, et al. Episodic radiations in the fly tree of life. Proc Natl Acad Sci U S A. 2011;108:5690–5.
46. Gojobori T. Codon substitution in evolution and the "saturation" of synonymous changes. Genetics. 1983;105:1011–27.
47. Smith JM, Smith NH. Synonymous nucleotide divergence: what is "saturation"? Genetics. 1996;142:1033–6.
48. Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M, et al. CTCF Genomic Binding Sites in *Drosophila* and the organisation of the bithorax complex. PLOS Genet. 2007;3(7):e112.
49. Hering L, Meyer G. Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in Panarthropoda. Genome Biol Evol. 2014;6(9):2380–91. *Bioinformatics* 14(9):755-763.
50. Hering L, Henze MJ, Kohler M, Kelber A, Bleidorn C, et al. Opsins in Onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods. Mol Biol Evol. 2012;29:3451–8.
51. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 2007;450:203–18.
52. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science. 2007;316:1718–23.
53. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. The genome sequence of the malaria mosquito anopheles gambiae. Science. 2002;298:129–49.
54. Goldsmith MR, Shimada T, Abe H. The genetics and genomics of the silkworm, *Bombyx mori*. Annu Rev Entomol. 2005;50:71–100.
55. Evans JD, Brown SJ, Hackett KJJ, Robinson G, Richards S, et al. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 2013;104:595–600.
56. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
57. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–63.
58. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
59. Altschul S, Gish W, Miller W. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
60. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:222–30.
61. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. Version 2.75.2011. 2015. http://mesquiteproject.org. Accessed Feb 2016.
62. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, et al. A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol. 2010;27(11):2451–64.
63. Guindon S, Gascuel O, Dufayard J-F, Lefort V, Anisimova M, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):1–37.
64. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao F, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. 2014;43:D250–6.
65. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
66. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:609–12.
67. Gaszner M, Vazquez J, Schedl P. The Zw5 protein, a component of the *scs* chromatin domain boundary, is able to block enhancer–promoter interaction. Genes Dev. 1999;13(16):2098–107.
68. Whitfield WG, Chaplin MA, Oegema K, Parry H, Glover DM. The 190 kDa centrosome-associated protein of Drosophila melanogaster contains four zinc finger motifs and binds to specific sites on polytene chromosomes. J Cell Sci. 1995;108:3377–87.
69. Omichinski JG, Pedone PV, Felsenfeld G, Gronenborn AM, Clore GM. The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. Nat Struct Biol. 1997;4:122–32.
70. Ohtsuki S, Levine M. GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. Genes Dev. 1998;12(21):3325–30.
71. Gerasimova TI, Gdula D a, Gerasimov DV, Simonova O, Corces VG. A *Drosophila* protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. Cell. 1995;82:587–97.
72. Dorn R, Krauss V. The modifier of mdg4 locus in *Drosophila*: functional complexity is resolved by trans splicing. Genetica. 2003;117:165–77.

BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design

Christoph Mayer,[†,1] Manuela Sann,[†,1,2] Alexander Donath,[1] Martin Meixner,[3] Lars Podsiadlowski,[4] Ralph S. Peters,[5] Malte Petersen,[1] Karen Meusemann,[1,6] Karsten Liere,[3] Johann-Wolfgang Wägele,[7] Bernhard Misof,[1] Christoph Bleidorn,[8,9,10] Michael Ohl,[*,2] and Oliver Niehuis[*,1]

[1]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Bonn, Germany

[2]Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

[3]Services in Molecular Biology GmbH, Rüdersdorf, Germany

[4]University of Bonn, Institute of Evolutionary Biology and Ecology, Bonn, Germany

[5]Department Arthropoda, Zoological Research Museum Alexander Koenig, Bonn, Germany

[6]Australian National Insect Collection, CSIRO National Research Collections Australia, Acton, Canberra, ACT, Australia

[7]Zoological Research Museum Alexander Koenig, Bonn, Germany

[8]Molecular Evolution and Systematics of Animals, Institute for Biology, University of Leipzig, Leipzig, Germany

[9]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[10]Museo Nacional de Ciencias Naturales, Spanish National Research Council (CSIC), Madrid, Spain

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: o.niehuis@zfmk.de; michael.ohl@mfn-berlin.de.

Associate editor: Michael Rosenberg

Abstract

Target DNA enrichment combined with high-throughput sequencing technologies is a powerful approach to probing a large number of loci in genomes of interest. However, software algorithms that explicitly consider nucleotide sequence information of target loci in multiple reference species for optimizing design of target enrichment baits to be applicable across a wide range of species have not been developed. Here we present an algorithm that infers target DNA enrichment baits from multiple nucleotide sequence alignments. By applying clustering methods and the combinatorial 1-center sequence optimization to bait design, we are able to minimize the total number of baits required to efficiently probe target loci in multiple species. Consequently, more loci can be probed across species with a given number of baits. Using transcript sequences of 24 apoid wasps (Hymenoptera: Crabronidae, Sphecidae) from the 1KITE project and the gene models of *Nasonia vitripennis*, we inferred 57,650, 120-bp-long baits for capturing 378 coding sequence sections of 282 genes in apoid wasps. Illumina reduced-representation library sequencing confirmed successful enrichment of the target DNA when applying these baits to DNA of various apoid wasps. The designed baits furthermore enriched a major fraction of the target DNA in distantly related Hymenoptera, such as Formicidae and Chalcidoidea, highlighting the baits' broad taxonomic applicability. The availability of baits with broad taxonomic applicability is of major interest in numerous disciplines, ranging from phylogenetics to biodiversity monitoring. We implemented our new approach in a software package, called BaitFisher, which is open source and freely available at https://github.com/cmayer/BaitFisher-package.git.

Key words: **hybrid enrichment, comparative genomics, phylogenetics, phylogenomics, Hymenoptera.**

Article

Introduction

Target DNA enrichment combined with high-throughput sequencing technology is a highly promising approach to studying and characterizing a large number of loci in genomes, at reasonable costs. Target DNA enrichment comprises various molecular techniques that augment target DNA in a given next-generation sequencing (NGS) library by means of oligonucleotide probes (hereafter also synonymously referred to as baits), either in solution (Faircloth et al. 2012; Lemmon et al. 2012) or on an array (Albert et al. 2007; Hodges et al. 2007, 2009; Liu et al. 2016). The nucleotide sequences of these baits are selected for high nucleotide sequence similarity to target DNA sequence sections of interest. The baits can then be hybridized to the target sequence sections in a DNA sample, which allows enriching these sequence sections. This technique has been named differently depending on which target regions are enriched (e.g., exome or gene capture when exons/coding DNA sequences are enriched [Ng et al. 2009; Cosart et al. 2011; Fisher et al. 2011; Li et al. 2013]; anchored hybrid enrichment when the flanking region of [ultra] conserved regions are of interest [Bejerano et al. 2004; Crawford et al. 2012; Faircloth et al. 2012, 2014; Lemmon et al. 2012; Bragg et al. 2015; Hawkins et al. 2015; Vinner et al. 2015];

hyRAD when specific RAD segments are enriched [Suchan et al. 2016]). Various laboratory protocols for enriching target loci have been developed (Bashiardes et al. 2005; Blumenstiel et al. 2010; Meyer and Kircher 2010; Bodi et al. 2013; Peñalba et al. 2014). Furthermore, molecular procedures have been described, which allow the capture of more dissimilar target loci with a given set of baits and extend the reach of the method considerably (Li et al. 2013; Paijmans et al. 2016). However, because target locus enrichment efficacy decreases with increasing bait-to-target DNA sequence distance (Bragg et al. 2015; Hawkins et al. 2015; Paijmans et al. 2016; present study), design of bait sets to be applied across a range of distantly related species can still pose a challenge. For example, a given bait can exhibit a high nucleotide sequence similarity to the target DNA of species in one ingroup lineage and consequently effectively enrich the target DNA in species of this lineage. But if the same bait differs significantly from the target DNA in species of another ingroup lineage, it will not enrich it to the same extent (or at all) in the species of the second lineage. In such a situation, one might want to design more than one bait to cope with the significant ingroup target locus sequence divergence and thereby improving the odds that the target locus is evenly enriched across all ingroup species.

No software algorithm is available so far that allows formally optimizing the number of baits for enriching target loci across a diverse group of species by dynamically adjusting the number of baits to the known taxonomic ingroup target locus divergence. Ideally, baits are developed by exploiting target locus nucleotide sequence information from multiple reference species that representatively capture ingroup nucleotide sequence divergence of all target loci. Baits should then be designed in a way that 1) for every target locus and reference species there is a bait that differs in less than a user-defined nucleotide sequence similarity threshold value from the target DNA in the reference species and 2) the total number of baits that fulfil criterion 1) is minimized. There is a growing need for such an approach, because the costs for bait sets scale with the number of different baits in such a set and comparative (phylo-)genomic studies, in which target genes are sampled across a wide range of species, are frequently conducted (Bejerano et al. 2004; Faircloth et al. 2012, 2014; Lemmon et al. 2012; Li et al. 2013; McCormack, Harvey, et al. 2013; McCormack, Hird, et al. 2013; Bragg et al. 2015; Hawkins et al. 2015; Hugall et al. 2015; Prum et al. 2015; Vinner et al. 2015).

Here we present a novel approach for the design of hybridization baits to be applied to DNA of a range of species. It infers baits by exploiting user-provided nucleotide sequence information of target loci in a representative set of species. It optimizes the total number and the nucleotide sequences of baits so that for each target locus in a reference species there is exactly one designed bait that differs in less than the user-defined bait-to-target nucleotide sequence similarity threshold value from the target locus. It furthermore allows the user to specify any intended tiling design and to thus compensate edge effects that may arise from shifts of, for example, exon–intron boundaries and other local but substantial changes in the target DNA (Bi et al. 2012). We implemented this approach in a software package called BaitFisher, which comprises two programs. The first one, BaitFisher, provides all possible bait designs suitable for enriching a given target locus (e.g., a gene or the exon of a specific gene). The output from BaitFisher can be passed to the second program, BaitFilter, for selecting a specific bait set. BaitFilter enables choosing the optimal start position for a given tiling design in a given locus based on a user-specified optimality criterion (i.e., minimizing the number of baits required to enrich the target locus or maximizing the number of nucleotide sequences which baits were inferred from). BaitFilter is also able to assess whether or not baits are likely to bind to multiple genomic regions and whether baits are likely to bind to contiguous genomic DNA (e.g., in case the applied gene model of a reference species was not correct when extracting CDS regions from user-provided transcript sequences). Both procedures require a user-provided reference species genome assembly.

We empirically tested baits inferred with BaitFisher in a pilot project for a study on the phylogeny of apoid wasps and bees, exploiting the comprehensive transcriptome libraries of insects compiled in the international 1KITE project (www.1kite.org). We present the result from this pilot project, which demonstrates that the BaitFisher-inferred baits were able to efficiently enrich a major fraction of the target genes in the taxonomic target group. Our results furthermore provide insights into what bait-to-target distance threshold value to choose when designing baits with the BaitFisher software in order to ensure that target loci are consistently and effectively enriched across species when applying molecular procedures similar to those used by us.

New Approaches

Software Implementation

BaitFisher designs baits for target DNA enrichment on the basis of multiple nucleotide sequence alignments that contain contiguous template DNA. Suitable templates are 1) genomic DNA sequences (gDNA) for designing baits that are meant to enrich gDNA, 2) transcript-complementary DNA (cDNA) sequences for designing baits that are meant to enrich cDNA, or 3) cDNA sequences for designing baits that are meant to enrich gDNA (fig. 1). When using the latter as templates, the software requires nucleotide sequence alignments that additionally include the cDNA sequence of a user-defined reference species with an annotated genome. By providing the genome assembly and corresponding gene feature format (GFF) annotation file of the reference species to BaitFisher, the program is able to split aligned cDNA sequences into genome-feature-specific sections, such as exons or coding sequence (CDS) regions (fig. 1, step 1). Specifically, the program uses the gene ID of the reference species in the multiple cDNA alignment to fetch nucleotide sequences from each corresponding feature (e.g., exons or CDS sections) in the assembled genome by using the corresponding coordinates in the GFF file. Each retrieved genomic feature sequence is subsequently aligned to the cDNA sequence of the reference species with the Needleman–

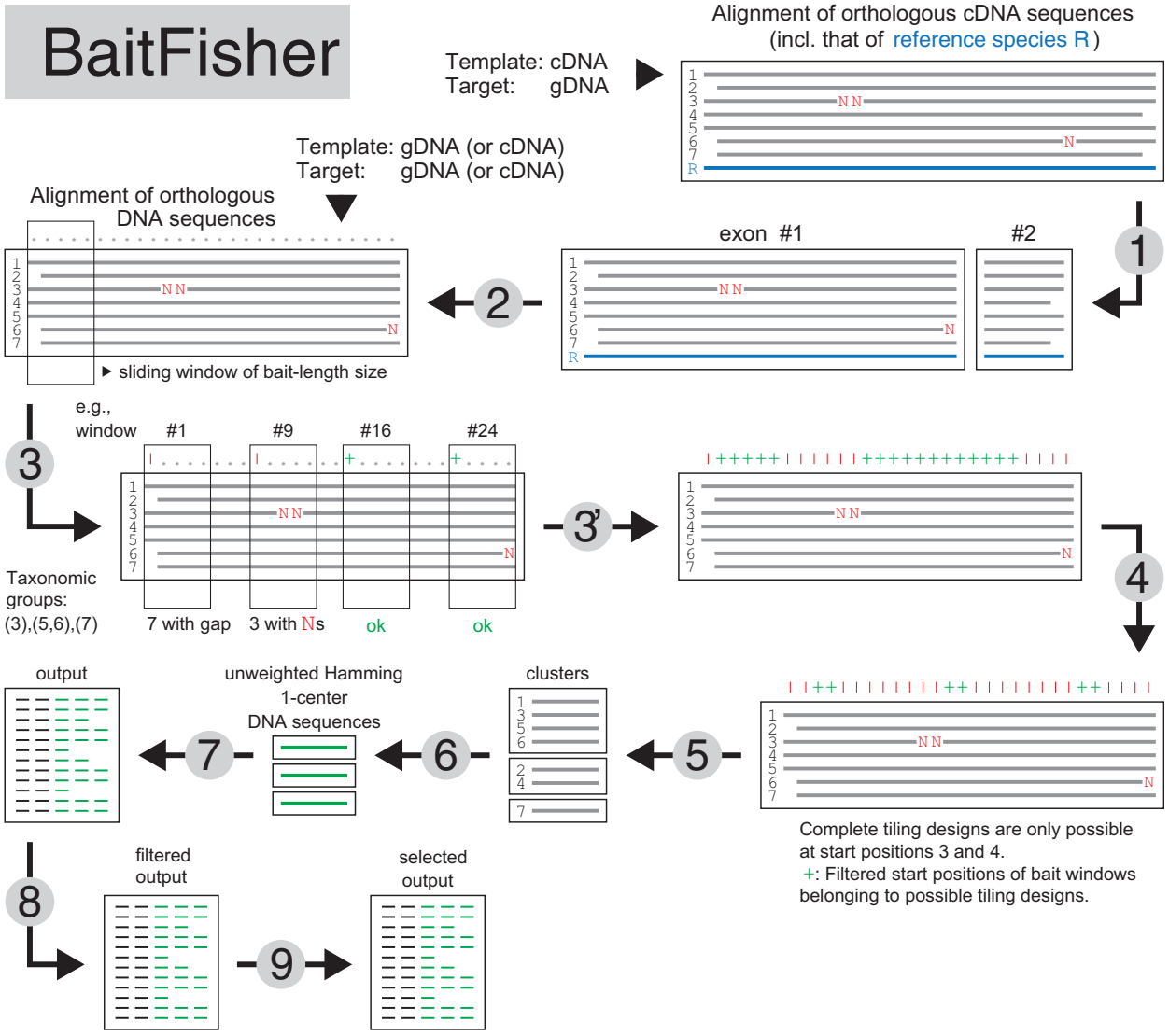

Fig. 1. Procedure for designing target DNA enrichment probes (= baits) as implemented in BaitFisher. MSAs that directly serve as templates for bait design are used as input. Alternatively, the user can provide MSAs of cDNA that are afterwards split into individual exons/CDS region based on the gene models of a user-defined reference species (R). To enable the latter approach, the MSAs must include the cDNA of a reference species with a sequenced and annotated genome. The cDNA sequence plus the genome assembly and official gene set of the reference species (all user-provided) allow identifying exon boundaries in the MSAs and split of the latter according to these boundaries (step 1). The cDNA sequence of the reference species can be optionally discarded (step 2). BaitFisher next identifies with a sliding window of bait-length size (5 bp in the illustrated example) start positions in the MSA that are deemed suitable for bait design (+) (step 3). Suitable start positions are those with windows in which user-defined taxonomic groups are represented by at least one gap- and ambiguity code-free nucleotide sequence. In the example shown, after having removed all sequences with gaps and/or ambiguity codes from a given window, it must still include the nucleotide sequences of taxon 3 and taxon 7 and the nucleotide sequence of taxon 5 or taxon 6. The nucleotide sequences of all remaining taxa (1, 2, and 4) are considered during bait design if they are gap- and ambiguity code-free, but their presence is not mandatory. After all windows have been analyzed (step 3'), BaitFisher filters positively evaluated start positions for those compatible with a user-defined tiling design (step 4). In the example given, the tiling design requires three consecutive baits of 5-bp length with a new bait every 10 bp. From the gap- and ambiguity code-free nucleotide sequences of each retained positively evaluated window, BaitFisher clusters sequences according to a user-defined degree of nucleotide sequence similarity (step 5). It then calculates the 1-center nucleotide sequence (= bait) of each cluster (step 6). Finally, information about all inferred baits is summarized (step 7). The inferred baits can be optionally searched with the BaitFilter helper program against the genome assembly of a user-selected reference species to identify and remove potentially non target-binding baits (step 8). BaitFilter program furthermore allows selecting one optimal set of tiled baits per exon/CDS region or gene, based on a user-selected optimality criterion (step 9).

Wunsch algorithm (Needleman and Wunsch 1970). The region in the multiple sequence alignment (MSA), to which the genomic feature sequence is aligned, is subsequently extracted and stored in a separate file. BaitFisher allows the user to decide whether the nucleotide sequence of the reference species is subsequently also considered for designing baits (fig. 1, step 2).

BaitFisher identifies regions in MSAs of contiguous potential target DNA that are deemed suitable for bait design (fig. 1, steps 3 and 4). Specifically, for every MSA window of the

user-defined bait length, BaitFisher first discards sequences with gaps and/or IUPAC ambiguity codes. BaitFisher then evaluates whether all user-defined taxonomic groups (fig. 1, step 3) are represented in a given MSA window. Only if they are present does BaitFisher mark this window as suitable for bait design. Finally, the software filters for those start positions that are compatible with the user-defined tiling design (e.g., three baits with a new bait every 10 bp; fig. 1, step 4).

In order to minimize the number of baits required to efficiently enrich all nucleotide sequences that are part of the MSA in a given window, BaitFisher infers baits in two steps: The software first calculates the uncorrected (= Hamming; p) distances between all sequences in a given window of bait-length size and clusters those sequences that differ by less than a user-defined maximum distance (e.g., 0.06) from each other (fig. 1, step 5). BaitFisher then infers from the nucleotide sequences of each cluster an artificial 1-center sequence (fig. 1, step 6). This 1-center sequence represents an artificial sequence that exhibits the smallest maximum distance to all nucleotide sequences in this cluster (Li et al. 2002). In case of multiple equivalent solutions, the software randomly picks a sequence from the pool of equivalent 1-center sequences. A detailed description of the 1-center problem and the algorithm used to compute the 1-center sequence is given in supplementary file S1, Supplementary Material online.

After having calculated all 1-center nucleotide sequences, BaitFisher provides the user a tab-delimited text file that contains the essential information about each possible bait region (fig. 1, step 7). A bait region is a sequence segment in the MSA that 1) hosts a complete tiling design and 2) fully contains the nucleotide sequences of all user-defined mandatory taxonomic groups (see BaitFisher manual for more information). The file lists for each possible start position in a given target DNA region (i.e., user-provided MSA, such as of a gene, or excised feature) an optimal set of baits compatible with the user-defined tiling design. A procedure to automatically select an optimal start position in a given target DNA region is described below.

If one or multiple baits of a given bait region exhibit a high sequence similarity to two or more regions in the user-provided reference genome assembly, it is likely that the inferred baits would also enrich nontarget DNA. Hence, the user might want to exclude such bait regions in favor of others that are more target specific. We therefore developed a helper program called BaitFilter, which is part of the BaitFisher software package. BaitFilter allows the user to identify and discard baits that are likely to bind to multiple regions, as judged from the baits' nucleotide sequence similarity to regions in a user-provided reference genome assembly (fig. 1, step 8). The sequence similarity search of baits against the reference genome assembly is accomplished using BLAST+ (Camacho et al. 2008). If a given bait shows no significant similarity to any region in the reference genome assembly, the bait and the corresponding bait region are retained. We hereby acknowledge the fact that the nucleotide sequence of the reference genome might not have been part of the sequence cluster from which the bait was inferred from. Using a similar approach, BaitFilter allows the user also to identify and remove baits that would likely not properly bind to the target DNA (e.g., because gene models used to splice MSAs consisting of cDNA were not correct), as judged from searching baits against the assembled genome of a reference species. Finally, BaitFilter enables selecting an optimal start position for a bait set in a given target DNA region (fig. 1, step 9). The user can apply one of the two optimality criteria for selecting the optimal start position: 1) Minimizing the number of baits required to enrich a given locus, which usually means placing the bait region (i.e, the genomic region spanned by all baits tiled across this region) in the most conserved segment of a given locus; or 2) maximizing the number of sequences that were considered when inferring baits, which results in selecting the bait region, in which the smallest number of nucleotide sequences is missing or contain gaps or ambiguous nucleotides.

The BaitFisher software package is written in the programing language C ++. It is open source and freely available at https://github.com/cmayer/BaitFisher-package.git.

Empirical Evaluation of the Bait Enrichment Capabilities

To assess the capability of baits designed by BaitFisher for enriching target DNA, we inferred a set of 57,650 baits for studying target DNA of apoid wasps (Hymenoptera: Crabonidae, Sphecidae) with the SureSelect Target Enrichment System offered by Agilent Technologies, Inc. (Santa Clara, CA). Specifically, we exploited transcriptomes of adults of 24 apoid wasp species sequenced in the international 1KITE project, and listed in supplementary file S2, Supplementary Material online. By querying the OrthoDB 5 database (Waterhouse et al. 2010), we identified a set of 5,561 genes that likely are single copy in apoid wasps, judged from the genes' presence in a representative set of six Hymenoptera with well-sequenced genomes (i.e., *Acromyrmex echinatior*, *Apis mellifera*, *Camponotus floridanus*, *Harpegnathos saltator*, *Linepithema humile*, *Nasonia vitripennis*; Weinstock et al. 2006; Bonasio et al. 2010; Werren et al. 2010; Nygaard et al. 2011; Smith et al. 2011) (supplementary file S2, Supplementary Material online). We next searched for transcripts that are orthologous to these 5,561 genes in the apoid wasp transcriptomes. For this purpose, we made use of HaMStRad (Misof et al. 2014), a modified version of HaMStR 8 (Ebersberger et al. 2009), following the procedure described by Misof et al. (2014). We used the software Orthograph 0.5.6, which became more recently available (Petersen et al. 2015), to later assign assembled contigs from enriched and sequenced next-generation DNA sequencing libraries to target loci. Orthograph and HaMStRad both rely on the best reciprocal genome/transcriptome-wide hit (BRH) criterion to infer gene-transcript orthology. We only considered transcripts, for which the BRH criterion was fulfilled for each of the six (see above) reference taxa in the reciprocal searches. We used the amino acid sequence output to align orthologous transcripts on the translational level with MAFFT 7.017 (L-INS-i iterative refinement method; Katoh and Standley 2013) and inferred the corresponding nucleotide sequence alignment with the program Pal2Nal (Suyama

et al. 2006) using a version modified as described by Misof et al. (2014).

We split the aligned cDNA sequences of apoid wasp into individual CDS regions using custom scripts which, in optimized form, are now integrated in BaitFisher. These scripts made use of the genome assembly and official gene set version 1.2 of the jewel wasp *N. vitripennis* (Werren et al. 2010) available from the Hymenoptera Genome Database (Muñoz-Torres et al. 2011). We specified a bait length of 120 bp and a tiling design of seven baits spanning a 240-bp window with a new bait every 20 bp. Furthermore, each bait region had to contain the cDNA sequence of at least one representative from each sampled taxonomic subfamily and tribe.

After all possible sets of baits had been inferred with the aid of BaitFisher and the above specified search parameters, we evaluated each bait for its potential to bind to nontarget regions with the BaitFilter program. For the present data set, we used the genomes of *Ap. mellifera* (assembly 4.0; Weinstock et al. 2006), *H. saltator* (assembly 3.3; Bonasio et al. 2010), and *N. vitripennis* (assembly 1.0; Werren et al. 2010) as references (supplementary file S2, Supplementary Material online). We discarded all bait regions that contained a bait that showed a significant match to $\geq$2 different loci in any of the reference genomes. To be more precise, the first BLASTN hit had to have an *E* value $<10^{-8}$ and the second BLASTN hit had to have an *E* value $<10^{-5}$ for the bait region to be considered to bind unspecifically. Finally, we removed baits of 131 CDS regions to lower the total number of baits to 57,650, the maximum number of baits to be included with the SureSelect Target Enrichment System at the time we ordered (July 31, 2013).

Results

Inference of Baits for Studying Target Genes in Apoid Wasps

We found orthologous transcripts to 5,555 selected single-copy target genes in 24 apoid wasp transcript libraries (with 2,767–4,406, average 4,033, genes per species). However, we discarded 256 of the resulting MSAs due to a missing *N. vitripennis* nucleotide sequence, which resulted in 5,299 target genes. Using the gene models of the *N. vitripennis* official gene set 1.2 as a basis for identifying CDS regions in the 5,299 MSAs suggested 10,854 CDS regions as suitable for bait design. Requiring the presence of at least one representative species per taxonomic subfamily and tribe (17 taxonomic groups in total) in each MSA resulted in 631 CDS regions in 424 single-copy genes as promising for bait design. When comparing the orthologous nucleotide sequences of the species included in the 631 CDS region MSAs, we found the maximum sequence distances to range between 6.7% and 68% when analyzing all possible 120-bp-long nucleotide sequence windows (i.e., the length of baits that we intended to design). Specifying a sequence similarity threshold of 6% for clustering the sequences of each given 120-bp-long sequence window, we inferred 12,177,558 promising baits likely to capture a total of 631 CDS regions. Searching the 12,177,558 bait sequences, referring to 79,174 bait regions against the genomes of the *Ap. mellifera*, *N. vitripennis*, and *H. saltator* (supplementary file S2, Supplementary Material online) indicated competing nontarget binding sites for baits in 23,910 bait regions. We deemed the remaining 55,264 bait regions suitable for capturing 509 CDS regions in 356 genes. Using BaitFilter to choose for each CDS region the bait region that requires the smallest number of baits resulted in 77,119 baits, which are required to enrich the 509 CDS regions under the requested tiling design and the cluster threshold parameter. However, given the maximum number of 57,650 baits that the SureSelect Target Enrichment System by Agilent Technologies, Inc. allowed to be designed on a single glass slide, we removed baits for enriching 131 CDS regions, thereby losing the ability to enrich 74 target genes. At this point, we were able to order 57,650 nonredundant baits to empirically test their capability to capture 378 CDS regions in 282 genes in various in- and outgroup species (supplementary file S3, Supplementary Material online). Due to later optimization of the BaitFisher code for extracting individual CDS regions in the MSAs (see Empirical Evaluation of the Bait Enrichment Capabilities), a small fraction (1.6%) of the ordered baits is not suggested by the current version of the software any more, because some baits do not full-length map to the target gDNA. Our subsequent empirical testing of the ordered baits (see Target DNA Enrichment Success) thus provides conservative estimates.

Computational Performance of BaitFisher and BaitFilter

We evaluated the computational performance of BaitFisher and BaitFilter on a 2.66 GHz Linux desktop computer with 36 GB of RAM using the above design of baits for enriching single-copy genes (see Inference of Baits for Studying Target Genes in Apoid Wasps). For this purpose, BaitFisher was provided the 5,299 MSAs (consisting on average of 19 sequences) specified in Inference of Baits for Studying Target Genes in Apoid Wasps and was run with the parameters outlined in Empirical Evaluation of the Bait Enrichment Capabilities. When applying various distance threshold values (0.06–0.30) for clustering of the nucleotide sequences, the total number of baits required to enrich the target loci significantly decreased when increasing the nucleotide sequence clustering threshold (supplementary file S1, Supplementary Material online). At the same time, the run-time for computing baits only slightly increased when increasing the nucleotide sequence clustering threshold.

To assess the impact of the number of nucleotide sequences within a given MSA on BaitFisher's memory consumption and computation time, we analyzed MSAs with an arbitrary set of 1) 500, 2) 1,500, and 3) 2,500 nonredundant and publicly available nucleotide sequences of the barcoding gene cytochrome c oxidase I (COI) using the same hardware as specified above. The required computation times were 6.2, 28.6h, and 289 h, respectively. The observed increase in run-time is roughly in line with the expected run time for a hierarchical clustering algorithm, which scales with the order of $O(N^2)$, where *N* is the number of nucleotide sequences. We also found the memory consumption to scale roughly with

$O(N^2)$: 18, 180, and 396 MB, respectively. The software implementation limit for the maximum number of sequences in a MSA for BaitFisher to be able to handle is 32,767. The practical limit for the maximum number of sequences in a MSA is determined by the available computation time: Analyzing a MSA with 3,000 nucleotide sequences required BaitFisher about 16 days on the described hardware.

Applying BaitFilter on the output files from the apoid wasp data set with different cluster threshold values (see above and supplementary file S1, Supplementary Material online) showed that even for large output files of up to 1 GB in size, BaitFilter extracts the requested information in less than 2 min (supplementary file S1, Supplementary Material online). Users are thus unlikely to experience any practical limitations when filtering BaitFisher output files.

When BaitFilter was invoked for removing bait sets that contain baits with nontarget binding sites in a reference genome, the filtering took several hours, as BaitFisher relies on the BLAST+ software for searching baits against the reference genome. The time required for this step is thus primarily determined by the number of baits and the size of the reference genome through which it searches (supplementary file S1, Supplementary Material online).

Target DNA Enrichment Success

We collected between 1.38 and 2.25 M (deeply sequenced Illumina DNA sequencing test libraries; four species in total) and between 0.35 and 0.97 M (shallowly sequenced Illumina DNA sequencing libraries; nine species) quality-trimmed raw reads per species. These reads assembled into 4,508–19,100 (deeply sequenced libraries) and 1,884–13,035 (shallowly sequenced libraries) contigs with lengths between 414 and 803 bp (table 1).

When searching the 13 obtained assemblies with Orthograph for the 378 target CDS regions in 282 target genes, we identified 203–303 target CDS regions (average: 263; median: 275) and 26–279 (average: 253; median: 262) target genes (table 1). The fewest target CDS regions and target genes were identified in the *Crabro peltarius* sample, which had been stored in Vitzthum's solution for 21 years. We found no striking difference in the target DNA recovery between samples of ingroup species preserved in pure ethanol (274 and 262 target genes; the first value found for a deeply sequenced sample, the second for a shallowly sequenced sample) and those preserved in approximately 70% ethanol (251–276 target genes; values refer to both deeply and shallowly sequenced samples) (table 1).

The base-coverage depth of contigs that referred to target genes, C_t, was on average 38–94 in species with deeply sequenced libraries and 3–51 in species with shallowly sequenced libraries (table 1). The base-coverage depth of contigs that contained nontarget DNA, C_n, was on average 0.15 of that of contigs with target DNA, suggesting a relative enrichment coefficient of 6.8 (table 1). When comparing C_t with the base-coverage depth that one would expect to find in the assembled contigs if DNA fragments of the genome of the investigated species were randomly sequenced (C_g), we found C_t to be on average 71.1 times higher than C_g (table 1).

Table 1. Sequencing Depth and Target Gene Recovery Statistics Obtained from Tests of BaitFisher-Inferred Baits on Genomic DNA of Various Apoid Wasps and Selected Outgroup Species.

Species	Number of Clean Raw Reads	Number of Assembled Contigs	Number of Assembled and Cleaned Contigs	Number of Sequenced Target Genes	Length of Contigs Referring to Target DNA (bp) (min–median–max)	C_t	C_n	$C_t \times C_n^{-1}$	$C_t \times C_g^{-1}$	C_g	Sample Tissue Preservation History
Dynatus burmeisteri	1,379,306	7,407	7,261	270	328–588–1,717	54	7	7.7		NA	7 years in 70% ethanol
Isodontia mexicana[a]	2,084,280	17,567	17,374	274	323–792–3,202	94	6	15.7		NA	12 years in 70% ethanol
Stangeella cyaniventris	1,723,712	4,508	4,495	251	308–553–2,674	38	18	2.1		NA	7 years in 70% ethanol
Stictia heros	2,248,824	19,101	19,091	279	180–584–1,719	73	5	14.6		NA	<1 year in 96% ethanol
Ampulex compressa	596,982	7,926	7,731	253	329–821–2,659	31	4	7.8	64.8	0.08	<1 year in 96% ethanol
Apis mellifera	488,786	2,190	2,120	188	319–441–1,070	51	6	8.5	98.1	0.22	<1 year in 96% ethanol
Crabro peltarius[a]	707,220	7,264	7,020	26	334–472–1,092	3	2	1.5		NA	21 years in Vitzthum's solution
Clypeadon sculleni	833,770	13,035	12,876	271	333–803–1,967	29	19	1.5		NA	2 years in 96% ethanol
Dinetus pictus[a]	350,062	1,935	1,702	262	319–445–101	28	5	5.6		NA	9 years in 70% ethanol
Eremnophila melanaria	381,072	1,884	1,658	276	327–614–1,686	23	5	4.6		NA	7 years in 70% ethanol
Harpegnathos saltator	548,090	5,154	4,904	236	321–698–2,313	24	4	6.0	57.8	0.07	<1 year in 96% ethanol
Nasonia vitripennis	964,536	9,933	9,742	203	333–581–1,971	41	6	6.8	63.6	0.13	<1 year in 96% ethanol
Sphex funerarius[a]	394,392	3,495	3,203	278	383–757–2,456	28	5	5.6		NA	10 years in 70% ethanol

NOTE.—The libraries of the first four species were deeply sequenced, those of the following nine species were shallowly sequenced. C_t, average base-coverage depth of contigs referring to target genes; C_n, average base-coverage depth of contigs referring to nontarget genes; C_g, average base-coverage depth of contigs expected if DNA fragments of the genome of the investigated species were randomly sequenced (only given for species whose genome size was known); NA, not applicable.
[a]Target DNA of the species was known and exploited during bait design.

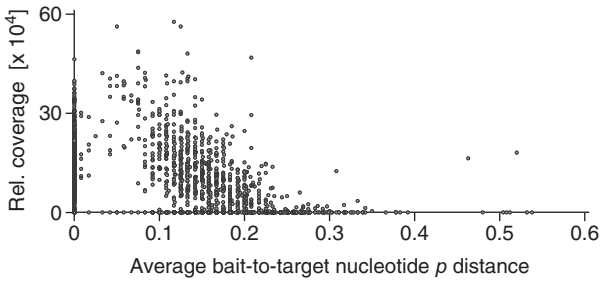

FIG. 2. Correlation between average Hamming (p) distances of baits designed by BaitFisher for enriching a given locus to the respective locus' actual nucleotide sequence and the relative base-coverage depth (normalized by dividing the base-coverage depth by the total amount of sequenced nucleotides) by which the locus was sequenced after applying the designed baits for enriching the target DNA. Shown are the results from analyzing *Apis mellifera*, *Dinetus pictus*, *Harpegnathos saltator*, *Isodontia mexicana*, *Nasonia vitripennis*, and *Sphex funerarius*.

To assess the impact of the bait-to-target sequence similarity on the enrichment efficiency, we plotted the base-coverage depth of contigs referring to target genes, C_v, normalized by dividing it by the total number of sequenced nucleotides, against the average bait-to-target sequence similarity in species with known target DNA (fig. 2). We found a strong negative correlation between bait-to-target sequence similarity and the relative base-coverage depth of the target DNA (fig. 2).

Discussion

The ability to selectively study the nucleotide sequences of hundreds or thousands of loci of interest in the genomes of different species can be considered as one of the most significant steps forward in targeted genomic data acquisition, relevant to many research disciplines (Bejerano et al. 2004; Hodges et al. 2007; Ng et al. 2009; Crawford et al. 2012; Brandley et al. 2015; Jones and Good 2016). Although the inference of oligonucleotides that serve as baits for enriching target DNA in a single species, whose genome has been sequenced, is well established, the design of baits to enrich target DNA in a wider range of species still remains a challenge. Researchers have applied different strategies to capture target loci across species. Li et al. (2013), for example, designed baits to capture target genes by using baits designed from analyzing the genome of a single species. By tuning the wet laboratory procedures (e.g., hybridization temperature profile; see below), the authors were able to extend the reach of the method considerably despite the potentially substantial bait-to-target distances associated with the applied bait design strategy. This approach is reasonable if no additional nucleotide sequence information of taxonomic ingroup species is available. Other authors considered nucleotide sequence information from in- and outgroup species in search for (ultra) conserved nucleotide sequence sections that can serve as anchors to capture and study (typically more variable) flanking nucleotide sequences across species (Crawford 2012; Faircloth et al. 2012; Lemmon et al. 2012; McCormack, Harvey, et al. 2013; McCormack, Hird, et al. 2013). Given the high conservation of the target nucleotide sequence, this approach also allows using the nucleotide sequence of a single species for bait design. Unfortunately, the approach cannot easily be used to study loci that are spatially distant from conserved sequence sections.

To capture and study variable exonic sequences in an entire class of marine invertebrates (Ophiuroidea), Hugall et al. (2015) recently suggested and applied an intriguing approach by exploiting transcriptomes. They inferred a phylogenetic tree from transcriptomes, which had been sampled in species across the class Ophiuroidea. The tree was then used to infer the ancestral nucleotide sequences of single-copy target genes in subordinated clades within Ophiuroidea. Sections of the inferred ancestral nucleotide sequences subsequently served as baits to capture the corresponding loci in other species of these clades. The number and size of the clades, from which one ancestral nucleotide sequence per locus was inferred, was chosen in a manner that the majority (>80%) of the resulting baits for capturing the target loci in species of a given clade did not differ in more than 12% from the known transcript sequences of species in this clade. The clustering of species and the inference of ancestral nucleotide sequences served two purposes: 1) To reduce redundancy in the taxonomic sampling by clustering species that share a high nucleotide sequence similarity and 2) to traceably select a single representative nucleotide sequence per locus and clade from which baits are designed.

Our approach to optimize the number and the efficacy of baits required to capture target loci across species relies on a strategy comparable with the one proposed by Hugall et al. (2015): It exploits user-provided nucleotide sequence information of target loci in different species for designing baits and automatically reduces (taxonomic) redundancy by clustering nucleotide sequences that differ in less than a user-defined threshold value from each other. In contrast to the approach applied by Hugall et al. (2015), our approach performs the clustering of nucleotide sequences, from which baits are inferred, for each nucleotide sequence window of bait-length size separately. By not clustering the reference species' nucleotide sequences by the species' phylogenetic relationships, but by clustering them according to the sequences' distances separately in each sequence section of bait-length size, we are able to reduce redundancy and bait-to-target distances even further (as compared with the approach applied by Hugall et al. 2015). We subsequently infer one artificial bait sequence per sequence window of bait-length size and group of clustered nucleotide sequences to reduce the bait-to-target distance across species, while Hugall et al. (2015) inferred an ancestral sequence for this purpose. The 1-center sequence guaranties that the bait-to-target sequence distance (as judged from the baits' sequence distance to the corresponding clustered nucleotide sequences; see supplemental file S1, Supplementary Material online) is indeed minimized. An ancestral sequence, a randomly picked ingroup sequence, or a consensus sequence, in contrast, do not guarantee to minimize the maximum bait-to-target sequence distance. For example, the nucleotide sequence of a

locus can be highly derived in some of the sampled species in a clade. Using the presumed ancestral sequence of this locus as bait would thus possibly result in the bait's nucleotide sequence being more similar to that of species with more plesiomorphic sequences than to those of species with more derived sequences.

The availability of the nucleotide sequences of a representative set of ingroup species is an important prerequisite when designing baits that are meant to effectively enrich target DNA across species of the ingroup. Our empirical evaluation of 120-bp-long baits to enrich target DNA with known nucleotide sequence similarity using the molecular procedure outlined in Taxon Sampling and Molecular Procedures and in which we applied constant hybridization and posthybridization washing temperatures suggests that the bait-to-target DNA sequence distance should not exceed 15–20% for the enrichment to be efficient (fig. 2). Our results are in line with those reported by Bragg et al. (2015), Hawkins et al. (2015), and Paijmans et al. (2016). Li et al. (2013) reported the capture of nucleotide sequences exhibiting a bait-to-target distance of up to 39%. The libraries that we enriched contained target loci that differed in up to 52% from the corresponding baits, but there is a clear negative correlation between enrichment efficacy and bait-to-target nucleotide sequence dissimilarity (fig. 2). Thus, although it may appear that our experiments resulted in successful enrichment of target loci differing in up to 52% from the nucleotide sequence of the applied capture baits, we interpret these distant target nucleotide sequences as outliers (fig. 2). This is because any enriched library still contains nontarget nucleotide sequences with both low and high read coverage. Based on the currently available data and the applied wet laboratory protocol (Taxon Sampling and Molecular Procedures; see also discussion further below), we suggest using a cluster threshold of not more than 30% when designing baits with a length of 120 bp. Although this value may appear at first glance conservative, given that the nucleotide sequence of a bait would generally not differ in more than 15% from any nucleotide sequence in a given cluster, this value acknowledges that the sequences of some target species may have historically undergone accelerated evolutionary change and that a higher enrichment efficacy requires less deep sequencing of the enriched library. However, future experiments should investigate the relationship between bait-to-target DNA sequence similarity and enrichment efficacy as a function of the length of the baits. We decided to design baits with a length of 120 bp due to promising results in studies that employed baits of this length for in-solution target capture (Faircloth et al. 2012; Lemmon et. al. 2012); however, other investigators successfully applied baits with a length of 60–90 bp on capture microarrays (Hodges et al. 2009; Mamanova et al. 2010; Hancock-Hanser et al. 2013).

Depending on the research question, target locus specificity of baits can be important. BaitFilter allows evaluating the probability of baits to bind to nontarget DNA by searching all inferred baits against a user-provided genome assembly. Search of bait sequences against an ingroup genome assembly may also prove valuable for evaluating the enrichment success of target loci. BaitFilter therefore also allows the user to assess whether at least one bait, of a given stack of baits (see BaitFisher manual for details) that is meant to bind at a specific position of a target locus across species, indeed exhibits a high nucleotide sequence similarity to a unique locus in a user-provided reference genome assembly. This feature is useful when baits are designed for enriching specific genomic features, such as individual CDSs. If the identification of these genomic features in the nucleotide sequences of the ingroup species relied on gene models in an outgroup species, chances are higher that these features are not applicable to ingroup species. We use sequence nucleotide similarity as a proxy to assess the propensity of baits to bind to target and off-target nucleotide sequence stretches, but acknowledge that the hybridization of oligonucleotides to DNA is determined by thermodynamic properties, such as the number of hydrogen bonds. Consideration of these properties when searching tens or hundreds of thousands of baits against a reference genome of up to several giga base pairs in size is computationally challenging and would result in a reduction of BaitFilter's computational performance. Given the tight correlation between DNA hybridization energy and nucleotide sequence similarity (Wallace et al. 1979) and the fact that baits designed by BaitFisher are meant to enrich loci in species, whose nucleotide sequence is expected to be different from that of the reference species, consideration of thermodynamic properties is expected to result only in a marginal improvement of the predictive power. We therefore deliberately refrained from considering hybridization properties in the current version of the BaitFilter software. However, a promising approach to cope with this shortcoming could be combining nucleotide sequence similarity search-guided identification of reference genome candidate regions, to which baits could potentially bind, and an in-depth analysis of the thermal stability of bait-target DNA duplexes in these candidate regions. BaitFisher currently does not consider the baits' propensities for folding and dimerization either. Although DNA binding and folding energy calculations are often considered by polymerase chain reaction (PCR) oligonucleotide primer design software (Mann et al. 2009), the large size (in bp) and the disproportionately large number of disparate oligonucleotides typically employed in target DNA enrichment exacerbate explicit contemplation of these properties in the latter context.

Our empirical evaluation of baits inferred with the aid of BaitFisher and BaitFilter on DNA of apoid wasps showed that the baits worked very well. In fact, the overall enrichment coefficient (C_t/C_g) achieved by using the baits proved to be in the magnitude of 58- to 98-fold when comparing the base-coverage depths of target loci (C_t) with the base-coverage depths expected if no enrichment had taken place (C_g) (table 1). The recovery rate of the target DNA from samples that had been long-term stored in approximately 70% ethanol was also very high (table 1) and opens a wide range of new areas for the application of target DNA enrichment. We see, for example, a particular profit of target DNA enrichment in the field of museomics and biodiversity monitoring. In the former, investigators seek to recover the DNA from unique

and often old samples stored in museum collections (Guschanski et al. 2013). The classical procedure of PCR-amplifying target loci and subsequent sequencing of the obtained amplicons using Sanger sequencing technology often cannot be applied, because the target DNA is too degraded (Hofreiter et al. 2015; Paijmans et al. 2016). Our recovery rate of target DNA from samples that had been stored for up to 12 years in approximately 70% ethanol, which resulted in a substantial degradation of the samples' DNA, was very high and is extremely promising (table 1). These results have been obtained by applying the molecular procedures outlined in Taxon Sampling and Molecular Procedures. The procedures involved constant hybridization and posthybridization temperature profiles and only a single round of target locus capture. Li et al. (2013) and Paijmans et al. (2016) assessed alternative temperature profiles in the hybridization step and in posthybridization steps and suggest modifications of the wet laboratory protocols that allow extending the reach of the method. Li et al. (2013) also suggested conducting a second round of target locus capture to further increase the enrichment success. We refer the reader to these two excellent articles when planning their wet laboratory procedures.

Materials and Methods

Taxon Sampling and Molecular Procedures

DNA Extraction and Library Preparation for Next-Generation DNA Sequencing

We tested the enrichment capacity of the 57,650 inferred baits on DNA extracts of nine ingroup species (i.e., apoid wasps, excluding cockroach wasp) and four outgroup taxa (including cockroach wasp; supplementary file S4, Supplementary Material online). Specifically, we selected four species of crabronid wasps and five species of sphecid wasps. The cDNA sequences of four of these species (i.e., *C. peltarius*, *Dinetus pictus*, *Isodontia mexicana*, *Sphex funerarius*) had also been used for bait design. Hence, these four species served as a positive control (i.e., the nucleotide sequences of a specific fraction of the designed baits were known to differ in less than 6% from that of the target genomic DNA of the four species). As outgroup taxa, we chose the cockroach wasp *Ampulex compressa* (Ampulicidae), the honeybee (*Ap. mellifera*), an ant (*H. saltator*), and a parasitoid wasp (*N. vitripennis*). The genomes of the latter three are sequenced (Weinstock et al. 2006; Bonasio et al. 2010; Werren et al. 2010) and enabled us to estimate the degree of target DNA enrichment in more distantly related taxa (as compared with ingroup species). The enriched DNA of these three taxa plus that of the four ingroup species serving as controls were also considered when exploring the correlation between bait-to-target DNA distance and target locus base-coverage depth. The DNA quality differed across the analyzed samples: Although we extracted some DNA from tissues that were short-term stored in absolute ethanol (i.e., the four outgroup species plus *Clypeadon sculleni* and *Stictia heros*), other DNAs were extracted from tissues that had been stored over much longer time (9–21 years) in either approximately 70% ethanol (i.e., *Di. pictus*, *Dynatus burmeisteri*, *Eremnophila melanaria*, *I. mexicana*, *Sph. funerarius*, *Stangeella cyaniventris*) or in Vitzthum's solution (80 g of 75% ethanol, 16 g glycerol, 4 g acetic acid glacial; Öttingen 1938) (i.e., *C. peltarius*).

The genomic DNA of all investigated species was extracted either with the Qiagen DNeasy Blood & Tissue Kit (Qiagen GmbH, Hilden, Germany) or by applying the CTAB DNA extraction protocol by Rogers and Bendich (1985) in combination with a DNA purification step using AMPure XP beads (Beckman Coulter GmbH, Krefeld, Germany). All extracted DNAs were dissolved in 100 µl of nuclease-free water and next-generation DNA sequencing libraries were prepared from the extracted DNA following the protocol given in supplementary file S5, Supplementary Material online.

Target DNA Enrichment and Illumina MiSeq Paired-End DNA Sequencing

We followed the SureSelect Target Enrichment System Kit protocol by Agilent Technologies, Inc. for Illumina Multiplexed Sequencing, published in 2013 (pp. 60–70), to capture target DNA fragments from the amplified NGS libraries using a pool of 57,650 baits that was designed by BaitFisher and synthesized by Agilent Technologies, Inc. when ordering the SureSelect Target Enrichment System Kit. Hybridization of the baits to the target DNA was allowed for 18 h at 65 °C in a GeneAmp PCR System 2700 thermocycler (Applied Biosystems, Inc., Waltham, USA). Posthybridization PCR amplification of the target-enriched libraries was also conducted with a GeneAmp PCR System 2700 thermocycler using the PCR Primer Cocktail and PCR Mastermix as described in supplementary file S5, Supplementary Material online, for NGS library PCR amplification. No additional indexing was done because we had already ligated indices to the DNA fragments of the NGS libraries during library preparation (supplementary file S5, Supplementary Material online). We applied the PCR protocol (consisting of 12 cycles) recommended by Agilent Technologies, Inc. for capturing >1.5 Mbp of DNA in all posthybridization PCRs. All amplified enriched libraries were purified with Agencourt AMPure XP beads, and the quality and quantity of the purified DNA fragments assessed with a Fragment Analyzer (Advanced Analytical Technologies GmbH, Heidelberg, Germany) and a Qubit 2.0 Fluorometer (Thermo Fisher Scientific Inc., Waltham, USA). In those cases in which the total yield of DNA in the range 200–800 bp proved to be too low (i.e., DNA concentration <1.5 ng/µl) for Illumina paired-end sequencing, we repeated the posthybridization PCR amplification as described above. Illumina MiSeq paired-end DNA sequencing of the enriched next-generation DNA sequencing libraries followed the protocol given in supplementary file S5, Supplementary Material online. In the first four samples sequenced (i.e., *Dy. burmeisteri*, *I. mexicana*, *Sta. cyaniventris*, *Sti. heros*), we collected 1.4–2.3 Mbp per species. Given the high base-pair coverage depth in the target genes achieved from assembling these data, we subsequently

lowered the amount of raw data collected per species to 0.35–0.96 Mbp.

De Novo Assembly of Reads

The quality of all obtained NGS raw reads was checked with FastQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptor and poor-quality regions were clipped with Trimmomatic 0.32 (Bolger et al. 2014; seed mismatches: 2, palindrome clip threshold: 30, simple clip threshold: 10, minimum quality required to keep a leading base: 3, minimum quality required to keep a trailing base: 3, sliding window size: 4, required average quality in window: 15, minimum length of reads to be kept: 25). The filtered paired-end reads were then assembled with the IDBA-UD de novo assembler 1.1.1 (Peng et al. 2012). The assembler is optimized for assembling contigs sequenced to a very uneven base-coverage depth. We recompiled IDBA-UD after applying slight changes in the source code, as suggested by the software developers, so that the assembler was able to handle reads of up to 320 bp in length. The iterative assembly process started with a *k*-mer size of 20 bp. The *k*-mer size was increased in steps of 5 bp during each iteration, until a *k*-mer size of 120 bp was reached.

Identification and Removal of Possible Contaminant Contigs

We discovered in context of the 1KITE project that single index-tagged libraries pooled on the same Illumina lane often exhibit a small percentage of cross contamination. To cope with this problem in the present investigation, we searched the contigs of those reduced-representation libraries that were sequenced on the same Illumina lane against each other with the BLASTN search engine of BLAST+ 2.2.29 (Camacho et al. 2008). In those instances, in which we identified nucleotide sequences that shared over a length of $\geq$200 bp a similarity of $\geq$98% with each other, we proceeded as follows: 1) If the relative read-coverage depths of the two contigs in question differed more than 2-fold, we removed the contig with the lower relative read-coverage depth from the corresponding assembly; and 2) if the relative read-coverage depth of the two contigs in question were sequenced to roughly the same depth (less than 2-fold difference), we conservatively removed both of the contigs from the corresponding assembly. If multiple highly similar contigs were found (because we searched the contigs of all assemblies in question simultaneously against each other; see above), we retained only the contig with the best coverage, given that its coverage was more than 2-fold higher than the coverage of the second-best matching contig. We defined as "read-coverage depth" of a given contig, the number of reads (as provided in IDBA-UD output) of this contig divided by the total amount of nucleotides sequenced from the corresponding library.

Target DNA Recovery and Enrichment Efficiency

To assess the coverage of the enriched target regions, we used the software segemehl 0.1.7 (Hoffmann et al. 2009) for mapping all raw sequencing reads to the assembled and contamination-filtered contigs. The mapping results were exported in the SAM file format, which was then imported for further analysis in tablet 1.14.10.20 (Milne et al. 2013). Tablet allowed us to conveniently calculate the number of reads that mapped to each specific contig. Exploiting this information, we calculated the actual average base-coverage depth of those contigs that contain a 250-bp-long bait-binding sequence section using the formula $C_t = N_t \times L_t \times S_t^{-1}$, in which N_t is the number of reads that mapped to a given contig containing the target DNA, L_t is the length (250 bp) of the reads that mapped to the contig containing the target DNA, and S_t is the length (in bp) of the contig containing the target DNA. We analogously calculated the average base-coverage depth of contigs that do not contain target DNA (C_n). Finally, we compared C_t with C_n, and calculated the ratio $C_t \times C_n^{-1}$ as one measure of target DNA enrichment degree.

To further assess the extent to which target DNA was enriched, we calculated the average base-coverage depth, C_g, that one would expect the sequenced and assembled fragments of the genome of a given species to exhibit if no enrichment had taken place. C_g was calculated using the formula $C_g = N_g \times L_g \times S_g^{-1}$, in which N_g is the total number of sequenced reads, L_g is the length (250 bp) of all sequenced reads, and S_g is the genome size (haploid nuclear DNA content in bp; Lander and Waterman 1988) of the investigated species. C_g was consequently only calculated in species whose genome size is reliably known. This is the case for *Am. compressa* (374 Mbp; Niehuis O, unpublished data), *Ap. mellifera* (235 Mbp; Ardila-Garcia et al. 2010), *N. vitripennis* (312 Mbp; Beukeboom et al. 2007), and *H. saltator* (330 Mbp; Bonasio et al. 2010). Finally, we compared C_t with C_g, and calculated the ratio $C_t \times C_g^{-1}$ as second measure of target DNA enrichment degree.

To shed light on the relationship between bait-to-target nucleotide sequence similarity and the relative target DNA base-coverage depth, we calculated the lowest observed distance between baits of a given bait set and the target DNA per CDS region. This has been calculated for *Di. pictus*, *I. mexicana*, and *Sph. funerarius* (in-group species whose target DNA sequence was known to us from the transcript library DNA sequences) and for *Ap. mellifera*, *H. saltator*, and *N. vitripennis* (outgroup species whose genome is sequenced) using a custom C++ program. We did not consider values referring to *C. peltarius* in this analysis due to the low overall recovery of target genes from sequencing the library of this species. Furthermore, we only considered bait-to-target DNA distance values that are based on MSAs, in which the entire length of the target DNA was known for each bait. The relative base coverage of target loci was obtained by dividing the base coverage of each target locus by the total number of nucleotides sequenced per library.

Supplementary Material

Supplementary files S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Leo W. Beukeboom, Jürgen Gadau, Jürgen Liebig, Robin Moritz, Dieter Schulten, and Silke Stoll for providing samples of *Ampulex compressa*, *Apis mellifera*, *Harpegnathos saltator*, and *Nasonia vitripennis*. M.O. and O.N. further acknowledge the German Research Foundation (DFG) for supporting this research (OH81/9-1; NI 1387/1-1). C.B. is a "Ramon y Cajal" fellow supported by the Spanish Ministry of Science and Education (MEC) (RYC-2014-15615). The authors are furthermore grateful to the 1KITE initiative, which produced the data that made this research possible, as well as to Elise Laetz and two anonymous reviewers for help in further improving the manuscript. All analyzed species were collected before October 2014. The study is part of the doctoral thesis of M.S. M.M. holds a company providing DNA sequencing services. K.L. is an employee of this company.

References

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*. 4:903–905.

Ardila-Garcia AM, Umphrey GJ, Gregory TR. 2010. An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. *Insect Mol Biol*. 19:337–346.

Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. 2005. Direct genomic selection. *Nat Methods*. 2:63–69.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Beukeboom LW, Kamping A, Louter M, Pijnacker LP, Katju V, Ferree PM, Werren JH. 2007. Haploid females in the parasitic wasp *Nasonia vitripennis*. *Science* 315:206.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403.

Blumenstiel B, Cibulskis K, Fisher S, DeFelice M, Barry A, Fennell T, Abreu J, Minie B, Costello M, Young G, et al. 2010. Targeted exon sequencing by in-solution hybrid selection. *Curr Protoc Hum Genet*. Chapter 18, Unit 18.4.

Bodi K, Perera AG, Adams PS, Bintzler D, Dewar K, Grove DS, Kieleczawa J, Lyons RH, Neubert TA, Noll AC, et al. 2013. Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech*. 24:73–86.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* 329:1068–1071.

Bragg JG, Potter S, Bi K, Moritz C. 2015. Exon capture phylogenomics: efficacy across scales of divergence. *Mol Ecol Res*. Advance Access published August 20, 2015, doi:10.1111/1755-0998.12449

Brandley MC, Bragg JG, Singhal S, Chapple DG, Jennings CK, Lemmon AR, Lemmon EM, Thompson MB, et al. 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian *Eugongylus* group scincid lizards. *BMC Evol Biol*. 15:62.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2008. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* 12:347.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Lett*. 8:783–786.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for ortholog ESTs. *BMC Evol Biol*. 9:157.

Faircloth BC, Branstetter MG, White ND, Brady SG. 2014. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Mol Ecol Res*. 15:489–501.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*. 61:717–726.

Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, Young G, Fennell TJ, Allen A, Ambrogio L, et al. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol*. 12:R1.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, et al. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Syst Biol*. 62:539–554.

Hancock-Hanser BL, Frey A, Leslie MS, Dutton PH, Archer FI, Morin PA. 2013. Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Mol Ecol Res*. 13:254–268.

Hawkins MT, Hofman CA, Callicrate T, McDonough MM, Tsuchiya MT, Gutiérrez EE, Helgen KM, Maldonado JE. 2015. In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol Ecol Res*. Advance Access published August 24, 2015, doi:10.1111/1755-0998.12448

Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc*. 4:960–974.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, et al. 2007. Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet*. 39:1522–1527.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*. 5:e1000502.

Hofreiter M, Paijmans JL, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ. 2015. The future of ancient DNA: technical advances and conceptual shifts. *BioEssays* 37:284–293.

Hugall AF, O'Hara TD, Hunjan S, Nilsen R, Moussalli A. 2015. An exon-capture system for the entire class Ophiuroidea. *Mol Bio Evol*. 33:281–294.

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. *Mol Ecol*. 25:185–202.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–780.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231–239.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol*. 61:727–744.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ. 2013. Capturing protein-coding genes across highly divergent species. *Biotechniques* 54:321–326.

Li M, Ma B, Wang L. 2002. On the closest string and substring problems. *J ACM*. 49:157–171.

Liu S, Wang X, Xie L, Tan M, Li Z, Su X, Zhang H, Misof B, Kjer KM, Tang M, et al. 2016. Mitochondrial capture enriches mito-DNA 100 folds enabling PCR-free mitogenomics biodiversity analysis. *Mol Ecol Res*. 16:470–479.

Mann T, Humbert R, Dorschner M, Stamatoyannopoulos J, Noble WS. 2009. A thermodynamic approach to PCR primer design. *Nucleic Acids Res*. 37:e95.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next generation sequencing. *Nat Methods*. 7:111–118.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One* 8:e54848.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol*. 66:526–538.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. 2010:pdb.prot5448.

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 14:193–202.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.

Muñoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elsik CG. 2011. Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res*. 39:D658–D662.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 48:443–453.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276.

Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M, et al. 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res*. 21:1339–1348.

Öttingen Hv. 1938. Erfahrungen über das Arbeiten mit Thysanopteren. *Arb Phys Augew Ent Berlin-Dahlem*. 5:178–182.

Paijmans JL, Fickel J, Courtiol A, Hofreiter M, Förster DW. 2016. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol Ecol Res*. 16:42–55.

Peñalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RC, Moritz C. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol Ecol Res*. 14:1000–1010.

Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428.

Petersen M, Meusemann K, Donath A, Dowling D, Liu S, Peters RS, Podsiadlowski L, Vasilikopoulos A, Zhou X, Misof B, Niehuis O. 2015. Orthograph 0.5.6. Available from: http://mptrsen.github.io/Orthograph.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569–573.

Rogers SO, Bendich AJ. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol Biol*. 5:69–76.

Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al. 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc Natl Acad Sci U S A*. 108:5673–5678.

Suchan T, Pitteloud C, Gerasimova NS, Kostikova A, Schmid S, Arrigo N, Pajkovic M, Ronikier M, Alvarez N. 2016. Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *PLoS One* 11:e0151651.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34:W609–W612.

Vinner L, Mourier T, Friis-Nielsen J, Gniadecki R, Dybkaer K, Rosenberg J, Langhoff JL, Cruz DF, Fonager J, Izarzugaza JM, et al. 2015. Investigation of human cancers for retrovirus by low-stringency target enrichment and high-throughput sequencing. *Sci Rep*. 5:13201.

Wallace RB, Shaffer J, Murphy RF, Bonner J, Hirose T, Itakura K. 1979. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res*. 6:3543–3557.

Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. 2010. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res*. 39:D283–D288.

Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Robinson GE, Worley KC, Evans JD, Maleszka R, Robertson HM, Weaver DB, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJ, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327:343–348.

B

Other co-authored publications

PLOS ONE

RESEARCH ARTICLE

Towards a DNA Barcode Reference Database for Spiders and Harvestmen of Germany

Jonas J. Astrin[1]*, Hubert Höfer[2]*, Jörg Spelda[3]*, Joachim Holstein[4]*, Steffen Bayer[2], Lars Hendrich[3], Bernhard A. Huber[1], Karl-Hinrich Kielhorn[5], Hans-Joachim Krammer[1], Martin Lemke[6], Juan Carlos Monje[4], Jérôme Morinière[3], Björn Rulik[1], Malte Petersen[1], Hannah Janssen[1], Christoph Muster[7]

1 ZFMK: Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany, **2** SMNK: Staatliches Museum für Naturkunde Karlsruhe, Karlsruhe, Germany, **3** ZSM: Zoologische Staatssammlung München, München, Germany, **4** SMNS: Staatliches Museum für Naturkunde Stuttgart, Stuttgart, Germany, **5** Karl-Hinrich Kielhorn, Berlin, Germany, **6** Martin Lemke, Lübeck, Germany, **7** Zoologisches Institut und Museum, Universität Greifswald, Greifswald, Germany

* j.astrin.zfmk@uni-bonn.de (JJA); hubert.hoefer@smnk.de (HH); spelda@zsm.mwn.de (JS); joachim.holstein@smns-bw.de (JH)

Citation: Astrin JJ, Höfer H, Spelda J, Holstein J, Bayer S, Hendrich L, et al. (2016) Towards a DNA Barcode Reference Database for Spiders and Harvestmen of Germany. PLoS ONE 11(9): e0162624. doi:10.1371/journal.pone.0162624

Editor: Matja Kuntner, Scientific Research Centre of the Slovenian Academy of Sciences and Art, SLOVENIA

Received: June 3, 2016

Accepted: August 25, 2016

Published: September 28, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Sequences are appended as alignment, their trace files are available from BOLD (http://www.boldsystems.org/). The DiStats script is available through GitHub (https://github.com/mptrsen/distats) and through the ZFMK homepage (www.zfmk.de/en/research/research-centres-and-groups/distats).

Funding: The GBOL project is financed by the German Federal Ministry of Education and Research (BMBF#01LI1101). Part of the sequences contributed by ZSM resulted from funding by the Bavarian State

Abstract

As part of the German Barcode of Life campaign, over 3500 arachnid specimens have been collected and analyzed: ca. 3300 Araneae and 200 Opiliones, belonging to almost 600 species (median: 4 individuals/species). This covers about 60% of the spider fauna and more than 70% of the harvestmen fauna recorded for Germany. The overwhelming majority of species could be readily identified through DNA barcoding: median distances between closest species lay around 9% in spiders and 13% in harvestmen, while in 95% of the cases, intraspecific distances were below 2.5% and 8% respectively, with intraspecific medians at 0.3% and 0.2%. However, almost 20 spider species, most notably in the family Lycosidae, could not be separated through DNA barcoding (although many of them present discrete morphological differences). Conspicuously high interspecific distances were found in even more cases, hinting at cryptic species in some instances. A new program is presented: DiStats calculates the statistics needed to meet DNA barcode release criteria. Furthermore, new generic COI primers useful for a wide range of taxa (also other than arachnids) are introduced.

Introduction

Long-term monitoring of biodiversity is one of the most important challenges in conservation biology. To evaluate the conservation status and anthropogenic impact of habitats, sufficient knowledge on species composition of natural environments is needed on a regional level. For many if not most invertebrate taxa, we are still far from achieving this goal. One promising approach to meet this challenge is DNA barcoding [1], a technique that uses the easy to homologize, well-quantifiable, discrete taxonomic characters contained in DNA sequence data for standardized, rapid, and relatively cheap species identification. DNA barcoding depends on

Ministry of Education and Culture, Science and the Arts (Barcoding Fauna Bavarica, BFB). The sequencing work of ZSM was supported, in part, by funding from the Government of Canada to Genome Canada through the Ontario Genomics Institute, while the Ontario Ministry of Research and Innovation and NSERC supported development of the BOLD informatics platform. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

low levels of intraspecific variation coupled with marked genetic differentiation between species (the 'barcoding gap', investigated in spiders in [2–4]).

With more than 45,800 described species [5], spiders are among the most diverse animal orders [6]. They are abundant in all terrestrial habitats. As ubiquitous predators, they occupy a key position in food webs. Many species show preferences for specific habitat structures or environmental factors, e.g. temperature, humidity, shading [7], which turns them into potential indicators [8]. Easy to observe and document, spiders are seen as a model group for ecological studies [9,10].

The spider fauna of Germany, comprising approximately 1000 species [11], is well known, and checklists and red lists of endangered species have been published for Germany and most of its federal states (see [11] and references therein). The 'Arachnologische Gesellschaft e.V.' (www.arages.de) offers regularly updated occurrence maps, based on a steadily growing database. Therefore, spiders are regularly used in habitat assessments, biodiversity inventories, and ecological studies (e.g. [12–18]). Spiders are particularly promising as indicators of sustainable forest management [19], habitat structure [20], successional stages [21,22], or conservation value [23,24]. There have been several attempts to classify spiders according to their habitat or niche preferences in Germany or Central Europe [7,25–30] and to use these data to classify habitats or assess habitat quality by identifying the proportion of rare, endangered, stenotopic, or character species (e.g. [18,31,32]). Identification of German spiders is facilitated by the online keys for spiders of Europe at www.araneae.unibe.ch [33]. However, morphological identification to species level requires adult specimens in most instances. About 80–200 spider species can occur in a near-natural habitat in Germany, of which only a small fraction can be directly recorded and identified in the field (pers. obs., H. Höfer, C. Muster). For an ecologically meaningful assessment or a close to complete inventory, much more time needs to be invested to capture, process (often meaning dissection of sexual organs) and identify the (adult) spiders, requiring considerable expertise. Regularly, several specimens remain that have to be checked by the few available taxonomic specialists with sufficient knowledge on morphological variability in the respective species and with access to reference collections.

With some 6500 species worldwide, harvestmen (Opiliones) constitute the third-largest order of arachnids [34]. Currently, 52 species have been recorded from Germany [35]. The omnivorous harvestmen constitute a regular component of terrestrial faunas, with highest densities in damp and shaded habitats [36]. Their use in applied and ecological studies is explained by the existence of both stenotopic species with strict microhabitat requirements (and often limited geographic ranges) and invasive species that exhibit immense colonization potential [37,38]. Determination of most German taxa is reliably achievable using the work of [39]. However, recent studies have revealed high levels of cryptic diversity in Central Europe [40–42], suggesting a promising perspective for DNA barcoding in this taxon.

The use of mitochondrial COI barcodes [43] from an extensive reference database of spider and harvestmen species will aid non-specialists in the determination of these groups. Species that have hitherto been problematic or even impossible to identify morphologically–either in general or for a particular sex–may be reliably discriminated. Even though not frequent in Germany, there are still many spider species in which one of the (dimorphic) sexes is still unknown, and barcoding can provide the link between sexes (demonstrated e.g. in [44]). Moreover, disputed instances of synonymy may be resolved [45]. Not least, a considerable advantage of barcoding is the possibility to identify juvenile specimens [3,46–49]. This will not only make inventories more complete, but will also allow species-level inclusion of juveniles into ecological analyses. Tapping into this rich material resource will allow studying more ecological questions without the necessity for exhaustive and expensive sampling. A future broad application of routine DNA barcoding in spiders is facilitated through mass-trapping, since some of the

sampling solutions employed in traps preserve DNA well enough for barcoding [50,51]. The method further holds the potential to reveal cryptic species or to identify cases where morphological plasticity may have been over-interpreted. Barcoding may thus act as a catalyst for alpha taxonomy [52]. While introgression events [53], retention of ancestral polymorphisms [49], nuclear mitochondrial pseudo-genes [54] or endosymbionts (*Wolbachia* bacteria etc.) [55,56] all pose potential problems to DNA barcoding approaches, a growing number of studies show the general feasibility of DNA barcoding for arachnids [2–4,46,47,57–66].

The German Barcode of Life (GBOL) campaign is implemented by a national network of ca. 20 biodiversity research institutions and more than 200 taxon specialists [67]. It pursues the goal to establish a DNA barcode library of as many animal, fungal and plant species as possible that occur in Germany. The project aims at collecting, if possible, ten specimens per species, from locations as distinct as possible throughout the country in order to capture genetic variability. Some species with wider ranges may also include specimens collected in neighboring countries.

Natural history collections constitute the core infrastructure of GBOL, taking into account that barcoding projects produce a valuable legacy of vouchers (morphological specimens and molecular samples alike) which become relevant in subsequent studies due to the high quality of the underlying taxonomic assignments and granularity of the metadata. These vouchers form the physical foundation that future monitoring projects will be based on, warranting continuous testability, validation, and coherent expansion of the barcoding reference database–ideally for centuries to come.

Within the GBOL consortium, arachnids have received wide attention, as no less than four GBOL institutes and their respective external arachnologist partners collaborate intensively on compiling a national molecular inventory of spiders and harvestmen: Staatliches Museum für Naturkunde Karlsruhe (SMNK), Staatliches Museum für Naturkunde Stuttgart (SMNS), Zoologisches Forschungsmuseum Alexander Koenig (ZFMK), and Zoologische Staatssammlung München (ZSM). Acari are being investigated in another GBOL subproject (by Senckenberg Museum für Naturkunde Görlitz, SMNG).

For spiders, country-focused, taxonomically broad barcoding datasets have so far been published for Canada [61] (1018 species covered), for Slovenia and for Switzerland [66] (together 298 species) and, in a pilot project, for the Netherlands [68] (31 species) (for a list of ongoing European projects, see http://www.araneae.unibe.ch/barcoding/content/15/Barcoding-of-European-spiders). The present study contributes the first dataset of a spider barcoding campaign for Germany and the first dataset worldwide of this kind for harvestmen.

Materials and Methods

Sampling

For this study, 3537 arachnid specimens, 3339 Araneae and 198 Opiliones, were sampled from Germany (91% of the material) and neighboring countries. Within Germany, 24% of the specimens were collected in Baden-Württemberg and 13% in Schleswig-Holstein. Most other German states were represented by 6–10% of the German specimens each. Thuringia (0.1%), Hesse (1%) and Rhineland-Palatinate (2%) were less well represented, as were the city-states and the comparatively small Saarland (1%). Fig 1 illustrates the sampling pattern.

SMNK and external partners are responsible for and contributed 14% of the specimens, SMNS and partners 10%, ZFMK and partners 56%, ZSM and partners 20%.

To date 598 morphological arachnid species (561 spp. in spiders *vs*. 37 in harvestmen) in 269 genera (246 *vs*. 23) and 50 families (44 *vs*. 6) could be integrated. Setting this into relation

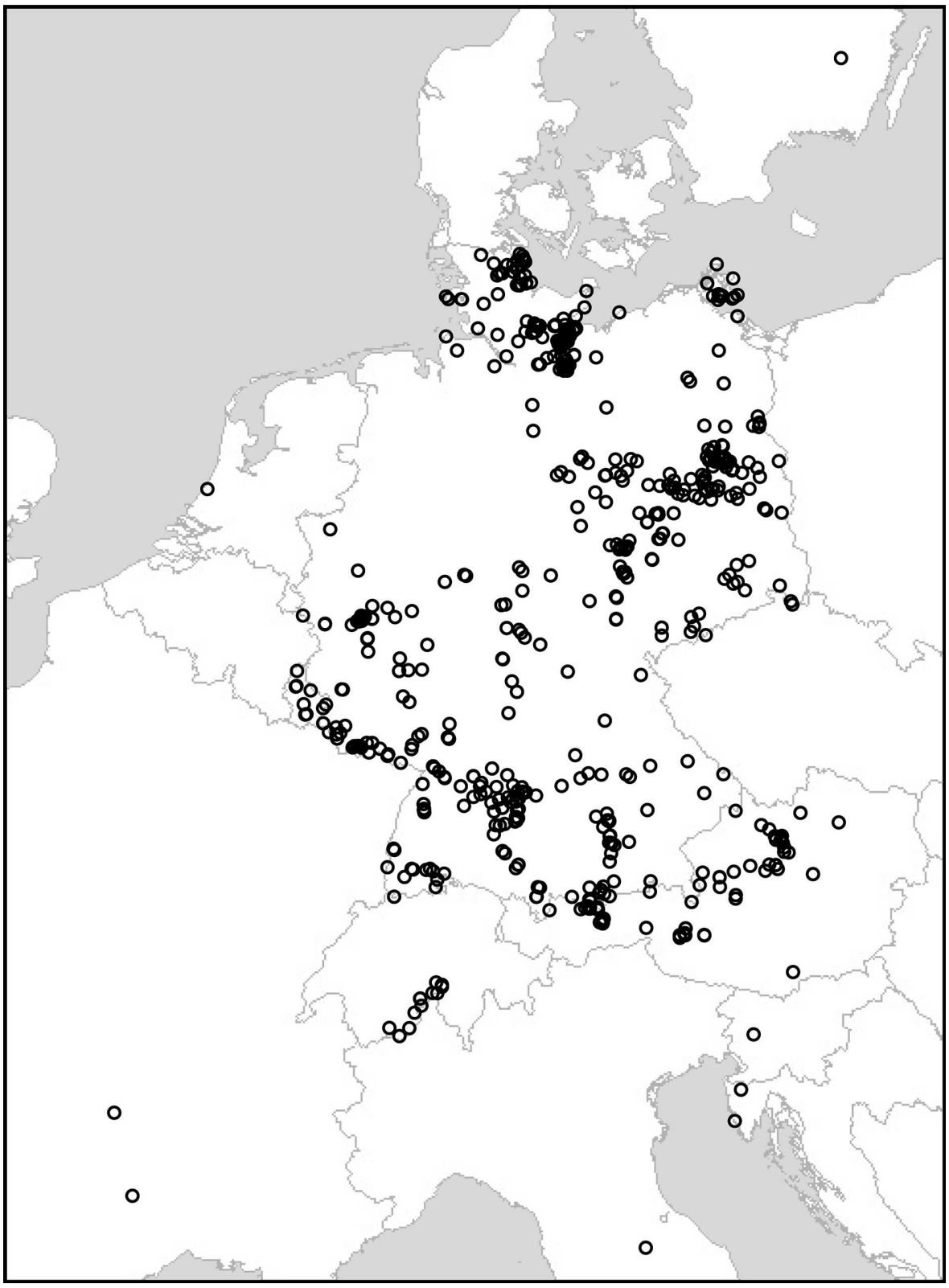

Fig 1. Geographic sampling of arachnid specimens underlying the present study. Image produced using GPS Visualizer (www.gpsvisualizer.com).

doi:10.1371/journal.pone.0162624.g001

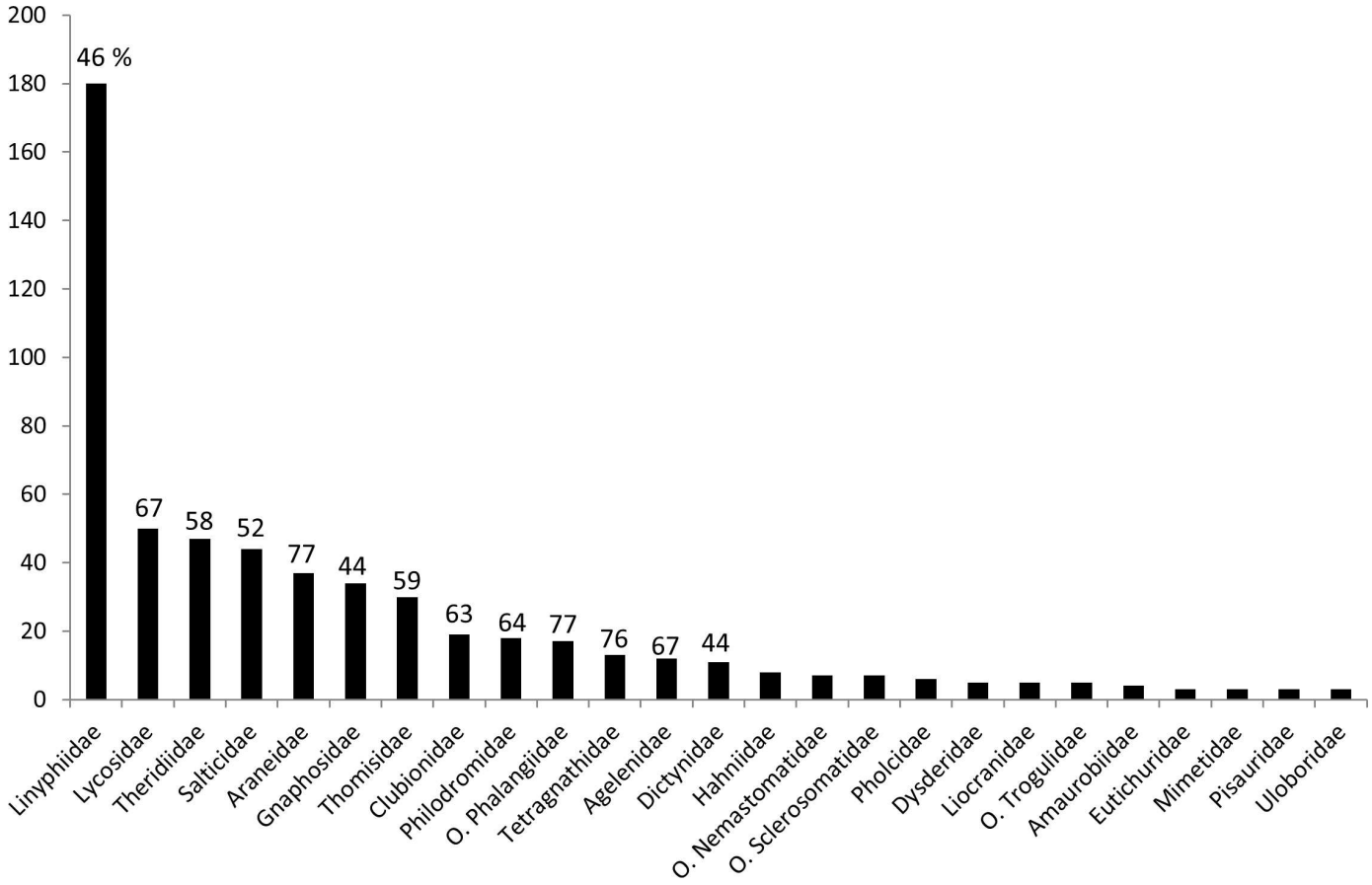

Fig 2. Number of species sampled per family (excluding families represented only by 1 or 2 species in this study). Numbers above bars are percentages showing species coverage for Germany, as derived from the checklists mentioned in the text. Family names prefixed with "O." belong to Opiliones, all others are spider families.

doi:10.1371/journal.pone.0162624.g002

with the German checklists [11,35], species coverage is 57% for spiders and 71% for harvestmen. Species numbers are plotted for the more frequent families in Fig 2.

The species in the dataset were represented by 6 individuals on average (median: 4). Almost 19%, i.e. 112 species, were 'singletons'. With 48 specimens, *Pardosa lugubris* was the species with most individuals; all other species were represented by 30 or fewer individuals (see Fig 3).

Most individuals (98%) were collected specifically for GBOL between the years 2011 and 2015. The oldest specimen processed in this study was collected in 2003.

Collecting was mostly done by hand, and most specimens were killed and preserved directly in 96% or 100% ethanol. 7% of the specimens were initially preserved in 70% water-diluted ethanol and 8% were collected in propylene glycol. The latter was used as capture fluid in pitfall traps; soon after identification, tissue for DNA extraction was transferred to absolute ethanol.

All material used in this study is property of the federal states of the involved institutions. Material acquired by these institutions is only accepted after a check that it was collected in compliance with national and international laws, regulations and conventions and that the material is free from third party rights. Furthermore, in order to become certified as a GBOL collector, it is required to accept the project's general terms and conditions, which demand abiding by the regulations of the Convention on Biological Diversity and national legislations. Field work permits were issued by the following authorities: Bayerisches Staatsministerium für

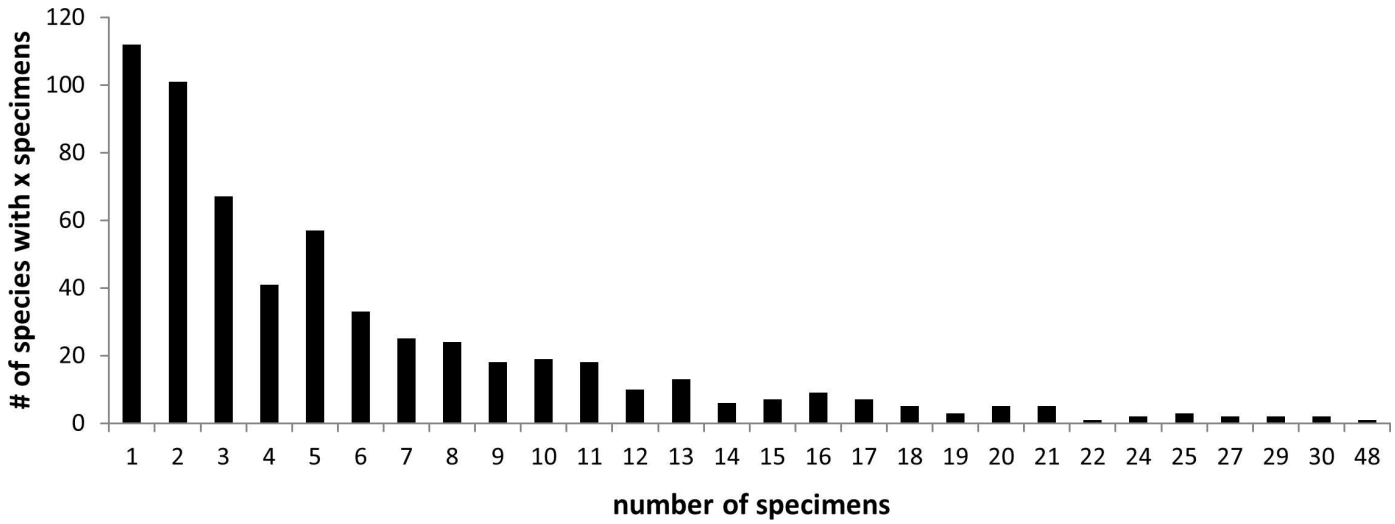

Fig 3. Number of specimens sampled per species. 19% of the species were 'singletons', while the median representation lay at 4 individuals per species.

doi:10.1371/journal.pone.0162624.g003

Umwelt und Gesundheit, München; Regierungspräsidium Stuttgart; Regierungspräsidium Karlsruhe; Struktur- und Genehmigungsdirektion Koblenz; Kreisverwaltung Rhein-Sieg-Kreis, Amt für Natur- und Landschaftsschutz; Amt für Umwelt, Verbraucherschutz und Lokale Agenda, Untere Landschaftsbehörde, Bonn; Nationalparkamt Müritz, Hohenzieritz; Biosphärenreservatsverwaltung Niedersächsische Elbtalaue, Hitzacker; Nationalparkforstamt Eifel, Schleiden-Gemünd; Amt für das Biosphärenreservat Südost-Rügen, Putbus; Landesamt für Umwelt, Naturschutz und Geologie Mecklenburg-Vorpommern, Güstrow; Landesamt für Landwirtschaft, Umwelt und ländliche Räume Schleswig-Holstein, Flintbek; Landrat Kreis Herzogtum Lauenburg; Landrat Kreis Rendsburg-Eckernförde, Fachdienste untere Naturschutzbehörde. The permits cover state forests, public land and protected areas as well as the five species of Arachnida protected in Germany: *Arctosa cinerea*, *Dolomedes fimbriatus*, *Dolomedes plantarius*, *Eresus cinnabarinus* and *Philaeus chrysops*.

Altogether, over 100 collectors contributed material. Field data for all analyzed specimens can be accessed in S1 Table. Juvenile specimens analyzed belong to taxa that are easily identifiable also in juvenile stage (e.g. based on coloration) or for which problematic ('look-alike') congeners do not occur in the study area. Juveniles that clustered conspicuously in the tree were removed from the dataset.

All morphological specimen vouchers and also molecular vouchers (DNA and often tissue) are deposited at and are available from the following four German public collections (permanent repositories): Staatliches Museum für Naturkunde Karlsruhe (SMNK), Staatliches Museum für Naturkunde Stuttgart (SMNS), Zoologisches Forschungsmuseum Alexander Koenig (ZFMK), Bonn, Zoologische Staatssammlung München (ZSM). All voucher numbers are given in S1 Table. The voucher IDs in S1 Table as well as the names in the trees include the institutional code, so that the association of a given sample to one of the GBOL partner institutes can be easily established.

Sequence data are available on BOLD [69] via DOI dx.doi.org/10.5883/DS-GBOLARA, and on GenBank. Specimen data will also be accessible, alongside specimen images, through the GBOL portal (www.bolgermany.de).

Molecular methods

Analyses were performed mostly in three separate laboratories: ZFMK, SMNS, and Canadian Centre for DNA Barcoding (CCDB) in Guelph. SMNS and ZFMK used their own facilities (up to the point of sequencing), SMNK samples were processed at both SMNS and ZFMK, ZSM samples at CCDB.

Total genomic DNA was usually isolated from legs. In very small specimens (especially many Linyphiidae) at ZFMK and SMNS, DNA was extracted non-destructively from whole specimens which were recovered after lysis (cf. [70]).

At ZSM, single legs were removed from each specimen and sent in 96 well lysis plates to CCDB for standardized DNA extraction, PCR amplification and bidirectional Sanger sequencing. CCDB lab protocols are available under www.ccdb.ca/resources.php.

At ZFMK and SMNS, silica-based methods were employed to extract DNA. SMNS followed the protocol by [71] with Pall AcroPrepTM 96 filter plates (Pall Corporation, Port Washington, NY, USA), while a Qiagen (Hilden, Germany) BioSprint96 magnetic bead extractor and corresponding kits were used at ZFMK.

Polymerase chain reaction for the 5' part of the mitochondrial cytochrome *c* oxidase subunit 1 (COI) gene was carried out, at ZFMK, in total reaction mixes of 20 μl, including 2 μl of undiluted DNA template, 0.8 μl of each primer (10 pmol/μl), and standard amounts of the reagents provided with the 'Multiplex PCR' kit from Qiagen (Hilden, Germany). At SMNS, PCR reactions of 25 μl volume contained 4 μl DNA, 5 units of Taq KAPA extra polymerase (KAPA-BIOSYSTEMS, Boston, USA), 10 μl of 5x KAPA Taq extra buffer, 3 μl of $MgCl_2$ (25 mM; both solutions provided by the manufacturer), 1 μl of each primer (10 pmol/μl), and 1 μl of 10 mM dNTP mix (Bioline, Luckenwalde, Germany).

The PCR primers used (also for sequencing) are given in Table 1.

Thermal cycling was performed, at ZFMK, on Applied Biosystems 2720 Thermal Cyclers (Life Technologies, Carlsbad, CA, USA), using a PCR program with two cycle sets, as a

Table 1. List of primers used for amplification and sequencing of the 5' part of the mitochondrial COI gene.

Primer name	Sequence	Publication	Used at
LCO1490	5'-GGTCAACAAATCATAAAGATATTGG	Folmer et al. 1994	SMNS, CCDB for ZSM, ZFMK
HCO2198	5'-TAAACTTCAGGGTGACCAAAAAATCA	Folmer et al. 1994	SMNS, CCDB for ZSM, ZFMK
LepF1	5'-ATTCAACCAATCATAAAGATATTGG	Hebert et al. 2004	CCDB for ZSM
LepR1	5'-TAAACTTCTGGATGTCCAAAAAATCA	Hebert et al. 2004	CCDB for ZSM
C_LepFolF	cocktail of LepF1 and LCO1490	www.boldsystem.org/index.php/Public_Primer_PrimerSearch	CCDB for ZSM
C_LepFolR	cocktail of LepR1 and HCO2198	www.boldsystem.org/index.php/Public_Primer_PrimerSearch	CCDB for ZSM
LCO1490-JJ	5'-CHACWAAYCATAAAGATATYGG	Astrin & Stüben 2008	ZFMK
HCO2198-JJ	5'-AWACTTCVGGRTGVCCAAARAATCA	Astrin & Stüben 2008	ZFMK
LCO1490-JJ2	5'-CHACWAAYCAYAARGAYATYGG	new	ZFMK
HCO2198-JJ2	5'-ANACTTCNGGRTGNCCAAARAATCA	new	ZFMK
LCO1490-JJ4a	5'-CNACNAAYCAYARRGAYATYGG	new	ZFMK
HCO2198-JJ4a	5'-AIACYTCNGGRTGICCAAARAATC	new	ZFMK
LCO1490-JJ4	5'-CIACIAAYCAYAARGAYATYGG	new	ZFMK
HCO2198-JJ4	5'-ANACTTCNGGRTGNCCAAARAATC	new	ZFMK

Species—also many others than arachnids—with strongly modified binding sites could usually be successfully amplified at ZFMK with a set of newly (manually) designed, highly degenerate primers (most often using combination LCO1490-JJ2 & HCO2198-JJ2). The combination LCO1490-JJ and HCO2198-JJ constitutes the standard set of primers used at ZFMK. The standard set of primers used at CCDB for ZSM was the combination C_LepFolF and C_LepFolR, a cocktail consisting of the primers listed above.

doi:10.1371/journal.pone.0162624.t001

combination of a 'touchdown' and a 'step-up' routine: first cycle set (15 repeats): 35 s denaturation at 94°C, 90 s annealing at 55°C (−1°C per cycle) and 90 s extension at 72°C. Second cycle set (25 repeats): 35 s denaturation at 94°C, 90 s annealing at 45°C, and 90 s extension at 72°C. At SMNS, PCR amplification was carried out in a Labcycler by SensoQuest (Göttingen, Germany). PCR conditions were: 35 cycles of 60 s denaturation at 93°C, 90 s annealing at 50°C and 60 s extension at 72°C.

PCR products were subsequently sent for bidirectional Sanger sequencing to various companies: ZFMK to BGI (Hong Kong, China) and Macrogen (Amsterdam, Netherlands), SMNS to LGC Genomics (Berlin, Germany) in 2013 and from 2014 on to GATC Biotech (Konstanz, Germany).

DNA sequence alignment was performed using parallelized MAFFT ver. 7.123 [72]. PAUP * ver. 4.0b10 [73] was used for p-distance transformations and for evaluating base composition and information content. Statistical parsimony networks [74] were calculated with the TCS algorithm [75] in PopART (http://popart.otago.ac.nz). Statistical evaluation of the data was performed using SPSS, R (box plots), Species Identifier ver. 1.7.7–3 [76] (extraction of 'splits' and 'lumps'), and the Perl script DiStats (intraspecific distances, individualized data on closest species pairs and on most distant congeners). DiStats has been developed for this study and is available, including documentation, under GitHub (https://github.com/mptrsen/distats) and through the ZFMK homepage (www.zfmk.de/en/research/research-centres-and-groups/distats). The script uses FASTA as input format, calculates *p*-distances or K2P distances and can be parallelized in order to process large datasets. It can produce two output files: a table with statistics for each species and optionally also the matrix of all pairwise distances in the dataset. For an alignment containing 1000 COI barcode sequences, the analysis will take around 6 minutes when using a single thread (on a 3.4 GHz processor). DiStats has an algorithmic complexity (*O*) of approximately $O(n^2)$, which means that run time increases exponentially with the number of input sequences (*n*). Using multiple CPU threads reduces the run time by a factor of 1/*c*, where *c* is the number of threads.

PAUP was also used for reconstruction of a phenetic neighbor-joining (NJ; [77]) tree as a quick molecular identification check. Phylogenetic reconstructions using Maximum Likelihood (ML; [78]) were performed with RAxML ver. 7.3.0 [79]. Evolutionary model selection for the ML analysis was implemented, using hierarchical likelihood ratio testing, in ModelGenerator ver. 0.85 [80] and indicated GTR + I + Γ as the best-fitting model [81]. The COI dataset was partitioned to treat 3rd codon positions separately from 1st and 2nd positions. The analysis was run for 1 million generations and included 1000 bootstrap replicates. For tree rooting purposes in NJ and ML analyses, we chose a mite sequence from BOLD as outgroup (see S1 Alignment).

Results

Average COI sequence length for the 3537 sequences was 650 bp. To accommodate many slightly shorter sequences, while avoiding genetic distance artifacts, alignment length was set to 653 bp. Sequences shorter than 500 bp were excluded from the analysis. The shortest included sequence was composed of 509 residues.

The dataset comprised 2099 distinct haplotypes, meaning that 1438 sequences were non-unique.

Among nucleotides, there was a compositional bias towards AT: 67.5%, which is close to levels previously reported for spiders (e.g. [4,82,83]). In detail, overall base composition was: A 25.3, C 13.3, G 19.2, T 42.2%.

Altogether, 5,572,791 pairwise distances were computed for spiders; of these, 17,867 were intraspecific distances. For Opiliones, there were 19,503 pairwise distances of which 896 were

Table 2. Estimators used to characterize genetic distance structure in the dataset.

[%]	intraspecific				interspecific			
	median	mean	range	95th perc.	median	mean	range	5th perc.
Araneae	0.3	0.7	0.0–10.1	2.5	17.5	17.4	0* - 28.2	13.6
Opiliones	0.2	1.3	0.0–8.9	8.1	19.3	19.4	7.0–30.1	13.6
Aran. K2P	0.3	0.7	0.0–11	2.6	20.0	20.0	0.0–35.6	15.0
Opil. K2P	0.2	1.3	0.0–9.5	8.7	22.6	22.7	7.5–38.6	15.1

The upper two rows indicate uncorrected distances for spiders and harvestmen, respectively, while the third and fourth rows give K2P distances (as required for a barcode data release). Median and mean distances are given for both intraspecific and interspecific distances, along with the range between the smallest and largest observation in the respective data category. *: There were cases of shared haplotypes among species, see text.

doi:10.1371/journal.pone.0162624.t002

distances between conspecific specimens. Table 2 gives an overview of intra- and interspecific distances, separated by order. Table 3 summarizes distances for all closest species pairs and for the most distant congeneric pairs; S2 and S3 Tables illustrate these in more detail, giving individual statistics by taxon. S2 Table (spiders) and S3 Table (harvestmen) furthermore indicate intraspecific distance ranges and central tendencies for all analyzed species. S4 Table individually lists the highest intraspecific distances in the dataset, while S5 Table gives the lowest interspecific distances.

The range covered by intraspecific (*p*-)distances was similar for spiders (0–10%) and harvestmen (0–9%), with an arithmetic mean of 0.7% in spiders and 1.3% in harvestmen. The influence of high outliers was stronger in the comparatively small harvestman dataset (median at 0.2% vs. mean at 1.3%), caused mostly by the deep splits within *Mitopus morio* (Fabricius, 1779) and *Phalangium opilio* Linnaeus, 1758 (both discussed below), but also in *Nemastoma lugubre* (Müller, 1776) (see S4 Table).

The interspecific distance range varied for the two arachnid orders: 0.0–28% in spiders, 7–30% in harvestmen. In four cases, haplotypes were shared among nominal spider species (see S5 Table, all discussed below): *Enoplognatha latimana* Hippa & Oksala, 1982 / *E. ovata* (Clerck, 1757); *Pardosa lugubris* (Walckenaer, 1802) / *P. saltans* Töpfer-Hofmann, 2000; *Tibellus maritimus* (Menge, 1875) / *T. oblongus* (Walckenaer, 1802); *Xysticus audax* (Schrank, 1803) / *X. cristatus* (Clerck, 1757). Interspecific arithmetic means were 17% for spiders, 19% for harvestmen.

Table 3. Statistics for closest species pairs and most distant congeneric pairs.

[%]	closest species pairs		most distant congener pairs	
	range	median (of all dist. for species pairs)	range	median (of largest distances)
Araneae	0.0* - 20.1 (22.0)	9.2	1.8–20.2	11.8
Opiliones	7.0–21.8 (22.8)	13.8	12.6–18.8	14.9
Araneae K2P	0.0–21.0 (28.6)	9.9	1.9–23.8	12.8
Opiliones K2P	7.5–24.0 (29.0)	15.5	13.7–22.0	16.5

The upper two rows indicate uncorrected distances for spiders and harvestmen, respectively, while the third and fourth rows give K2P distances (as required for a barcode data release). The range for the closest species pairs indicates the minimum and maximum among all *smallest* pairwise distances between closest species pairs. If a closest species pair is represented by several individuals, there may be larger distances as well: the maximal closest species distance in the respective dataset is indicated in parentheses. While for the closest species pairs no classificatorial (genus) background information was used, the last two columns in this table orient themselves at distances from representatives within the *same genus* (but different species). The range for the most distant congener pairs extracts the extremes among these *maximal* pairwise distances between farthest congeneric species pairs. Information on the individual closest species pairs and on the respective most distant congeneric species pairs is given in S2 Table for spiders and S3 Table for harvestmen.
* There were cases of shared haplotypes among species, see text.

doi:10.1371/journal.pone.0162624.t003

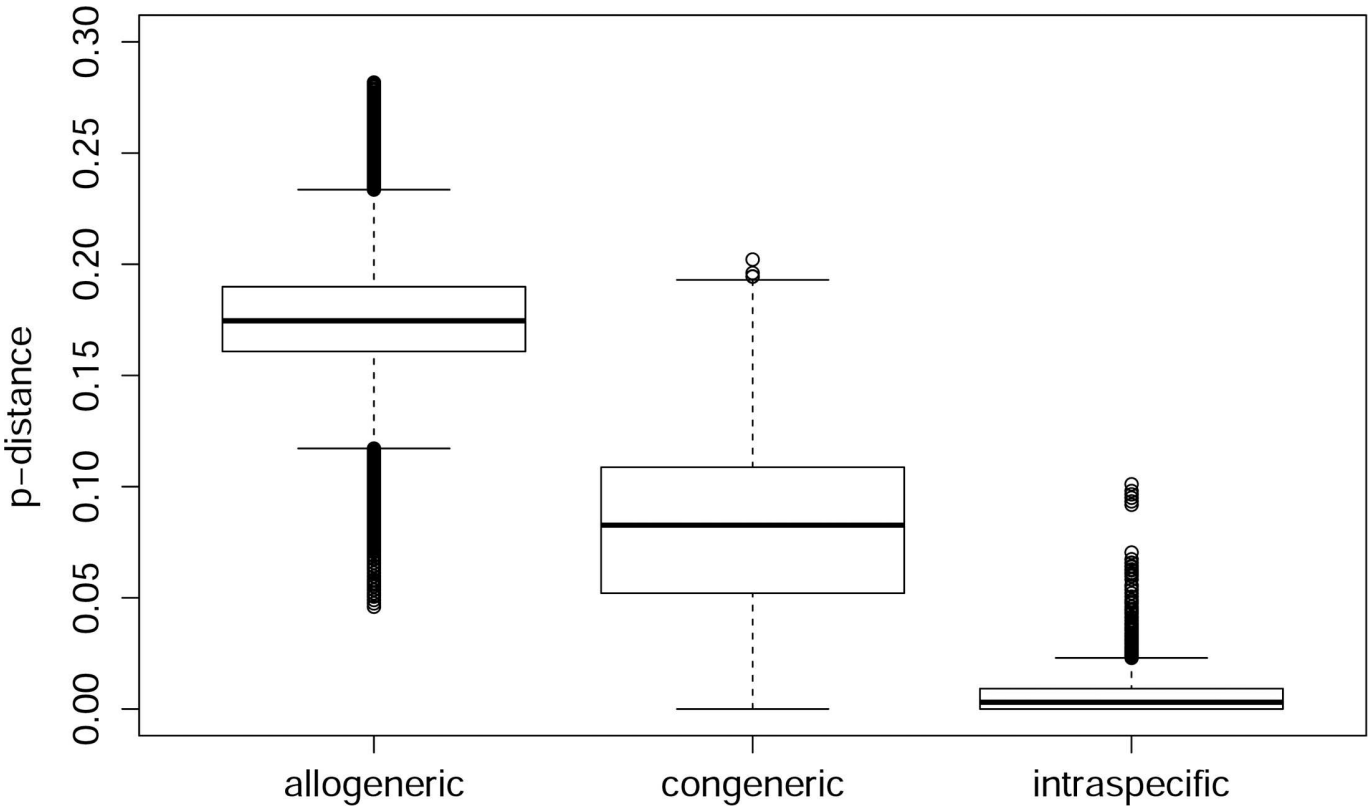

Fig 4. Box plot of *p*-distances for the order Araneae. Sorted by distance category: between specimens of different genera (allogeneric), between specimens belonging to different species, but to the same genus (congeneric), and between specimens that belong to the same species (intraspecific). Boxes indicate interquartile range (IQR: between upper [Q3] and lower [Q1] quartile). Black bars designate medians, whiskers indicate values within 1.5 × IQR beneath Q1 or 1.5 × above Q3. Circles depict outliers (above or below 1.5 × IQR).

doi:10.1371/journal.pone.0162624.g004

Mean intraspecific distances varied considerably among families. For the best-represented spider families, these lay between 0.4% (Agelenidae, Lycosidae, Philodromidae) and 1.0% (Tetragnathidae). Clubionidae, Linyphiidae and Theridiidae had mean intraspecific distances of 0.6%, Gnaphosidae 0.7%, Araneidae 0.8%, Salticidae and Thomisidae 0.9%.

Table 2 and the box plots (Figs 4 and 5) indicate that a universal barcoding gap is absent from the dataset. However, most of the species separate well; when ignoring the 5% most extreme outliers (a hypothetical scenario not surpassing the usual significance threshold), the barcoding gap for harvestmen would span 5.5% and for spiders even 11% (see Table 2). The median distance for closest species pairs was 9% in spiders and 13% in harvestmen.

A phenetic reconstruction using Neighbor Joining (NJ; S1 Fig) and a phylogenetic reconstruction using Maximum Likelihood (ML; S2 Fig) delivered trees in which the species overwhelmingly formed monophyletic clusters (but see discussion –several of the cases with conspicuous distances were also recovered as paraphyletic and polyphyletic). ML bootstrap analysis predominantly indicated very high support for species-level nodes, but did usually not allow much insight into deeper tree topology.

Discussion

Many previous studies using COI have shown that species differentiation via DNA barcoding is generally feasible and promising in arachnids (mostly spiders: e.g. [2–4,46,47,57–66]–but see

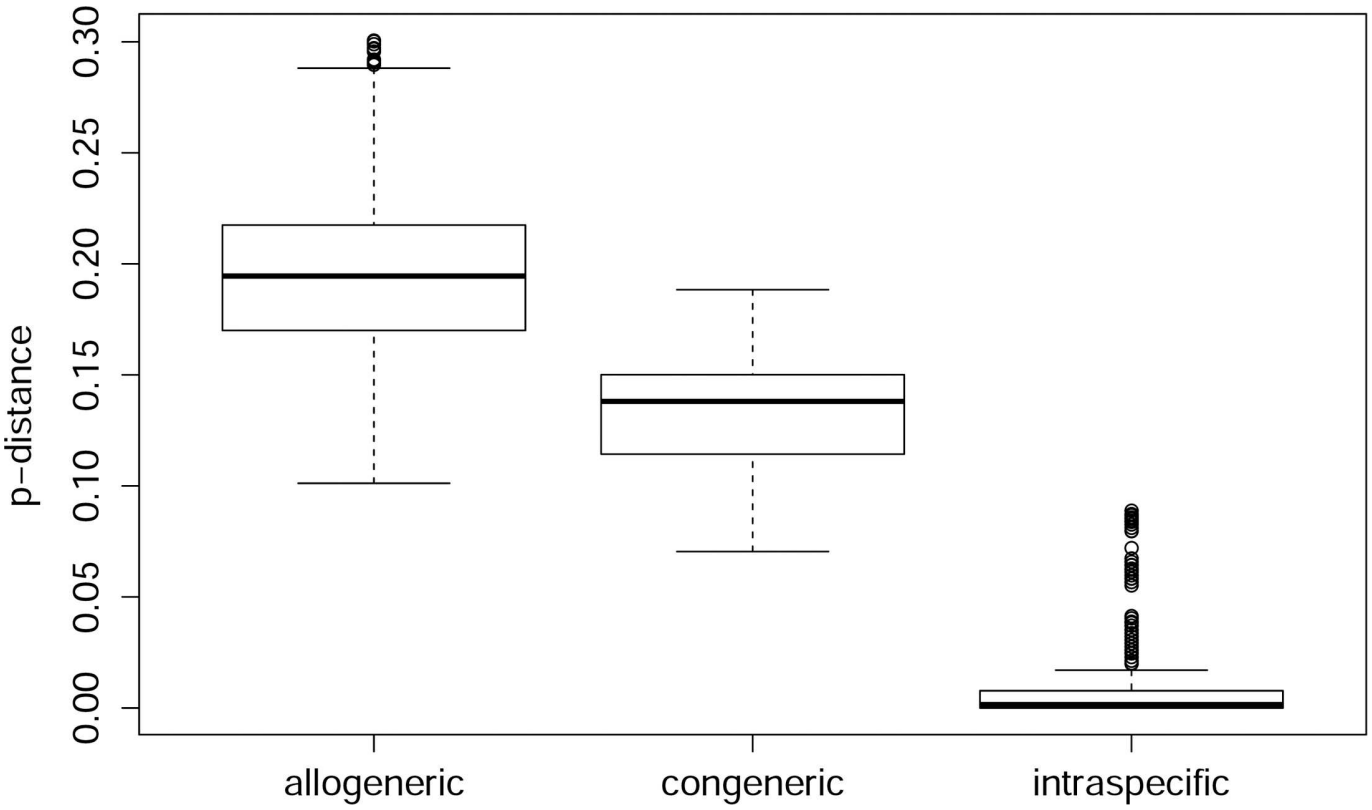

Fig 5. Box plot of *p*-distances for the order Opiliones. Legend: see Fig 4.

doi:10.1371/journal.pone.0162624.g005

e.g. [84–87]). This also applies to the present dataset. For most analyzed species, a so-called 'barcoding gap' exists: intraspecific sequence divergence levels are clearly lower than interspecific divergence to the nearest neighbor taxon in the dataset. This general tendency becomes evident from Table 2 and Table 3: the medians (and also the arithmetic means) of all distances between closest species lie around 9% in spiders and 13% in harvestmen, while in 95% of the cases, intraspecific distances are below 2.5% and 8%, with intraspecific medians at 0.7% and 0.2%.

However, despite the overall high suitability for barcoding of the dataset, we also encountered 19 currently valid species (3% of the dataset, all of them spiders) that are neither recovered monophyletic in the trees, nor in which the maximum intraspecific distance exceeds the distance to the nearest neighbor. Species determination via DNA barcoding fails in these instances. Since many if not most of the involved species pairs show discrete morphological differences, the explanation of such discrepancies between morphology and molecules should be regarded as a chance rather than a nuisance: it demands differentiated evolutionary hypotheses and directs further in-depth study that may result in intriguing biological insights [88].

Overall, the dataset contains 26 species with *p*-distances to the nearest interspecific neighbors below 2%. The most striking examples for difficult taxon separation from the GBOL dataset concern wolf spiders (Lycosidae). Wolf spiders alone contribute half of the 'barcode-resistant' cases mentioned above. The species pair *Pardosa lugubris/saltans*, for example, shows a pattern of completely intermixed haplotypes (Fig 6). It has been noted previously that "individuals of the *P. lugubris* group [containing additional species, e.g. *P. alacris*] cannot be identified by DNA barcoding, nor by ITS2 and 28S" [89]. The species in this complex are arguably

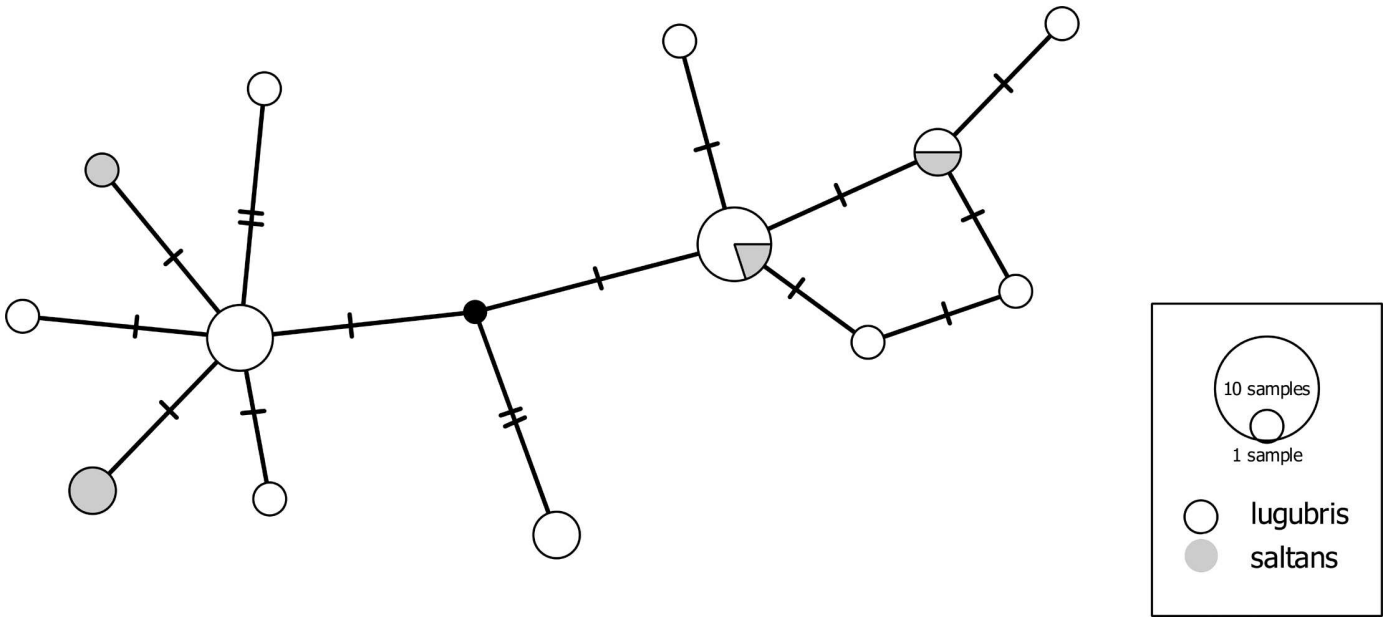

Fig 6. Haplotype network of the species pair *Pardosa lugubris/saltans*. To guarantee unequivocal morphological determination, only males were included. Small black dot indicates a hypothetical haplotype.

doi:10.1371/journal.pone.0162624.g006

well isolated by courtship behavior, while females, in particular, pose challenges also to morphological identification [90] (the latter were identified based mostly on [91]). To our surprise, a similar pattern of nonexistent haplotype segregation was detected in *Alopecosa cuneata* (Clerck, 1757) and *Alopecosa pulverulenta* (Clerck, 1757) (Fig 7). These are two of the most abundant spider species in Central European grassland ecosystems. Males are readily distinguished, even in the field, by the distinctive swelling of the front tibiae in *A. cuneata*. Furthermore, they show differences in details of the sexual organs and in courtship behavior [92]. Further examples of very low COI differentiation in Lycosidae include species within the *Pardosa pullata* group (S3 Fig) and several species pairs in the *P. monticola* group (*P. agrestis* (Westring, 1861) / *P. torrentum* Simon, 1876; *P. agrestis* / *P. palustris* (Linnaeus, 1758); *P. agrestis* / *P. monticola* (Clerck, 1757); *P. palustris* / *P. torrentum*). We speculate that the shallow mitochondrial divergence in many of the analyzed Lycosidae (but see [2]) may be related to the complex courtship behavior of these spiders [93]. A plausible mechanism is accelerated speciation through sexual selection. This could lead to fixation rates in male behavioral traits that exceed those of (putatively) neutral mitochondrial genes, as demonstrated for some jumping spiders (Salticidae) by [94]. These findings offer a promising perspective for detailed evolutionary and ethological studies.

In families other than wolf spiders, we encountered considerably fewer cases without barcoding gaps. In crab spiders (Thomisidae), the species pair *Xysticus audax/cristatus* is notorious for the difficult separation of females, while the male palps are clearly distinct [95]. In our data, the haplotypes of *X. audax* and *X. cristatus* are intermingled, while they are separated from the related species *X. gallicus* Simon, 1875 and *X. kochi* Thorell, 1872 (S4 Fig). Equally, *Tibellus maritimus* and *T. oblongus* (Philodromidae) are not separable by their COI sequences (S5 Fig), although morphological discrimination is rather straightforward. This result is supported by Canadian specimens as well [61]. In the comb-footed spiders (Theridiidae) we encountered two examples of very limited or even absent COI differentiation. *Enoplognatha*

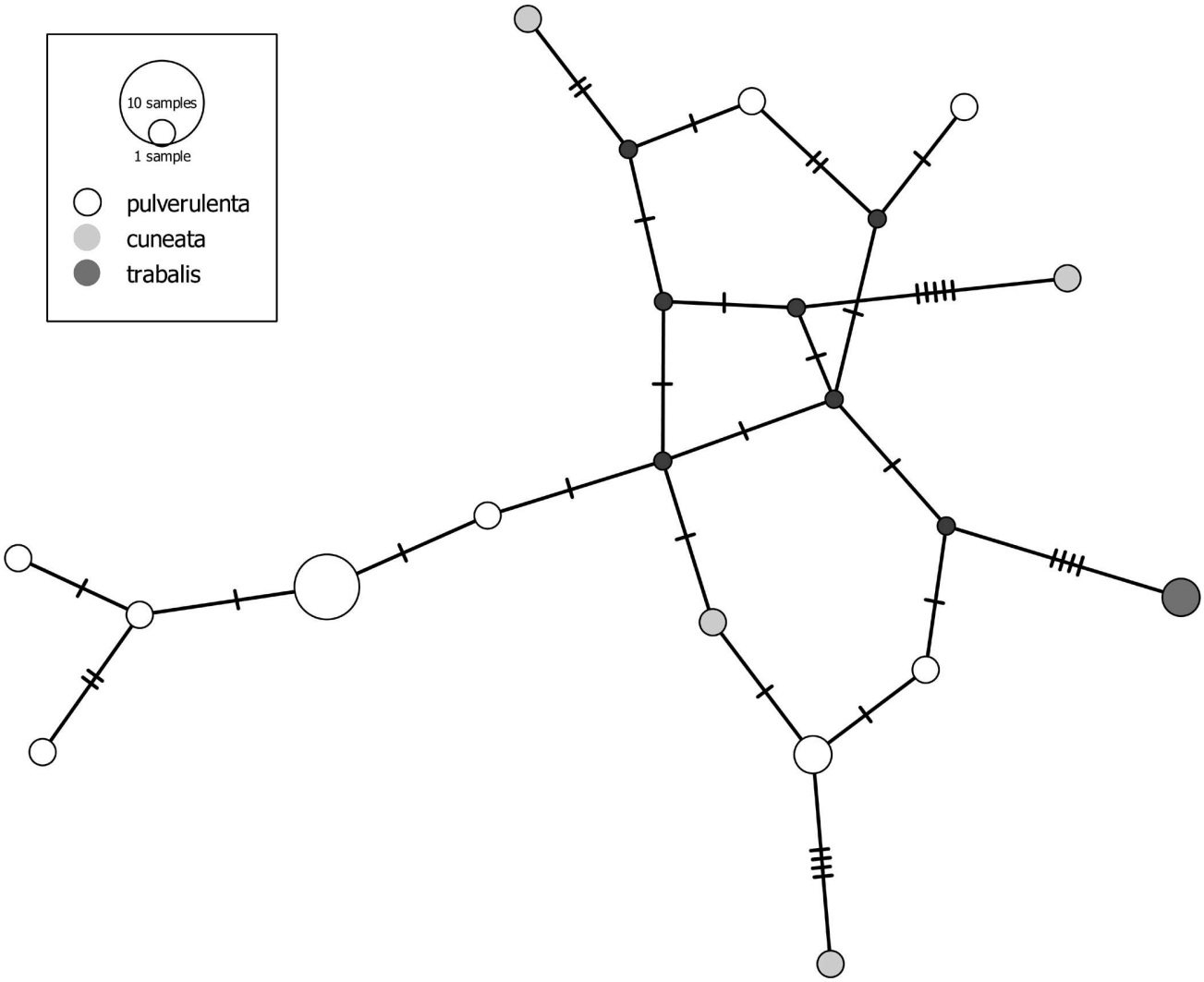

Fig 7. Haplotype network of three species of the *Alopecosa pulverulenta* group. To guarantee unequivocal morphological determination, only males were included (except for one female of *A. trabalis*). Small black dots indicate hypothetical haplotypes.

doi:10.1371/journal.pone.0162624.g007

latimana and *E. ovata* were not separated until 1982 [96]. Both species are widespread and abundant vegetation dwellers in Central Europe, often occurring syntopically. Although differences in the male palps are distinct and constant, the COI haplotypes are not segregated. An ongoing study at the University of Bern with more comprehensive sampling and comprising additional molecular markers questions the taxonomic status of these two nominal species [89]. Strikingly similar species are *Parasteatoda tepidariorum* (C. L. Koch, 1841) and *P. simulans* (Thorell, 1875). In the tree, the only sequence of *P. simulans* is nested within the relatively homogenous clade of *P. tepidariorum*. The only more or less solid morphological differences between these two species are the size dimensions (*P. tepidariorum* being significantly larger than *P. simulans*). Several authors have doubted the species status of *P. simulans* or treated it as subspecies of *P. tepidariorum* (e.g. [97–99]). In Germany, *P. tepidariorum* is usually found in buildings, while *P. simulans* also occurs outside, e.g. on the bark of trees. It is well conceivable that specimens living in marginal habitats stay smaller and develop a slightly different coloration pattern, representing ecological morphs within a species.

Linyphiidae, the most speciose spider family in Central Europe, contains a few examples of conspicuously shallow COI divergences among congeneric species (although most do not infringe a barcoding gap). These include *Agyneta ressli* (Wunderlich, 1973) / *A. rurestris* (C. L. Koch, 1836); *Gongylidiellum murcidum* Simon, 1884 / *G. vivum* (O. Pickard-Cambridge, 1875); *Hypomma bituberculatum* (Wider, 1834) / *H. cornutum* (Blackwall, 1833) / *H. fulvum* (Bösenberg, 1902); and *Tapinocyba affinis* Lessert, 1907 / *T. pallens* (O. Pickard-Cambridge, 1872). In all these instances, the species are distinguishable by consistent differences in at least the male sexual organs, even though distinction is subtle in some cases.

The processes behind the incomplete mitochondrial segregation in species of the latter families are possibly different from wolf spiders, which have a complex visual and acoustic courtship behavior. Alternative evolutionary explanations include the existence of distinct morphs within polymorphic species (e.g. [100,101]) or mitochondrial introgression, which has so far rarely been reported from spiders [102,103]. Detailed studies are required for each individual taxon to uncover the underlying mechanisms.

In recent years, great attention has been paid to the detection of cryptic diversity as reflected in deep intraspecific splits. Many new species have been described based on deep COI divergence within morphologically similar taxa (e.g. [104–109]). The GBOL dataset contains 48 species with a maximum intraspecific barcode divergence of $>$ 3% (26 species when looking at a maximal intraspecific distance of $>$ 4%). Interestingly, the proportion of species with conspicuously large intraspecific variation is considerably higher in Opiliones than in Araneae: 27% in harvestmen versus 8% in spiders (or 15% vs. 4% when using 4% as cutoff). This finding suggests that more cryptic diversity is to be expected in harvestmen than in spiders, a result that may be related to the comparatively reduced character complexity in the sexual organs of harvestmen.

A frequently observed pattern in our data is a single outlier haplotype found alongside a cluster of closely related sequences (e.g. in *Aelurillus v-insignitus* (Clerck, 1757); *Nemastoma lugubre* (Müller, 1776); *Steatoda bipunctata* (Linnaeus, 1758); *Clubiona corticalis* (Walckenaer, 1802); *Hypsosinga albovittata* (Westring, 1851); *Pardosa hortensis* (Thorell, 1872); *Centromerus pabulator* (O. Pickard-Cambridge, 1875); *Tetragnatha obtusa* C. L. Koch, 1837; *Steatoda albomaculata* (De Geer, 1778); *Robertus lividus* (Blackwall, 1836); *Xysticus lanio* C. L. Koch, 1835; in order of descending divergence). In the case of the largest intraspecific barcode divergence, *Aelurillus v-insignitus* (maximum *p*-distance 10.1%), we can trace back the deep split to differences between specimens of the gray and black morphs, which are well distinguished morphologically [110] and may represent separate species. Likewise, the split within *Steatoda bipunctata* (maximum *p*-distance 7%) is corroborated by external evidence, as our outlier specimen from Berlin shares an identical barcode with two specimens from Canada/Nova Scotia (submitted to BOLD by G. Blagoev and colleagues), hinting at a so far unrecognized sibling species.

In other cases of single outlier sequences we refrain from further interpretation. Although we took greatest care in the detection of numts (nuclear mitochondrial DNA) [111] and processing errors, sequencing artefacts cannot be completely ruled out.

Nonetheless, numerous examples remain of currently valid species where (multiple) sequences fall into two or more clearly distinct COI clusters. A representative case is the opilionid *Mitopus morio* (Fabricius, 1779), the harvestman species with the widest geographic distribution and the highest abundance in European mountain ecosystems. The taxon shows a remarkable altitudinal variation in leg length and dorsal coloration pattern. [41] investigated the genetic structure along two altitudinal transects in the Alps and found three deeply diverged lineages which, however, did not correspond to leg morphometric variants. The 15 GBOL sequences of *Mitopus morio* fell into four deeply diverged clades (Fig 8 and S1 Fig).

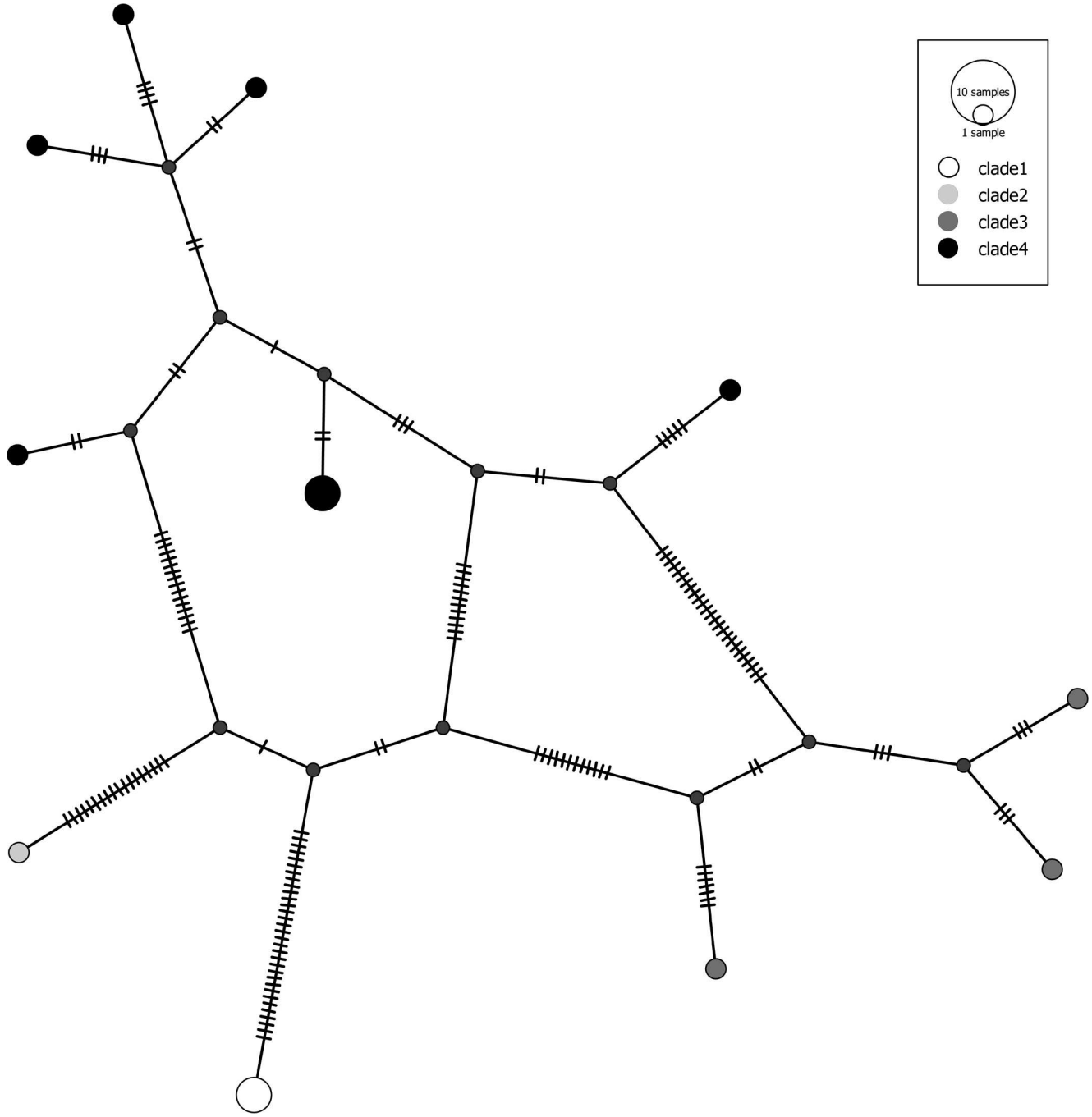

Fig 8. Haplotype network of *Mitopus morio*.

doi:10.1371/journal.pone.0162624.g008

Sequences of clade 1 originate from the German Alps and the Bavarian Forest, the single specimen of clade 2 comes from the surroundings of Berlin, clade 3 is restricted to the Alps (Karwendel and Wallis), while clade 4 appears widespread in Central Europe. Thus, specimens of three clades occurr in the Alps and may well correspond to the lineages described by [41]. However, the true diversity in Central Europe may be even higher and available names

currently in synonymy of *M. morio* may deserve revalidation (e.g. *Mitopus ericaeus* Jennings, 1962 from Great Britain). A similar situation applies to the common harvestman *Phalangium opilio* Linnaeus, 1758. This species is known for extreme variation in body size and in length of the conspicuous process on the second cheliceral segment, but due to the apparently continuous variation, all morphological variants have been considered conspecific [39]. The 12 GBOL sequences of *Phalangium opilio* split into two clades that are separated by *p*-distances of 8.3–8.7% and show a sympatric distribution in Germany.

Also for several spider species, sequences fall into two deeply diverged clusters: *Tmarus piger* (Walckenaer, 1802); *Micaria pulicaria* (Sundevall, 1831); *Nigma walckenaeri* (Roewer, 1951); *Sitticus pubescens* (Fabricius, 1775); *Hahnia nava* (Blackwall, 1841); *Euophrys frontalis* (Walckenaer, 1802); *Salticus scenicus* (Clerck, 1757); *Xysticus kochi* Thorell, 1872; *Theridion familiare* O. Pickard-Cambridge, 1871; *Haplodrassus dalmatensis* (L. Koch, 1866); *Heliophanus flavipes* (Hahn, 1832), or even into multiple clusters: *Tetragnatha extensa* (Linnaeus, 1758), *Drassodes lapidosus* (Walckenaer, 1802), *Neon reticulatus* (Blackwall, 1853), *Haplodrassus signifer* (C. L. Koch, 1839) (ordered by descending genetic divergence, see S5 Table for the respective distance values). All these species show pairwise intraspecific distances between 3% and 7%. One plausible explanation for comparably high intraspecific divergence is isolation by distance in dispersal-limited species. We did not find indications for allopatric distribution of clusters in any of these species, but the limited size of the datasets in most species precludes more detailed analyses at this stage. Without doubt, all these taxa deserve a thorough taxonomic reconsideration. The GBOL dataset can be a convenient starting point to that end; it offers useful guidance for taxonomists to select promising study objects.

Finally, the GBOL data provide five new records for Germany: one on national scale, and seven at federal state level. The species *Sibianor larae* Logunov, 2001 and *Evansia merens* O. Pickard-Cambridge, 1900 have been recorded for the first time in Baden-Württemberg, in mountainous, relatively humid heathland in the Black Forest. *S. larae* has been recorded in the same type of open country habitat in the Netherlands [112]. *Oreonetides glacialis* (L. Koch, 1872) could be recorded for the first time in Bavaria (arguably also for the whole country). It was collected–in both sexes–as the dominant spider species on a barren karst plateau on the Zugspitze (at 2600 m.a.s.l.; leg. J. Spelda, S. Friedrich & R. Melzer) among scree, the typical habitat for this species. The crab spider *Xysticus acerbus* Thorell, 1872 is the first record for Mecklenburg-Vorpommern (leg. C. Muster). Finally, for Schleswig-Holstein, *Hahnia ononidum* Simon, 1875, *Mermessus trilobatus* (Emerton, 1882), *Glyphesis servulus* (Simon, 1881) (leg. M. Lemke) all represent new records.

Pholcus Walckenaer, 1805 is the most species-rich genus in Pholcidae, with most of the currently 329 species in tropical and subtropical regions (www.pholcidae.de). Only two widespread species have previously been recorded in Germany: the cosmopolitan synanthropic *Pholcus phalangioides* (Fuesslin, 1775) and the Mediterranean to Central Asian anthropophilic *P. opilionoides* (Schrank, 1781). Several further representatives of the genus occur in and around human buildings and have probably for this reason attained wide distributional ranges [113]. Among them is the East European to Central Asian *P. alticeps* Spassky, 1932, whose most western record so far was from Poland [113]. Our sequenced specimen originates from eastern Germany (Sachsen, Dresden-Kaditz). At the same locality, a vital population of *P. alticeps* (adult males and females as well as juveniles) was observed in June 2015 (leg. C. Muster), co-existing with *P. phalangioides*. Specimens were collected from a cellar, a barn, and outdoors from the wall of a building. Thus, the species is probably well established at this locality.

Conclusion

For ca. 60% of the German spider fauna and ca. 70% of the country's harvestman fauna, the dataset and material basis provided through this study enable fast, reliable and reproducible species identification via barcoding and highlight the species with inherent problems connected to this type of identification or to current taxonomy.

Building extensive, carefully validated reference databases ('libraries') is the most relevant prerequisite for successful DNA barcoding applications. In this context, our project should considerably facilitate DNA-based species identification of Araneae and Opiliones in Germany, for non-specialists as well as for large-scale biodiversity monitoring endeavors. The latter is envisioned in a campaign currently proposed for Germany and is already implemented (on a much smaller scale) in the second phase of the German Barcode of Life Project. Within this project, compiling the reference database and reference collections is still an ongoing effort, for arachnids as well as for many other taxa.

Supporting Information

S1 Alignment. Sequence data for the 3538 analyzed arachnid specimens. Includes the mite outgroup retrieved from BOLD: BOLDMSACA57112_OG_Acari. FASTA-formatted. See S1 Table for more details on specimens.
(FAS)

S1 Fig. Neighbor Joining tree. PDF can be searched for species names. Apart from ID and species name, life stage, sex and coordinates of collecting locality are given. See S1 Table for more details on individual specimens in the tree.
(PDF)

S2 Fig. Maximum Likelihood tree with bootstrap values. The analysis was run for 1 million generations and includes 1000 bootstrap replicates. Apart from ID and species name, life stage, sex and coordinates of collecting locality are given. See S1 Table for more details on individual specimens in the tree.
(PDF)

S3 Fig. Haplotype network of three species of the *Pardosa pullata* group. Small black dots indicate hypothetical haplotypes.
(TIF)

S4 Fig. Haplotype network of the species complex *Xysticus audax* and *X. cristatus*, along with their closest relatives. To guarantee unequivocal morphological determination, only males were included. Small black dots indicate hypothetical haplotypes.
(TIF)

S5 Fig. Haplotype network of the species pair *Tibellus maritimus* and *T. oblongus*. Small black dots indicate hypothetical haplotypes.
(TIF)

S1 Table. Field data and IDs for all analyzed specimens. This table lists collecting date and location (incl. GPS coordinates), collector, taxonomy, identifier, preservation fluid, life stage and sex for the specimens analyzed. Sample IDs in this table correspond to those given in S1 Fig (NJ tree) and S2 Fig (ML tree), as well as in S4 Table ('splits') and S5 Table ('lumps'). Please note that while working on the release dataset, some species names have changed: *Dictyna civica -> Brigittea civica*, *Dictyna latens -> Brigittea latens*, *Hahnia difficilis -> Iberina difficilis*, *Hahnia montana -> Iberina montana*, *Lepthyphantes keyserlingi -> Ipa keyserlingi*,

Titanoeca psammophila -> Titanoeca spominima. For these species, the old names are used throughout the article and related materials.
(XLSX)

S2 Table. List of closest species pairs and most distant congeneric species pairs for spiders. Statistics, individual by species, for three types of distance comparisons (intraspecific, closest interspecific, largest congeneric): minimal, maximal, mean and median *intraspecific* genetic distances; *closest species* (by distance) and minimal, maximal and median distance separating the two species; genetically most distant species within the same genus along with maximal distance in separating the two species. In case of identical distances to reference species, two or more rows under the same species name are used for listing all these cases–one line for each allogeneric comparison (note: the DiStats script used to compute these values considers the full number of decimal places during comparison/sorting of distances, even if output is set to contain only 2 decimal places, as in DiStats default mode).
(XLSX)

S3 Table. List of closest species pairs and most distant congeneric species pairs for harvestmen. Legend: see caption for S2 Table.
(XLSX)

S4 Table. Highest intraspecific distances, 'splits'. This table contains the 352 pairwise comparisons with the highest conspecific *p*-distances in the dataset, ranging from 10 to 3% (range 5 to 3% given in gray, denoting an 'uncertainty zone' for average species limits in this scenario). 164 comparisons have values above 4%. Specimens are identified through species name and ID (see S1 Table for more details).
(XLSX)

S5 Table. Lowest interspecific distances, 'lumps'. This table contains the 731 pairwise comparisons with the lowest allospecific (but congeneric) *p*-distances in the dataset, ranging from 0 to 5%. 353 comparisons have values below 3%. Specimens are identified through species name and ID (see S1 Table for more details).
(XLSX)

Acknowledgments

The GBOL project is financed by the German Federal Ministry of Education and Research (BMBF #01LI1101). Part of the sequences contributed by ZSM resulted from funding by the Bavarian State Ministry of Education and Culture, Science and the Arts (Barcoding Fauna Bavarica, BFB). The sequencing work of ZSM was supported, in part, by funding from the Government of Canada to Genome Canada through the Ontario Genomics Institute, while the Ontario Ministry of Research and Innovation and NSERC supported development of the BOLD informatics platform.

We thank the lab crews for unfailingly producing lots of high-quality results: Laura von der Mark and Jana Thormann (ZFMK), Hong Shen (SMNS) and the CCDB team (ZSM). Gerry Blagoev (CCDB) provided valuable data for sequence comparisons, engaged us in insightful discussions and critically reviewed the manuscript. We equally thank Matjaž Kuntner and an anonymous reviewer for improving the manuscript. The 1KITE project (1KITE consortium: www.1kite.org) kindly provided access to mitogenome data for hundreds of samples that greatly helped in designing the new degenerate primers used in GBOL/ZFMK. Thank you to Peter Grobe and Dirk Steinke for orchestrating the data transfer from GBOL to BOLD. Thank you to Adam Schneider for providing GPS Visualizer to the community. Last but not least,

many thanks also to all collectors and identifiers who support the GBOL Arachnida subproject by contributing specimens (see S1 Table) or species identifications. To list (alphabetically) just those with more—often many more—than 10 contributed samples: P. Bergmann, Th. Blick, J. Esser, M. Freudenschuss, L. Friman, V. Hemm, T. Klug, T. Kothe, L. Krogmann, S. Leidenroth, T. Maier, E. Merches, F. Meyer, H.P. Reike, A. Sch0078nhofer, W. Walbaum, D. Weber, I. Wendt, T. Wesener, J. Wunderlich (please excuse if in the meantime, with more material accumulating after the 'cut' for the release dataset, more specimens have been contributed and more arachnologists have joined the team).

Author Contributions

Conceptualization: JJA SB LH HH JH JCM JM CM BR JS.

Data curation: JJA SB HH JH HJK JCM BR JS.

Formal analysis: JJA CM JS.

Funding acquisition: JJA LH HH JH JS.

Investigation: JJA SB HH JH BAH KHK HJK ML CM JS.

Methodology: JJA SB CM BR.

Project administration: JJA LH HH JH HJ HJK JCM JM BR JS.

Resources: SB HH JH BAH KHK HJK ML CM JS.

Software: HJ MP.

Validation: SB HH JH BAH HJK CM JS.

Visualization: JJA CM BR.

Writing – original draft: JJA SB HH BAH CM MP BR JS.

Writing – review & editing: JJA SB LH HH JH BAH HJ KHK HJK ML JCM JM CM MP BR JS.

References

1. Hebert P, Ratnasingham S, DeWaard J. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc London, Ser B Biol Sci. 2003; 270 Suppl1: 96–99.
2. Čandek K, Kuntner M. DNA barcoding gap: reliable species identification over morphological and geographical scales. Mol Ecol Resour. 2015; 15: 268–277. doi: 10.1111/1755-0998.12304 PMID: 25042335
3. Barrett RDH, Hebert PDN. Identifying spiders through DNA barcodes. Can J Zool. 2005; 83: 481–491.
4. Astrin JJ, Huber BA, Bernhard M, Klutsch CFC. Molecular taxonomy in pholcid spiders (Pholcidae, Araneae): evaluation of species identification methods using CO1 and 16S rRNA. Zool Scr. 2006; 35: 441–457.
5. World Spider Catalog. World Spider Catalog. Natural History Museum Bern. 2016; version 17.0. Available: http://wsc.nmbe.ch.
6. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on Earth and in the Ocean? PLOS Biol. 2011; 9: e1001127. doi: 10.1371/journal.pbio.1001127 PMID: 21886479
7. Entling W, Schmidt MH, Bacher S, Brandl R, Nentwig W. Niche properties of Central European spiders: Shading, moisture and the evolution of the habitat niche. Glob Ecol Biogeogr. 2007; 16: 440–448. doi: 10.1111/j.1466-8238.2006.00305.x
8. Gerlach J, Samways M, Pryke J. Terrestrial invertebrates as bioindicators: An overview of available taxonomic groups. J Insect Conserv. 2013; 17: 831–850. doi: 10.1007/s10841-013-9565-9
9. Wise DH. Spiders in ecological webs. Cambridge: Cambridge University Press; 1993.

10. Moya-Larano J, Foellmer M, Pekar S, Arnedo M, Bilde T, Lubin Y. Evolutionary ecology: linking traits, selective pressures and ecological factors. In: Penney D, editor. Spider Research in the 21st Century: trends and perspectives. Manchester: Siri Scientific; 2013. pp. 112–153.
11. Blick T, Finch O-D, Harms KH, Kiechle J, Kielhorn K-H, Kreuels M, et al. Rote Liste und Gesamtartenliste der Spinnen (Arachnida: Araneae) Deutschlands. Naturschutz und Biol Vielfalt. 2016; 70: 383–510.
12. Muster C. Biogeographie von Spinnentieren der mittleren Nordalpen (Arachnida: Araneae, Opiliones, Pseudoscorpiones). Verhandlungen des Naturwissenschaftlichen Vereins Hambg. 2001; 39: 5–196.
13. Loch R. Statistisch-ökologischer Vergleich der epigäischen Spinnentierfauna von Bann- und Wirtschaftswäldern. Berichte Freiburg Forstl Forsch. 2002; 38: 1–249.
14. Finch O-D, Blick T, Schuldt A. Macroecological patterns of spider species richness across Europe. Biodivers Conserv. 2008; 17: 2849–2868.
15. Höfer H, Blick T, Muster C, Paulsch D. Artenvielfalt und Diversität der Spinnen (Araneae) auf einem beweideten Allgäuer Grasberg (Alpe Einödsberg) und unbeweideten Vergleichsstandorten im Naturschutzgebiet Allgäuer Hochalpen. Andrias. 2010; 18: 53–78.
16. Blick T. Spider coenoses in strict forest reserves in Hesse (Germany). In: Nentwig W, Entling MH, Kropf C, editors. 24th European Congress of Arachnology. Natural History Museum Bern; 2010. pp. 11–29.
17. Blick T. Abundant and rare spiders on tree trunks in German forests (Arachnida: Araneae). Arachnol Mitteilungen. 2011; 40: 5–14.
18. Hemm V, Höfer H. Effects of grazing and habitat structure on the epigeic spider fauna in an open xerothermic area in southern Germany. Bull Br Arachnol Soc. 2012; 15: 260–268.
19. Pearce JL, Venier LA. The use of ground beetles (Coleoptera: Carabidae) and spiders (Araneae) as bioindicators of sustainable forest management: A review. Ecol Indic. 2006; 6: 780–793.
20. Buchholz S. Ground spider assemblages as indicators for habitat structure in inland sand ecosystems. Biodivers Conserv. 2010; 19: 2565–2595.
21. Haase H, Balkenhol B. Spiders (Araneae) as subtle indicators for successional stages in peat bogs. Wetl Ecol Manag. 2014; 22: 1–12.
22. Noreika N, Kotiaho JS, Penttinen J, Punttila P, Vuori A, Pajunen T, et al. Rapid recovery of invertebrate communities after ecological restoration of boreal mires. Restor Ecol. 2015; 23. doi: 10.1111/rec.12237
23. Scott AG, Oxford GS, Selden PA. Epigeic spiders as ecological indicators of conservation value for peat bogs. Biol Conserv. 2006; 127: 420–428. doi: 10.1016/j.biocon.2005.09.001
24. Muster C, Gaudig G, Krebs M, Joosten H. *Sphagnum* farming: the promised land for peat bog species? Biodivers Conserv. Springer Netherlands; 2015; 1989–2009. doi: 10.1007/s10531-015-0922-8
25. Martin D. Zur Autökologie der Spinnen (Arachnida: Araneae) 1. Charakteristik der Habitatausstattung und Präferenzverhalten epigäischer Spinnenarten. Arachnol Mitteilungen. 1991; 1: 5–26.
26. Platen R, Moritz M, Broen BV. Liste der Webspinnen- und Weberknechtarten (Arach.: Araneida, Opilionida) des Berliner Raums und ihre Auswertung für Naturschutzzwecke (Rote Liste). Landschaftsentwicklung und Umweltforsch. 1991; 6: 169–205.
27. Platen R. A method to develop an "indicator value" system for spiders using canonical correspondence analysis (CCA). Mem Queensl Museum. Brisbane: Proceedings of the 12th International Congress of Arachnology; 1993;33: 621–627.
28. Hänggi A, Stöckli E, Nentwig W. Lebensräume Mitteleuropäischer Spinnen. Charakterisierung der Lebensräume der häufigsten Spinnenarten Mitteleuropas und der mit diesen vergesellschafteten Arten. Misc Faun Helv. Centre suisse de cartographie de la fauna (CSCF); 1995; 4: 1–460.
29. Buchar J, Ružicka V. Catalogue of spiders of the Czech Republic. Praha: Peres; 2002.
30. Duffey E. Spider habitat classification and the development of habitat profiles. Bull Br Arachnol Soc. 2010; 15: 1–20.
31. Brand C, Höfer H, Beck L. Zur Biologie eines Buchenwaldbodens 16. Die Spinnenassoziation einer Windbruchfläche. Carolinea. 1994; 52: 61–74.
32. Riecken U. The importance of semi-natural landscape structures in an agricultural landscape as habitats for stenotopic spiders. Proc 17th Eur Colloq Arachnol Edinburgh 1997. 1998; 17: 301–310.
33. Nentwig W, Blick T, Gloor D, Hänggi A, Kropf C. Spiders of Europe. 2015. Available: www.araneae.unibe.ch.
34. Blick T, Harvey MS. Worldwide catalogues and species numbers of the arachnid orders (Arachnida). Arachnol Mitteilungen. 2011; 41: 41–43. doi: 10.5431/aramit4108

35. Muster C, Blick T, Schönhofer A. Rote Liste und Gesamtartenliste der Weberknechte (Arachnida: Opiliones) Deutschlands. Naturschutz und Biol Vielfalt. 2016; 70: 513–536.

36. Curtis DJ, Machado G. Ecology. In: Pinto-da-Rocha R, Machado G, Giribet G, editors. The Biology of Opiliones. 2007. pp. 280–308.

37. Komposch C. Rote Liste der Weberknechte (Opiliones) Österreichs. In: Zulka P, editor. Rote Liste gefährdeter Tiere Österreichs Checklisten, Gefährdungsanalysen Teil 3 Handlungsbedarf. Wien: Böhlau, Wien: Grüne Reihe des Lebensministeriums 14/3; 2009. pp. 397–483.

38. Muster C, Meyer M, Sattler T. Spatial arrangement overrules environmental factors to structure native and non-native assemblages of synanthropic harvestmen. PLoS One. 2014; 9 e90474: doi: 10.1371/journal.pone.0090474 PMID: 24595309

39. Martens J. Spinnentiere, Arachnida—Weberknechte, Opiliones. Die Tierwelt Deutschlands 64. Jena: G. Fischer; 1978.

40. Schönhofer AL, Martens J. Hidden Mediterranean diversity: Assessing species taxa by molecular phylogeny within the opilionid family Trogulidae (Arachnida, Opiliones). Mol Phylogenet Evol. 2010; 54: 59–75. doi: 10.1016/j.ympev.2009.10.013 PMID: 19840858

41. Arthofer W, Rauch H, Thaler-Knoflach B, Moder K, Muster C, Schlick-Steiner BC, et al. How diverse is *Mitopus morio*? Integrative taxonomy detects cryptic species in a small-scale sample of a widespread harvestman. Mol Ecol. 2013; 22: 3850–3863. doi: 10.1111/mec.12340 PMID: 23731459

42. Wachter GA, Muster C, Arthofer W, Raspotnig G, Fottinger P, Komposch C, et al. Taking the discovery approach in integrative taxonomy: decrypting a complex of narrow-endemic alpine harvestmen (Opiliones: Phalangiidae: *Megabunus*). Mol Ecol. 2015; 1–27.

43. Hebert PDN, Cywinska A, Ball SL, de Waard JR. Biological identifications through DNA barcodes. Proc R Soc B. 2003; 270: 313–321. PMID: 12614582

44. Tanikawa A. The first description of a male of *Paraplectana tsushimensis* (Araneae: Araneidae). Acta Arachnol. 2011; 60: 71–73.

45. Correa-Ramirez MM, Jimenez ML, Garcia-De Leon FJ. Testing species boundaries in *Pardosa sierra* (Araneae: Lycosidae) using female morphology and COI mtDNA. J Arachnol. 2010; 38: 538–554.

46. Bayer S, Schönhofer AL. Phylogenetic relationships of the spider family Psechridae inferred from molecular data, with comments on the Lycosoidea (Arachnida: Araneae). Invertebr Syst. 2013; 27: 53–80.

47. Blagoev GA, Nikolova NI, Sobel CN, Hebert DNP, Adamowicz SJ. Spiders (Araneae) of Churchill, Manitoba: DNA barcodes and morphology reveal high species diversity and new Canadian records. BMC Ecol. 2013; 13: 44. doi: 10.1186/1472-6785-13-44 PMID: 24279427

48. Paquin P, Hedin M. The power and perils of "molecular taxonomy": a case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. Mol Ecol. 2004; 13: 3239–3255. PMID: 15367136

49. Prendini L. Comment on "Identifying spiders through DNA barcodes." Can J Zool. 2005; 83: 498–504.

50. Stoeckle BC, Dworschak K, Gossner MM, Kuehn R. Influence of arthropod sampling solution on insect genotyping reliability. Entomol Exp Appl. 2010; 135: 217–223.

51. Höfer H, Astrin J, Holstein J, Spelda J, Meyer F, Zarte N. Propylene glycol–a useful capture preservative for spiders for DNA barcoding. Arachnol Mitteilungen. 2015; 50: 30–36.

52. Hebert PDN, Gregory TR. The promise of DNA barcoding for taxonomy. Syst Biol. 2005; 54: 852–859. doi: 10.1080/10635150500354886 PMID: 16243770

53. Croucher PJP, Oxford GS, Searle J. Mitochondrial differentiation, introgression and phylogeny of species in the *Tegenaria atrica* group (Araneae: Agelenidae). Biol J Linn Soc. 2004; 81: 79–89.

54. Buhay J. "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. J Crustac Biol. 2009; 29: 96–110.

55. Baldo L, Ayoub N, Hayashi C, Russell J, Stahlhut J, Werren J. Insight into the routes of *Wolbachia* invasion: high levels of horizontal transfer in the spider genus *Agelenopsis* revealed by *Wolbachia* strain and mitochondrial DNA diversity. Mol Ecol. 2008; 17: 557–569. doi: 10.1111/j.1365-294X.2007.03608.x PMID: 18179432

56. Smith M, Bertrand C, Crosby K, Eveleigh E, Fernandez-Triana J, Fisher B, et al. *Wolbachia* and DNA barcoding insects: patterns, potential, and problems. PLoS One. 2012; 7: e36514. doi: 10.1371/journal.pone.0036514 PMID: 22567162

57. Greenstone MH, Rowley DL, Heimbach U, Lundgren JG, Pfannenstiel RS, Rehner SA. Barcoding generalist predators by polymerase chain reaction: carabids and spiders. Mol Ecol. 2005; 14: 3247–3266. PMID: 16101789

58. Robinson EA, Blagoev GA, Hebert PDN, Adamowicz SJ. Prospects for using DNA barcoding to identify spiders in species-rich genera. In: Stoev P., Dunlop J., Lazarov S, editor. A life caught in a spider's web Papers in arachnology in honour of Christo Deltshev. Zookeys 16; 2009. pp. 27–46.

59. Huber BA, Fischer N, Astrin JJ. High level of endemism in Haiti's last remaining forests: a revision of *Modisimus* (Araneae: Pholcidae) on Hispaniola, using morphology and molecules. Zool J Linn Soc. 2010; 158: 244–299.

60. Slowik J, Blagoev GA. First description of the male spider *Pacifiphantes magnificus* (Chamberlin & Ivie) (Araneae: Linyphiidae). Zootaxa. 2012; 73–81.

61. Blagoev GA, de Waard JR, Ratnasingham S, de Waard SL, Lu L, Robertson J, et al. Untangling taxonomy: A DNA barcode reference library for Canadian spiders. Mol Ecol Resour. 2016; 16: 325–341. doi: 10.1111/1755-0998.12444 PMID: 26175299

62. Castalanelli MA, Teale R, Rix MG, Kennington WJ, Harvey MS. Barcoding of mygalomorph spiders (Araneae: Mygalomorphae) in the Pilbara bioregion of Western Australia reveals a highly diverse biota. Invertebr Syst. 2014; 28: 375–385.

63. Planas E, Ribera C. Description of six new species of *Loxosceles* (Araneae: Sicariidae) endemic to the Canary Islands and the utility of DNA barcoding for their fast and accurate identification. Zool J Linn Soc. 2015; 174: 47–73.

64. Xu X, Liu FX, Chen J, Li DQ, Kuntner M. Integrative taxonomy of the primitively segmented spider genus *Ganthela* (Araneae: Mesothelae: Liphistiidae): DNA barcoding gap agrees with morphology. Zool J Linn Soc. 2015; 175: 288–306.

65. Starrett J, Derkarabetian S, Richart CH, Cabrero A, Hedin M. A new monster from southwest Oregon forests: *Cryptomaster behemoth* sp. n. (Opiliones, Laniatores, Travunioidea). Zookeys. 2016; 11–35.

66. Coddington J, Agnarsson I, Cheng R-C, Čandek K, Driskell A, Frick H, et al. DNA barcode data accurately assign higher spider taxa. PeerJ. 2016; 4: e1633v1631.

67. Geiger MF, Astrin JJ, Borsch T, Burkhardt U, Grobe P, Hand R, et al. How to tackle the molecular species inventory for an industrialized nation–lessons from the first phase of the German Barcode of Life initiative GBOL (2012–2015) (2012–2015). Genome. 2016; in press. doi: 10.1139/gen-2015-0185

68. Miller JA, Beentjes KK, Van Helsdingen P, Ijland S. Which specimens from a museum collection will yield DNA barcodes? A time series study of spiders in alcohol. Zookeys. 2013; 365: 245–261. doi: 10.3897/zookeys.365.5787 PMID: 24453561

69. Ratnasingham S, Hebert PDN. BOLD: The Barcode of Life Data System. Available: http://www.barcodinglife.org. Mol Ecol Notes. 2007; 7: 355–364. PMID: 18784790

70. Porco D, Rougerie R, Deharveng L, Hebert P. Coupling non-destructive DNA extraction and voucher retrieval for small soft-bodied arthropods in a high-throughput context: the example of Collembola. Mol Ecol Resour. 2010; 10: 942–945. doi: 10.1111/j.1755-0998.2010.2839.x PMID: 21565103

71. Ivanova NV, Dewaard JR, Hebert PDN. An inexpensive, automation-friendly protocol for recovering high-quality DNA. Mol Ecol Notes. 2006; 6: 998–1002.

72. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics. 2010; 26: 1899–1900. doi: 10.1093/bioinformatics/btq224 PMID: 20427515

73. Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), Version 4.0b10. Sunderland, MA.: Sinauer Associates; 2002.

74. Templeton AR. Using phylogeographic analyses of gene trees to test species status and processes. Mol Ecol. 2001; 10: 779–791. PMID: 11298987

75. Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealogies. Mol Ecol. 2000; 9: 1657–1659. PMID: 11050560

76. Meier R, Shiyang K, Vaidya G, Ng PKL. DNA Barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. Syst Biol. 2006; 55: 715–728. PMID: 17060194

77. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4: 406–425. PMID: 3447015

78. Felsenstein J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool. 1973; 22: 240–249.

79. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005; 21: 456–463. PMID: 15608047

80. Keane T, Creevey C, Pentony M, Naughton T, Mclnerney J. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol. 2006; 6: 29. PMID: 16563161

81. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol. 1984; 20: 86–93. PMID: 6429346

82. Ayoub NA, Riechert SE, Small RL. Speciation history of the North American funnel web spiders, *Agelenopsis* (Araneae: Agelenidae): Phylogenetic inferences at the population-species interface. Mol Phylogenet Evol. 2005; 36: 42–57. PMID: 15904855

83. Wood HM, Griswold CE, Spicer GS. Phylogenetic relationships within an endemic group of Malagasy "assassin spiders" (Araneae, Archaeidae): ancestral character reconstruction, convergent evolution and biogeography. Mol Phylogenet Evol. 2007; 45: 612–619. PMID: 17869131

84. Boyer SL, Baker JM, Giribet G. Deep genetic divergences in *Aoraki denticulata* (Arachnida, Opiliones, Cyphophthalmi): a widespread "mite harvestman" defies DNA taxonomy. Mol Ecol. 2007; 16: 4999–5016. PMID: 17944852

85. Ros V, Breeuwer J. Spider mite (Acari: Tetranychidae) mitochondrial COI phylogeny reviewed: host plant relationships, phylogeography, reproductive parasites and barcoding. Exp Appl Acarol. 2007; 42: 239–262. PMID: 17712605

86. Huber BA, Astrin JJ. Increased sampling blurs morphological and molecular species limits: revision of the Hispaniolan endemic spider genus *Tainonia* (Araneae: Pholcidae). Invertebr Syst. 2009; 23: 281–300.

87. Harms D, Framenau VW. New species of Mouse Spiders (Araneae: Mygalomorphae: Actinopodidae: *Missulena*) from the Pilbara region, Western Australia. Zootaxa. 2013; 3637: 521–540. PMID: 26046218

88. Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH. Integrative Taxonomy: A multisource approach to exploring biodiversity. Annu Rev Entomol. 2010; 55: 421–438. doi: 10.1146/annurev-ento-112408-085432 PMID: 19737081

89. Lasut L. Testing the suitability of DNA barcoding in spiders (Araneae). Ph.D. Thesis, Universität Bern. 2015.

90. Töpfer-Hofmann G, Cordes D, von Helversen O. Cryptic species and behavioural isolation in the *Pardosa lugubris* group (Araneae, Lycosidae), with description of two new species. Bull Br Arachnol Soc. 2000; 11: 257–274.

91. Harvey P. *Pardosa lugubris* sensu stricto in Britain. Newsl Br Arachnol Soc. 2004; 101: 8–9.

92. Kronestedt T. Separation of two species standing as *Alopecosa aculeata* (Clerck) by morphological behavioural and ecological characters with remarks on related species in the *pulverulenta* group (Araneae: Lycosidae). Zool Scr. 1990; 19: 203–225.

93. Chiarle A, Kronestedt T, Isaia M. Courtship behavior in European species of the genus *Pardosa* (Araneae, Lycosidae). J Arachnol. 2013; 41: 108–125.

94. Masta SE, Maddison WP. Sexual selection driving diversification in jumping spiders. Proc Natl Acad Sci USA. 2002; 99: 4442–4447. doi: 10.1073/pnas.072493099 PMID: 11930004

95. Jantscher E. Diagnostic characters of *Xysticus cristatus*, *X. audax* and *X. macedonicus* (Araneae: Thomisidae). Bull Br Arachnol Soc. 2001; 12: 17–25.

96. Hippa H, Oksala I. Definition and revision of the *Enoplognatha ovata* (Clerck) group (Araneae: Theridiidae). Entomol Scand. 1982; 13: 213–222.

97. Wiehle H. Spinnentiere oder Arachnoidea. VIII. Theridiidae oder Haubennetzspinnen (Kugelspinnen). Tierwelt Deutschlands. 1937; 33: 119–222.

98. Denis J. Sur quelques *Theridion* appartenants á la faune de France. Bull la Société Entomol Fr. 1944; 49: 111–117.

99. Roberts MJ. Spinnengids. Netherlands: Tirion Natuur Baarn; 1998.

100. Maelfait J-P, de Keer R, De Meester L. Genetical background of the polymorphism of *Oedothorax gibbosus* (Blackwall) (Linyphiidae Araneae). Rev Arachnol. 1990; 9: 29–34.

101. Jocqué R. Genitalic polymorphism—a challenge for taxonomy. J Arachnol. 2002; 30: 298–306.

102. Croucher PJP, Jones RM, Searle JB, Oxford GS. Contrasting patterns of hybridization in large house spiders (*Tegenaria atrica* group, Agelenidae). Evolution (N Y). 2007; 61: 1622–1640.

103. Lattimore VL, Vink CJ, Paterson AM, Cruickshank RH. Unidirectional introgression within the genus *Dolomedes* (Araneae: Pisauridae) in southern New Zealand. Invertebr Syst. 2011; 25: 70–79.

104. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc Natl Acad Sci U S A. 2004; 101: 14812–14817. doi: 10.1073/pnas.0406166101 PMID: 15465915

105. Decaens T, Rougerie R. Descriptions of two new species of Hemileucinae (Lepidoptera: Saturniidae) from the region of Muzo in Colombia—evidence from morphology and DNA barcodes. Zootaxa. 2008; 34–52.

106. Vaglia T, Haxaire J, Kitching IJ, Meusnier I, Rougerie R. Morphology and DNA barcoding reveal three cryptic species within the *Xylophanes neoptolemus* and *loelia* species-groups (Lepidoptera: Sphingidae). Zootaxa. 2008; 18–36.

107. Astrin JJ, Stüben PE. Molecular phylogeny in "nano-weevils": description of a new subgenus Nanoacalles and two new species of *Calacalles* from the Macaronesian Islands (Curculionidae: Cryptorhynchidae). Zootaxa. 2009; 2300: 51–67.

108. Pauls SU, Blahnik RJ, Zhou X, Wardwell CT, Holzenthal RW. DNA barcode data confirm new species and reveal cryptic diversity in Chilean Smicridea (Smicridea) (Trichoptera: Hydropsychidae). J North Am Benthol Soc. 2010; 29: 1058–1074. doi: 10.1899/09-108.1

109. Clouse RM, Wheeler WC. Descriptions of two new, cryptic species of *Metasiro* (Arachnida: Opiliones: Cyphophthalmi: Neogoveidae) from South Carolina, USA, including a discussion of mitochondrial mutation rates. Zootaxa. 2014; 3814: 177–201. doi: 10.11646/zootaxa.3814.2.2 PMID: 24943422

110. Zabka M. Salticidae: Jumping spiders (Arachnida, Araneae). Fauna Pol. 1997; 3–183.

111. Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci U S A. 2008; 105: 13486–13491. doi: 10.1073/pnas.0803076105 PMID: 18757756

112. Vogels J. First record of the salticid spider *Sibianor larae* (Araneae: Salticidae) in The Netherlands. Entomol Ber. 2012; 72: 254–258.

113. Huber BA. Revision and cladistic analysis of *Pholcus* and closely related taxa (Araneae, Pholcidae). Bonn Zool Monogr. 2011; 58: 1–509.

Decay of Sexual Trait Genes in an Asexual Parasitoid Wasp

Ken Kraaijeveld[1,2,*], Seyed Yahya Anvar[2], Jeroen Frank[2], Arnoud Schmitz[2], Jens Bast[3], Jeanne Wilbrandt[4], Malte Petersen[4], Tanja Ziesmann[4], Oliver Niehuis[4], Peter de Knijff[5], Johan T. den Dunnen[2], and Jacintha Ellers[1]

1Animal Ecology, Department of Ecological Sciences, VU University Amsterdam, The Netherlands

2Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

3Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

4Zoological Research Museum Alexander Koenig, Center for Molecular Biodiversity Research, Bonn, Germany

5Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

*Corresponding author: E-mail: ken@kenkraaijeveld.nl.

Accepted: November 10, 2016

Data deposition: The data for this project can be browsed and downloaded at http://parasitoids.labs.vu.nl/. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession JUFY00000000. The version described in this paper is version JUFY01000000. Fastq files containing the Illumina GAIIx sequence reads for the sexual lineage are available at Dryad: doi: 10.5061/dryad.k3bh40bg. The transcriptome assembly has been released as part of the 1KITE project: http://www.ncbi.nlm.nih.gov/bioproject/219570.

Abstract

Trait loss is a widespread phenomenon with pervasive consequences for a species' evolutionary potential. The genetic changes underlying trait loss have only been clarified in a small number of cases. None of these studies can identify whether the loss of the trait under study was a result of neutral mutation accumulation or negative selection. This distinction is relatively clear-cut in the loss of sexual traits in asexual organisms. Male-specific sexual traits are not expressed and can only decay through neutral mutations, whereas female-specific traits are expressed and subject to negative selection. We present the genome of an asexual parasitoid wasp and compare it to that of a sexual lineage of the same species. We identify a short-list of 16 genes for which the asexual lineage carries deleterious SNP or indel variants, whereas the sexual lineage does not. Using tissue-specific expression data from other insects, we show that fifteen of these are expressed in male-specific reproductive tissues. Only one deleterious variant was found that is expressed in the female-specific spermathecae, a trait that is heavily degraded and thought to be under negative selection in *L. clavipes*. Although the phenotypic decay of male-specific sexual traits in asexuals is generally slow compared with the decay of female-specific sexual traits, we show that male-specific traits do indeed accumulate deleterious mutations as expected by theory. Our results provide an excellent starting point for detailed study of the genomics of neutral and selected trait decay.

Key words: *Leptopilina clavipes*, *Wolbachia*, parthenogenesis, deleterious variants, sexual trait decay.

Introduction

When selective pressures shift, traits may become redundant. Such redundant traits tend to degenerate over time and may eventually be lost entirely. Trait loss is widespread, both phylogenetically and in terms of trait types, and has important evolutionary consequences. For example, when a trait is lost because its function is compensated by an ecological interaction, the species may become dependent on the ecological partner (Ellers et al. 2012). Another common pattern of trait loss is seen when sexually reproducing organisms switch to asexual reproduction. Such lineages quickly lose their ability to attract mates and fertilize eggs, effectively blocking a reversal to sexual reproduction (van der Kooi and Schwander 2014).

The molecular causes of trait loss are diverse. First, trait loss may result from pseudogenization of key genes through deleterious amino acid changes or mutations that disrupt gene function. Examples of trait loss caused by such loss-of-function mutations are the loss of vitamin C synthesis in several groups of mammals (Cui et al. 2011; Drouin et al. 2011; Hiller et al. 2012), loss of taste receptor genes in whales (Feng et al. 2014) and loss of a phospholipid transporter in horses and guinea pigs (Hiller et al. 2012). Second, mutations in regulatory sequences may alter the expression of genes underlying the trait.

For example, the loss of pelvic spines in the three-spined stickleback *Gasterosteus aculeatus* is caused by deletion of a tissue-specific enhancer of the *Pitx1* gene (Chan et al. 2010). Comparable deletions of regulatory elements are responsible for the loss of penile spines and forebrain growth arrest in humans (McLean et al. 2011). Last, redundant genes may be lost from a genome completely. Ortholog losses appear to be widespread (Wyder et al. 2007; Suen et al. 2011), although the true absence of a (pseudo)gene is difficult to prove. For example, bird genomes appear to have lost several genes involved in insulin sensitivity, without leaving them as detectable pseudogenes (Dakovic et al. 2014).

Trait integrity may be selectively neutral or under negative selection. This distinction is often difficult to make in real systems, but it is relatively clear-cut in the loss of sexual traits in asexual organisms. Upon the switch from sexual to asexual reproduction, redundant female-specific sexual traits tend to decay rapidly and consistently, suggestive of negative selection (van der Kooi and Schwander 2014). Redundant male-specific traits, on the other hand, are not expressed in asexual females, are consequently not exposed to selection and tend to remain functional for extended lengths of time (van der Kooi and Schwander 2014). Asexual organisms thus provide excellent models to study the dynamics of selected vs. neutral trait decay. An important challenge is to identify the genetic changes underlying the decay of sexual traits in asexuals. Mutations resulting in the decay in female-specific sexual traits may enhance fitness of asexual females and thus have a high chance of getting fixed in the population. In contrast, mutations affecting neutral male-specific traits would only become fixed through genetic drift. As a result, mutations affecting female-specific traits may be more prevalent than mutations affecting male-specific traits in asexual lineages. The parasitoid wasp *Leptopilina clavipes* provides a promising study species in which to address this issue. *L. clavipes* features both sexual and asexual reproducing lineages and its asexual lineages have decayed female-specific as well as male-specific traits (Pannebakker et al. 2005; Kraaijeveld et al. 2009).

Here, we present a draft genome assembly of an asexual lineage of the parasitoid wasp *Leptopilina clavipes.* We aligned whole-genome shotgun sequences of a sexual lineage of the same species to this draft genome. Using this alignment, we compare the genetic load of the sexual and asexual lineages. Tissue-specific expression patterns of homologous genes in *Nasonia vitripennis* and *Drosophila melanogaster* were used to identify candidate genes underlying the observed decay of sexual traits in *L. clavipes*. Given this information, we address the question of whether negative selection on female-specific sexual traits results in fixation of a larger number of deleterious variants in the underlying genes than found in genes encoding selectively neutral male-specific sexual traits. We investigated single-nucleotide polymorphism and insertion–deletion (indel) variants and identified variants likely to decrease the function of a given gene product. For a small set of candidate loci, we additionally examined whether independently evolved asexual lineages of *L. clavipes* have accumulated identical or comparable trait-loss mutations. This represents the first genome-wide assessment of sexual trait decay in an asexual organism.

Material and Methods

Study System

We sequenced the genome of the haplodiploid wasp *Leptopilina clavipes* (Hymenoptera: Figitidae), a parasitoid of *Drosophila* larvae. Asexual reproduction in this species is caused by *Wolbachia* endosymbionts that induce diploidy through gamete duplication (Pannebakker et al. 2004b). This meiotic alteration results in completely homozygous *L. clavipes* offspring (Kraaijeveld et al. 2011). *L. clavipes* occurs in both haplodiploid sexual (arrhenotokous) and asexual (thelytokous) populations, which are geographically separated. Northern European populations of this species have diverged from a Spanish population about 12,000–43,000 generations ago (this species has one or two generations a year in Northern Europe) and have become infected with a parthenogenesis-inducing *Wolbachia* during this period (Kraaijeveld et al. 2011). *Wolbachia* has infected multiple female lineages and the northern populations of *L. clavipes* consequently comprises a series of genetically distinct clones (Kraaijeveld et al. 2011).

Isofemale lineages of *L. clavipes* were maintained at Leiden University (The Netherlands) as described previously (Kraaijeveld et al. 2009). Three females were used to initiate each subsequent generation for at least 65 generations, thus likely resulting in high inbreeding levels in these isofemale lineages. We chose one asexual lineage (GBW) for whole genome shotgun sequencing and genome assembly. For comparison, we also obtained whole-genome shotgun sequences for one sexual lineage (EPG), which were aligned to the draft reference assembly [see Kraaijeveld et al. (2011) for collection details].

Genome Size Estimation

Flow cytometric genome size estimation was done with an Accuri C6 system following a standard protocol (Hare and Johnston 2011). *D. melanogaster* (estimated genome size 175 Mb; Animal Genome Size Database; http://www.genomesize.com; last accessed November 15, 2016) was used as reference for co-staining. Heads were removed from frozen animals (−80 °C), transferred into Galbraith buffer and ground using a Dounce tissue ginder. Both *L. clavipes* and *D. melanogaster* samples were filtered through a 20 μm nylon mesh and stained with propidium iodide (50 μg/ml) by incubating for 2 h at 4 °C. To compare 2C (and 4C) peak fluorescence signals, samples were run both separately and combined. All flow cytometry estimates are based on minimum counts of at least 1,000 nuclei each (i.e., 2C peaks).

In addition to our flow cytometry estimate, we estimated genome size from the sequence data (see below for details). Scaffolds containing sequences matching the putatively single-copy genes Ef-1a and RNApolII were identified using blast (Altschul et al. 1990). Both scaffolds had a fairly even coverage by HiSeq data of 87×. Genome size can then be estimated as (number of reads * average read length)/87. Furthermore, kmer-based methods provide an alternative method for estimating genome size (Liu et al. 2013). We employed two such methods: SGA (Simpson 2014) and KmerGenie (Chikhi and Medvedev 2014).

Sequencing

DNA was extracted from pools of ten *L. clavipes* females for Illumina sequencing and 30 females for Pacific Biosciences SMRT sequencing using the DNAeasy Blood and Tissue Kit (Qiagen, Valencia, CA) according to the manufacturer's protocol.

All next-generation sequencing was performed at the Leiden Genome Technology Center (LGTC) at the Leiden University Medical Center (The Netherlands). The GBW and EPG lineages were first sequenced on Illumina GAIIx as described by (Kraaijeveld et al. 2012). To obtain a high-quality reference genome, the GBW lineage was additionally sequenced on Illumina HiSeq 2000 and Pacific Biosciences (see supplementary table S1, Supplementary Material online for details on output).

For Illumina sequencing, genomic DNA was sonicated using the Covaris Instrument (Covaris Inc., USA). Paired-end libraries were prepared following Illumina's protocol (Illumina DNA sample kit). Briefly, fragments were end-repaired, 3′-adenylated, and ligated with Illumina adapters. Ligation products of 600–700 bp were gel-purified and PCR-amplified using Illumina adapter-specific primers. Libraries were purified and quantified using a Qubit Fluorometer (Thermo Fisher, USA) and evaluated using an Agilent 2100 bioanalyzer (Agilent Technologies, USA). GBW and EPG libraries were sequenced using 75-bp paired-end read chemistry on an Illumina GAIIx (Illumina, USA). The subsequent GBW library was sequenced using 100-bp paired-end read chemistry on Illumina HiSeq 2000 (Illumina, USA).

For Pacific Biosciences SMRT sequencing of the asexual GBW lineage, SMRTbell DNA template libraries were prepared according to the manufacturer's specification after the fragmentation with G-tubes (Covaris, USA). SMRTbell template libraries of different insert sizes (1.5, 4, 6.4, and 7 kb) were prepared. The fragmented DNA was end-repaired and ligated to hairpin adapters. SMRT sequencing was carried out on the Pacific Biosciences RS according to standard protocols, 16 SMRT cells with the C1 chemistry (diffusion loading, 2 × 45 min, 1 kb fragment size) and four SMRT cells with XL-P4 chemistry (Magbead loading, 1 × 120 min, 1 kb fragment size). All runs were processed using the standard primary data analysis.

Genome Assembly

The Illumina HiSeq (HiSeq) and Pacific Biosciences RS I (PacBio) data were used to assemble the genome of the asexual GBW lineage. First, filtered PacBio subreads >500 bp with a read quality >0.80 were error corrected using the PacBioToCA pipeline available in Celera Assembler 7.0 (Myers et al. 2000) (parameters merSize = 14, utgErroRate = 0.25, utgErrorLimit = 4.5, cnsErrorRate = 0.25, cgwErrorRate = 0.25, ovlErrorRate = 0.25, doOverlapBasedtrimmin = 0). This procedure maps the short, high-quality Illumina HiSeq reads to the long, low-quality PacBio reads and determines the consensus sequence. From the raw PacBio data, read correction removed 24.6% of reads and 35.6% of bases and shortened the average read length by 14.6%. The error-corrected PacBio reads and the HiSeq reads were used for hybrid *de novo* assembly using the Celera Assembler 7.0 (parameters merSize = 14, unitigger = bogart, toggleNumInstances = 0, cgwDemoteRBP = 0).

As a first validation of the *de novo* assembly, we re-mapped the HiSeq reads that were used in the de novo assembly to the final assembly using Bowtie2 (Langmead and Salzberg 2012) (parameters –N 1, –mp 4).

To assess the completeness of the assembled gene space, we mapped a set of Core Eukaryotic Genes (CEGs) to the assembly using the Core Eukaryotic Gene-Mapping Approach (CEGMA) pipeline (Parra et al. 2007, 2009). CEGs are highly conserved and thought to be present in every genome of a multicellular eukaryote in low copy numbers (Parra et al. 2009). Therefore, the percentage of CEGs that are present in a given sequenced genome can be taken as an estimator for the completeness of the sequenced gene space. Furthermore, we compared the gene space of the draft assembly to that of the parasitoid wasp *N. vitripennis* (genome build nvit_2.1) using blastp at an e-value cut-off of 1e-5.

To characterize any co-sequenced symbionts, parasites and contaminants, we employed the Blobology pipeline (Kumar et al. 2013). Briefly, all scaffolds were compared with a local install of NCBI's nt database using BLASTn (megaBLAST, *e*-value cut-off = 1e−5). We aligned Illumina GAIIx reads from the sexual lineage and the asexual lineage [described in Kraaijeveld et al. (2012)] to the reference assembly using Bowtie2 (Langmead and Salzberg 2012) with parameters –N 1 –mp 4. Duplicate reads were removed using Picardtools (http://broadinstittute.github.io/picard; last accessed November 15, 2016) and indels were realigned using GATK (McKenna et al. 2010). The bam files from these two alignments were used to calculate coverage for each scaffold. These were then plotted against the GC content of the scaffolds. Scaffolds and parts of misassembled scaffolds matching prokaryotic endosymbionts were removed from the final assembly.

Annotation

Protein-coding genes in the genome of *L. clavipes* were automatically annotated using MAKER2 version 2.31.6 (Holt and Yandell 2011). MAKER2 is an annotation pipeline that uses a combination of *ab initio* and evidence-based approaches to infer gene models with high confidence. We applied a two-pass, iterative workflow that aims to maximize the number of true positives in both gene predictions and annotations. The following information was used as input for the first MAKER2 run: transcriptome data (74,639 transcript sequences) generated as part of the 1KITE project (http://www.1kite.org/; last accessed November 15, 2016); Uniprot reference proteomes for *Apis mellifera* and *Atta cephalotes* (17.04.2014, without isoforms); gene predictions generated using the tools CEGMA (version 2.4; Parra et al. 2007), GeneMark-ES (version 2.3c; Lomsadze et al. 2005) and SNAP (release 29.11.2013; Korf 2004), each with default settings; repeat libraries obtained from RepeatMasker (arthropods) and generated *de novo* using Recon, as implemented in RepeatModeler (version 1.0.7; http://www.repeatmasker.org/RepeatModeler.html; last accessed November 15, 2016); transposable element library provided by MAKER2. The results from the first MAKER2 run were used to train Augustus (version 3.0.1; Stanke and Waack 2003) and SNAP. MAKER2 was then run a second time using the same input files as in the first run, except that we used the improved Augustus and SNAP files.

Functional annotation was carried out using InterProScan 5.7.48 (Jones et al. 2014). We searched the proteins predicted in the *L. clavipes* genome in the following databases: TIGRFAM 13.0 (Haft et al. 2003), ProDom 2006.1 (Servant 2002), SMART 6.2 (Letunic et al. 2009), HAMAP 201311.27 (Pedruzzi et al. 2013), ProSitePatterns 20.97 (Sigrist et al. 2013), SuperFamily 1.75 (Wilson et al. 2007), PANTHER 9.0 (Mi et al. 2013), Gene3D 3.5.0 (Sillitoe et al. 2015), PIRSF 284 (Wu et al. 2004), Pfam-A 27.0 (Finn et al. 2015), ProSiteProfiles 20.97 (Sigrist et al. 2013), and Coils 2.2 (Lupas et al. 1991). For proteins with matches, we extracted the Gene Ontology (GO) terms. We used OrthoMCL-DB (Chen et al. 2006) to assess orthology of gene models. OrthoMCL conducts blastp (Altschul et al. 1990) searches of all proteins against themselves and against proteins in the OrthoMCL database (*e*-value cut-off: e−5, 50% match). Proteins with matches above the threshold are assigned to orthologous groups. The remaining proteins are then compared with each other to find putative paralogous pairs, which are then clustered into paralog groups.

Comparison of Coding Variants

To compare the genome of the asexual *L. clavipes* lineage to that of the sexual lineage, we generated a preliminary list of variants (SNPs and indels) in vcf format using samtools and bcftools from the aligments described above. The vcf file was then filtered for QUAL $\geq$ 20 (phred-scaled quality score for the variant call) and read depth $\geq$10. To limit the influence of sequencing or assembly artifacts, we removed all variants that were also present in the alignment of the HiSeq data of the asexual lineage.

Trait loss may result from disruptions at various places in the transcript, leading to loss-of-function variants. Disruptions may appear as premature stop codons, at splice-sites or as insertion/deletions (indels) that break the transcript's reading frame (Macarthur et al. 2012). We therefore annotated all variants using snpEff (Cingolani et al. 2012) and filtered the resulting list of candidate loss-of-function variants on highly repetitive sequences, variants affecting non-canonical splice sites and transcripts whose underlying gene model did not contain a start codon. We further removed candidates whose protein was predicted to be short ($<$100 amino acids), that showed no significant similarity to proteins of other hymenopteran insects (assessed via BLASTP search) or where such BLASTP hits were based on repetitive or transposase domains (manual curation). Variants found in the sexually reproducing lineage were considered to be potentially involved in trait loss in the asexual lineage if they removed a stop codon from or caused a frame shift in the reference sequence (of the asexual lineage). We further selected candidates in genes related to sexual functions. For this, we exploited the fact that tissue-specific gene expression is well conserved between insects (Baker et al. 2011), and selected only variants in genes for which the expression of *N. vitripennis* or *D. melanogaster* homologs was enriched in one of the tissues related to sexual functions. This expression enrichment was determined by identifying the top blastp hit among *N. vitripennis* and *D. melanogaster* genes in the Waspatlas (Davies and Tauber 2015) and Flyatlas (Chintapalli et al. 2007) databases, respectively. Expression data was available for testes in *N. vitripennis* and testes, accessory glands and spermathecae in *D. melanogaster.* We attempted to predict whether the variant carried by the sexual lineage would result in a more optimal protein than produced by the variant carried by the asexual lineage by investigating sequence conservation among hymenopteran insects, analogous to the SIFT analysis described below. This assumes that variations on conserved amino acid sequences will usually result in a sub-optimal protein.

In addition to loss-of-function mutations, non-synonymous base substitutions could result in suboptimal protein function. At a given residue, amino acids that optimize protein function should be favored by selection and thus show a higher degree of conservation among related species than amino acids that reduce protein function. To predict whether an amino acid substitution affects protein function, we generated a SIFT (Ng and Henikoff 2001) database for the *L. clavipes* reference genome. SIFT predicts whether an amino acid substitution is likely to be deleterious to protein function based on sequence homology and the physical properties of amino acids. SIFT uses multiple alignment information to calculate normalized probabilities for all possible substitutions. Positions with

normalized probabilities less than 0.05 are predicted to be non-tolerated (deleterious) and those greater than or equal to 0.05 are predicted to be tolerated. We then used SIFT 4G (http://sift4g.org; last accessed November 15, 2016) to annotate all single-nucleotide polymorphisms (SNPs) between the asexual and the sexual *L. clavipes* genomes. For the variants that were predicted to be non-tolerated in the asexual genome but not in the sexual genome, or vice versa, we searched the protein against the *N. vitripennis* and *D. melanogaster* genomes using blastp and determined tissue-specific expression enrichment as above.

For all non-synonymous amino acid differences between the asexual and the sexual genomes, we predicted whether either the asexual or the sexual variant would result in a more stable protein using MUpro (Cheng et al. 2006). MUpro uses machine learning to predict how a single-site amino acid mutation affects protein stability and achieves about 84% accuracy. A confidence score is calculated, taking values between −1 and 1. Negative values indicate a decrease in protein stability and positive values an increase in protein stability. Values closer to −1 or 1 have higher confidence than values closer to 0. Proteins that were predicted to be more stable in the sexual lineage versus the asexual lineage at high confidence were searched against the *N. vitripennis* and *D. melanogaster* genomes using BLASTP. Tissue-specific expression enrichment was then determined as above.

Downstream Analysis of Candidate Decayed Genes

To examine whether genetically different asexual lineages all carried the same putative trait-loss variants, we sequenced four variants (two in genes enriched in testes and two in genes enriched in accessory glands) identified from our SIFT analysis in twelve asexual and nine sexual lineages of *L. clavipes.* These lineages were selected from a larger set of lineages, because microsatellite analysis had previously identified them as between genetically different (Kraaijeveld et al. 2011).

Results

The *Leptopilina clavipes* Genome

The draft genome assembly of *L. clavipes* consists of 36,601 scaffold with a size larger 200 bp and spans 255 Mb. The largest scaffold had a size of 419,8 kb and N50 was 13,759. A summary of the assembly statistics is presented in supplementary tables S1 and S2, Supplementary Material online. Overall, 92.7% of Hiseq reads aligned to the genome assembly. 54.6% of read pairs aligned concordantly exactly once and 30.1% more than once. Of the 15.3% read pairs that did not align concordantly, 13.6% aligned discordantly once. Discordantly mapping reads were found on many (28,570) scaffolds and visual inspection showed most of these reads to be spread evenly within scaffolds. The read coverage was unimodal (supplementary fig. S1, Supplementary Material online).

Flow cytometry yielded a genome size estimate of 321 Mb for *L. clavipes* (supplementary fig. S2, Supplementary Material online). Our read-based method estimated genome size as 318 Mb, whereas the k-mer based methods SGA and KmerGenie yielded estimates of 293.8 Mb and 255.1 Mb, respectively. Based on these various estimates, the draft genome assembly represents 79.5–99.9% of the genome.

We found 230 (93%) of the 248 Core Eukaryotic Genes (CEGs) to be present and seemingly complete in the *L. clavipes* genome assembly. An additional 15 CEGs (6%) were found incomplete. These CEGs tend to occur as single copies in eukaryote genomes (Parra et al. 2009). The average number of orthologs identified for this set of CEGs in the *L. clavipes* genome assembly was 1.23 (1.38 when including incomplete CEGs), indicating that the level of redundancy was low. We found 90.1% of the predicted proteins of *N. vitripennis* to be represented in the *L. clavipes* genome assembly.

Most scaffolds exhibited local similarity (indicated by BLAST hits) to genomic sequences of eukaryotes (mostly Hymenoptera and other insects; fig. 1). A subset of 90 scaffolds was classified as Rickettsiales, and all but one of these matched various *Wolbachia* genomes. Most of these scaffolds ($n = 53$) had very low coverage ($<1\times$) in the sequenced sexual lineage (fig. 1), but above-average coverage ($>70\times$) in the asexual lineage (fig. 1), consistent with the absence of *Wolbachia* from the sexual lineage. A small number of scaffolds ($n = 37$) classified as Rickettsiales had coverage within the range of the scaffolds classified as insect in both the sexual and asexual lineage (fig. 1). In twelve of these scaffolds, the *Wolbachia* hit was flanked by hits to insect genomes, potentially indicative of horizontal transmission of *Wolbachia* DNA to the nucleus. However, closer inspection revealed that in 15 out of 37 cases, the region corresponding to the *Wolbachia* hit were not covered by reads from the sexual lineage, suggesting that these regions were not part of the sexual genome. Furthermore, these same regions showed above-average coverage by reads from the asexual lineage, suggesting that they were likely misassembled. The remaining regions were all short (<500 bp) and probably represented spurious hits to *Wolbachia.* In conclusion, we have no compelling evidence for horizontal transmission events from *Wolbachia* to the nuclear genome of *L. clavipes*. We also identified seven scaffolds and two partial (i.e., misassembled) scaffolds matching the WO phage of the wVitB *Wolbachia* of *N. vitripennis*. These sequences had $>200\times$ coverage in the asexual lineage, but no coverage in the sexual lineage. A further 18 scaffolds matched other bacteria and 220 scaffolds matched other viruses (mostly an Ichnovirus isolated from the wasp *Hyposoter didymator*) and had comparable coverage in the asexual and sexual lineage.

MAKER2 annotated a total of 49,568 genes, 50,004 transcripts, 186,194 exons and 15,426 untranslated regions

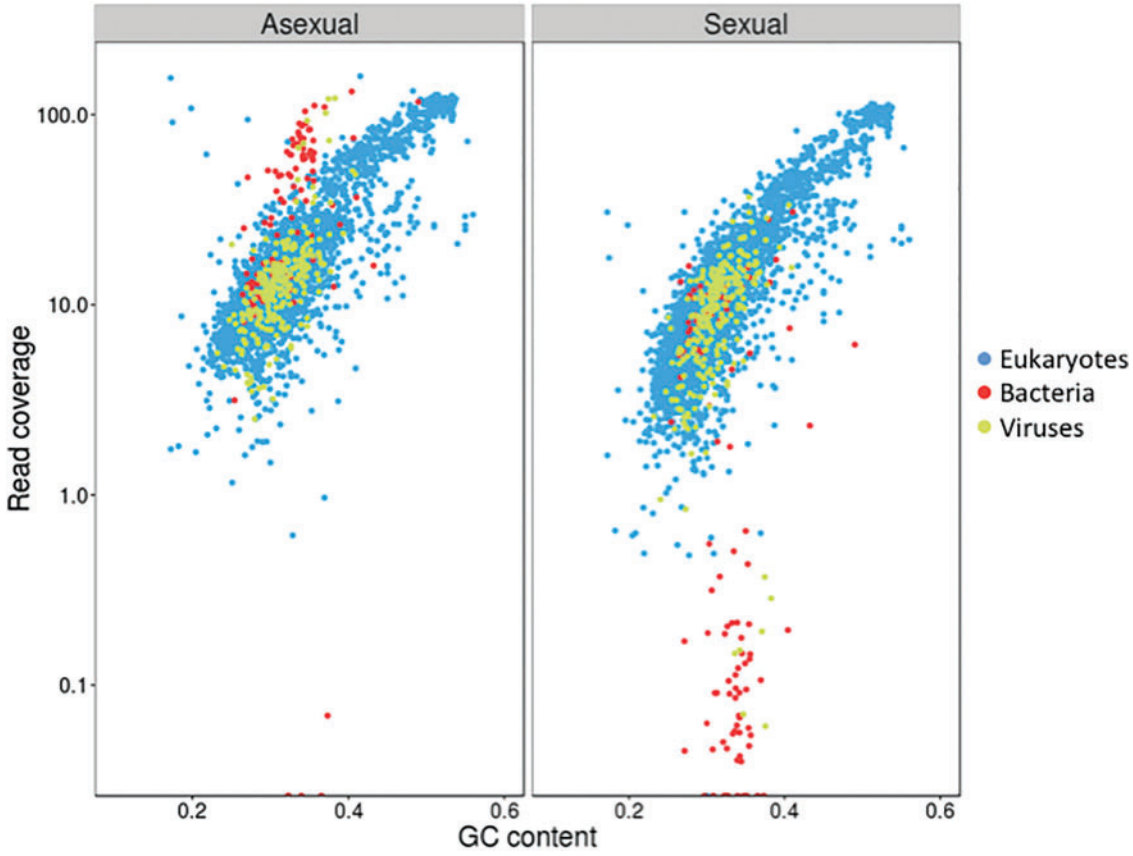

Fig. 1.—"Blobology" plots of read coverage against GC content per scaffold for *Wolbachia*-infected asexual lineage and uninfected sexual lineage. Dots are colored according the top hit from a BLAST search against the NCBI's nt database. Only scaffolds for which a significant BLAST hit was obtained are shown.

(UTRs). We found 16,562 predicted proteins that had at least one match with any of the protein databases (supplementary information, Supplementary Material online). A total of 8,243 orthologous groups were assigned to proteins in the *L. clavipes* genome. Furthermore, 1,571 groups of paralogous proteins were identified, each containing between 2 and 246 proteins.

Comparison of Coding Variants

Our initial list of possible loss-of-function variants comprised of 597 SNPs and 997 indels. After stringent filtering (see "Methods" section), we obtained a short-list of five genes that contained possible loss-of-function variants in the reference sequence and for which gene expression for putative homologs in *N. vitripennis* and *D. melanogaster* was biased to male reproductive tissue (table 1). We were not able to confirm bioinformatically whether variants carried by the sexual lineage would result in a more functional protein, because of a too low level of nucleotide sequence conservation among the investigated Hymenoptera insects.

We obtained SIFT scores for a total of 11,874 homozygous SNPs in protein-coding sequences (see fig. 2 for an example). Specifically, we found twelve variants for which the asexual genotype was deleterious, whereas the sexual genotype was not (table 1). The reverse was true for 671 variants, indicating that the sexual genome carried a heavier load of deleterious mutations compared with the asexual genome (Fisher exact test $P < 2.2 \times 10^{-16}$). We assessed the putative function of these genes affected by predicted deleterious variants in both the asexual or sexual lineage by identifying their homologs in *D. melanogaster* and determining the tissue in which the homologue was most expressed. The few deleterious variants identified using SIFT in the genome of the asexual lineage were found in genes expressed in testes, accessory glands, and spermathecae (fig. 3). While this distribution did not differ from random expectation (Fisher exact tests after FDR correction $P > 0.25$), it is noteworthy that these are all tissues whose functions are likely to be redundant in asexuals. We searched for homologs in the *N. vitripennis* genome and confirmed that the two genes for which *D. melanogaster* homologs were enriched in testes, showed the same pattern in

Table 1
Shortlist of Candidate Genes Involved in Sexual Traits Decay in Asexual *Leptopilina clavipes*

Mutation Type	Identified Using	Drosophila Homolog	Drosophila Tissue Enrichment	Nasonia Homolog	Nasonia Tissue Enrichment	Annotation	Notes
Enriched in reproductive tissue							
Loss-of-function	snpEff	NP_648446.1	Testis	XP_003425377.1	Female body	Pleckstrin homology-like domain family B member 1	Frame shift
Loss-of-function	snpEff	NP_001015401.1	Testis	XP_003426117.1	Testis	Tim17b	Stop codon removed
Loss-of-function	snpEff	NP_995777.1	Testis	XP_008217920.1	Testis	Ribonuclease H1	Frameshift
Loss-of-function	snpEff			XP_008216187.1	Testis	RNA-binding protein 4.1-like	Frameshift
Loss-of-function	snpEff	NP_610943.2	Testis	XP_008206136.1	Testis	Ubiquitin specific protease 20/33	Frameshift
Non-tolerated	SIFT	NP_788479.1	acc	XP_008207671.1	Testis	ergic53	validated
Non-tolerated	SIFT	NP_727442.1	spt	XP_008217640.1	Female body	Raspberry	
Non-tolerated	SIFT	NP_788565.1	acc	XP_001602982.1	Testis	Isoleucyl-tRNA synthetase	validated
Non-tolerated	SIFT	NP_611087.1	Tubule	XP_001606432.1	Testis	Cysteinyl-tRNA synthetase	
Non-tolerated	SIFT	NP_731238.1	Testis	XP_008205904.1	Testis	Dipeptidyl aminopeptidase III	Validated
Non-tolerated	SIFT	NP_608533.1	Testis	XP_003427673.2	Testis	Uncharacterized	Validated
Non-tolerated	SIFT	NP_649645.1	acc	XP_001607849.1	Testis	Small ribonucleoprotein particle protein SmD2	
Non-tolerated	SIFT	NP_477412.1	trachea	XP_001601436.1	Testis	nop5	
Non-tolerated	SIFT	NP_001261050.1		XP_008205733.1	Testis	Quaking related 54B	
Unstable protein	MU-pro	NP_611131.2	Fat body	XP_008208307.1	Testis	Uncharacterized	
Unstable protein	MU-pro	NP_611350.1	Tubule	XP_001067690.2	Testis	Autophagy-related 7	
Not enriched in reproductive tissue							
Unstable protein	MU-pro			XP_008204426.1	Female body	Uncharacterized	
Unstable protein	MU-pro						
Non-tolerated	SIFT	NP_611179.3		XP_008203900.1	Female body	Eps15 homology domain containing protein-binding protein 1	
Unstable protein	MU-pro	NP_611223.4	Trachea			anaphase promoting complex subunit 10	
Non-tolerated	SIFT	NP_725570.1	Fat body	XP_008208687.1	Female head	HMG coenzyme A synthase	
Non-tolerated	SIFT	NP_572695.2	Eye	XP_001604944.2	Female body	antdh	

N. vitripennis. We also searched for *N. vitripennis* homologs for two genes for which no flyatlas data was available. One of these genes was enriched in testes in *N. vitripennis*, adding an additional candidate trait-loss gene to our list (table 1). In contrast, genes containing deleterious variants in the sexual lineage were more often highly expressed in ovaries and less often in salivary glands than expected by chance (Fisher exact tests after FDR correction $P = 0.001$). This was not the case for genes expressed in testis (fig. 3). Ovarian genes are less likely to be expressed in males and deleterious mutations in these genes are therefore not purged in sexual haplodiploids.

MUpro analysis yielded comparable patterns as Sift analysis in the abundance and function of affected genes in the sexual and asexual lineage. Of the 9,579 non-synonymous differences found between the genomes of the sexual and the asexual lineages, MUpro predicted 379 differences to result in a less stable protein in the asexual lineage (1.3% predicted at > 0.8 confidence). Waspatlas data was available for three of the five genes predicted at high confidence to be less stable in the asexual lineage (table 1). Two of these were enriched in male reproductive tissue. Flyatlas data was also available for three of the five genes, but none was enriched in a tissue related to sexual function (table 1). In contrast, 9,200 (96%) were predicted to have resulted in a less stable protein in the sexual lineage (54.2% predicted at > 0.8 confidence). Again, the affected genes in the sexual lineage were biased towards those expressed in reproductive tissues (mainly ovaries; supplementary fig. S3, Supplementary Material online).

Downstream Analysis of Candidate Decayed Genes

Four of the putative trait-loss genes identified using SIFT (see above) were selected for further testing: in two of these, the *N. vitripennis* and/or *D. melanogaster* homologs were enriched in testes and in the other two, a homolog was enriched in accessory glands. We genotyped twelve asexual and nine sexual lineages of *L. clavipes* at these four loci. The genetically different asexual lineages did not carry the same putative trait-loss variants. Furthermore, the pattern of presence/absence of the variants across the 12 asexual lineages followed their phylogenetic relationships based on neutral microsatellite markers

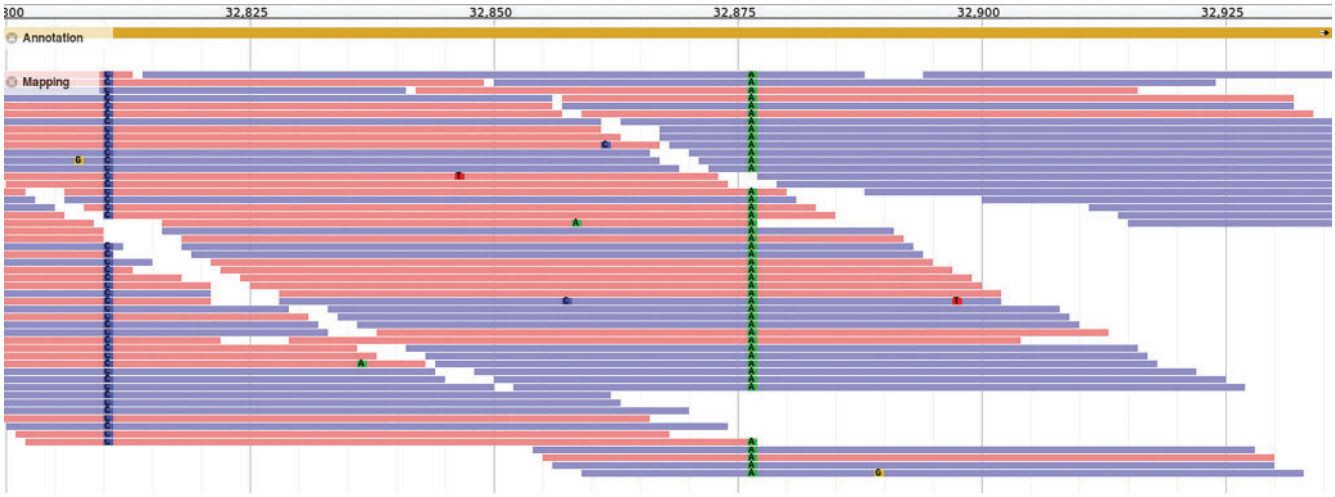

Fig. 2.—Alignment of reads from the sexual lineage against the reference genome of the asexual lineage, showing variants in a gene primarily expressed in testis. From parasitoids.labs.vu.nl.

(fig. 4), with more closely related lineages sharing more variants with the genome-sequenced lineage.

The occurrence of putative deleterious variants also differed between asexual and sexual lineages. Both of the putative trait loss variants in a gene enriched in the testes were unique to the asexual lineages (fig. 4). Of the variants in a gene enriched in the accessory glands, one also segregated among the sexual lineages, while the other was only found in the asexual lineages.

Discussion

We sequenced the genome of an asexual lineage of the parasitoid wasp *L. clavipes.* A small number of variants in coding regions were predicted to be deleterious in this asexual lineage, and these were concentrated in genes expressed in tissues related to redundant sexual functions. We identified a shortlist of deleterious variants in 16 genes that potentially contributed to the observed phenotypic decay of redundant sexual traits in this species. Subsequent analysis of four of these variants showed that not all asexual lineages carry the same deleterious variants.

The patterns of occurrence of deleterious variants in the genome of asexually reproducing *L. clavipes* are consistent with phenotypic patterns of trait decay observed in *L. clavipes*. Asexual lineages of this species have degenerated spermathecae (Kraaijeveld et al. 2009) and reduced male fertility (Pannebakker et al. 2005). The spermatheca-specific and testis-specific genes identified as carrying deleterious mutations thus represent candidates underlying these degenerated phenotypes. The genetic basis of reduced male fertility was previously mapped to a single QTL of large effect (Pannebakker et al. 2004a). Subsequent work should focus on the genomic location of the identified candidate genes, and test whether or not they overlap with the QTL region.

Our analysis of gene function is based on tissue-specific expression data of putative homologs in *N. vitripennis* and *D. melanogaster*. Tissue-specific expression data for *L. clavipes* is needed to confirm that our interpretations are correct. However, gene expression patterns tend to be conserved among insects (Baker et al. 2011). Tissue-specific expression data for *N. vitripennis* covers fewer tissues than that for *D. melanogaster*, but the patterns of enrichment match for most of our candidate genes (especially when assuming that accessory glands were co-extracted with the testes in *N. vitripennis*).

It is noteworthy that we identified 15 putatively deleterious variants in genes expressed mostly in male reproductive tissues, but only one in a redundant female-specific tissue (spermathecae). Spermathecae in asexual *L. clavipes* are heavily degraded and non-functional (Kraaijeveld et al. 2009). Males derived by curing asexual mothers from *Wolbachia* infection are still fertile—albeit to a reduced degree (Pannebakker et al. 2005). One possible explanation for this apparent discrepancy is that one or more genes crucial for spermathecal development may have been deleted mostly or entirely from the genome and we consequently were unable to detect them in our analysis. Although many genes are known to be upregulated or even specific to mature spermatheca in *Drosophila* (Prokupek et al. 2008; Schnakenberg et al. 2011), little is known about the genes involved in spermathecal development. The gene Hr39 was shown to be essential for normal spermathecal development in *Drosophila* (Allen and Spradling 2008) and a homolog of this gene is present in *L. clavipes*. Female-specific sexual function tends to degrade rapidly upon the switch to asexual reproduction (van der Kooi and Schwander 2014), which might indicate that female-specific trait decay is often caused by few mutations of large effect. Male-specific sexual functions, on the other hand, decay much more slowly (van der Kooi and Schwander 2014).

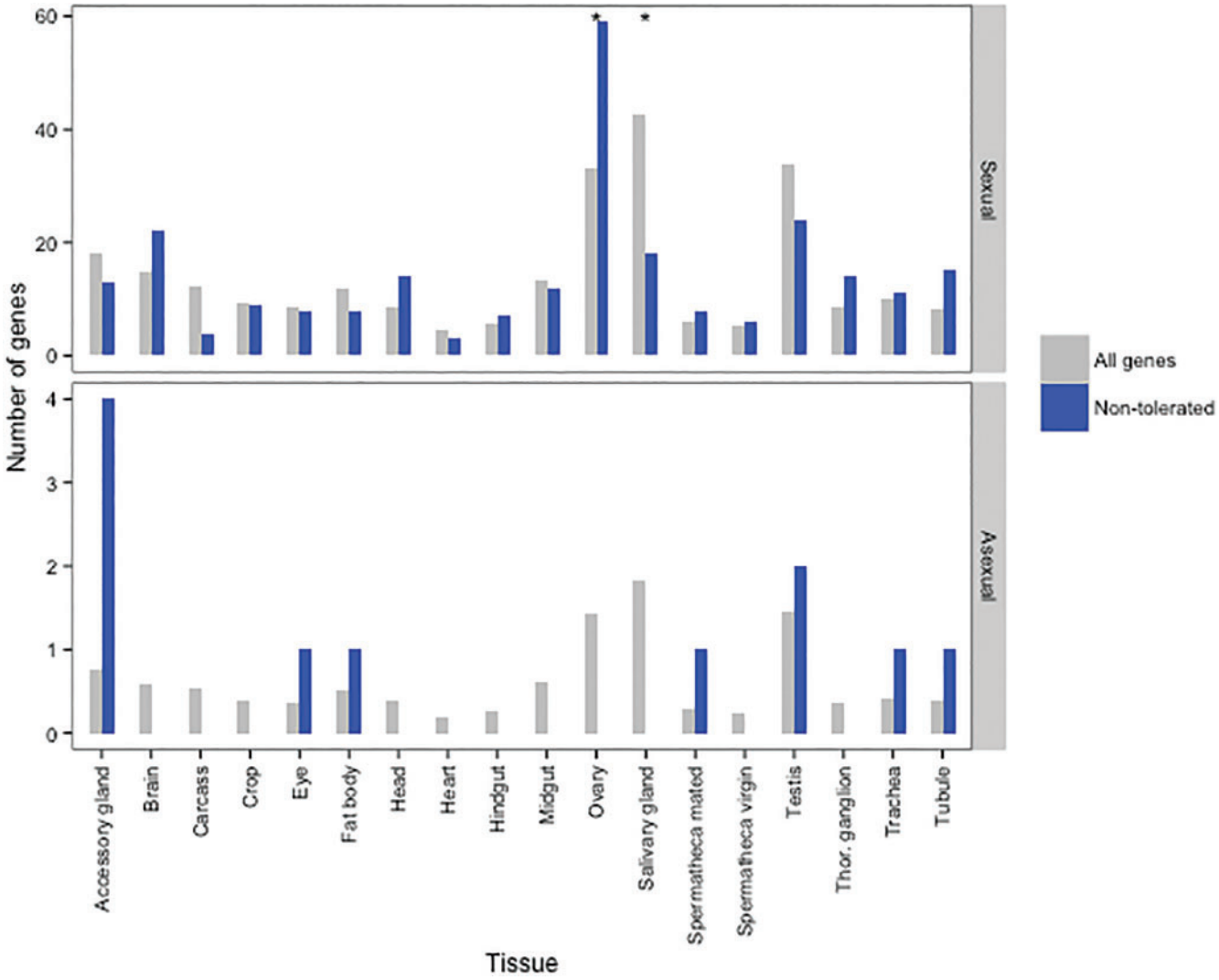

Fig. 3.—Deleterious variants in the *Leptopilina clavipes* genome are overrepresented in reproductive tissues. Deleterious (non-tolerated) variants were identified using SIFT and the orthologs of the genes in which they were found were searched for in the genome of *Drosophila melanogaster*. The tissue in which each of these orthologs show highest expression was identified in Flyatlas (Chintapalli et al. 2007) and is shown in blue for asexual and sexual *L. clavipes* lineages. The distribution of tissues with most abundant expression for all genes in Flyatlas is shown in grey. Significant Fisher exact P values following FDR correction are indicated with an asterisk.

Since we found several candidate variants that could contribute to the decay of male-specific sexual traits, our results suggest that sexual trait decay in *L. clavipes* males is the result of multiple mutations of small effect.

Our results suggest that the genome of a sexual *L. clavipes* lineage was more heavily loaded with deleterious variants than that of the asexual lineage. Deleterious variants in the sexual lineage were overrepresented in genes enriched in ovaries, which are probably only expressed in diploid females in which recessive alleles are partially shielded from selection. Our interpretation of the excess of deleterious variants is therefore that prolonged inbreeding exposed recessive deleterious variants that segregated in the ancestral sexual lineage. This interpretation would be consistent with inbreeding effects in other haplodiploid organisms (Brückner 1978; Henter 2003; Tortajada et al. 2009; Tien et al. 2015). Deleterious variants in female-specific tissues were not observed in the asexual lineage, suggesting that these alleles must have been purged by lineage selection during the transition from sexual to asexual reproduction.

We present the first genome-wide assessment of the genetic changes potentially underlying sexual trait decay in an asexual insect. Our results indicate that the genome of asexual *L. clavipes* was relatively free of deleterious variants and that damaging effects were concentrated in redundant sexual

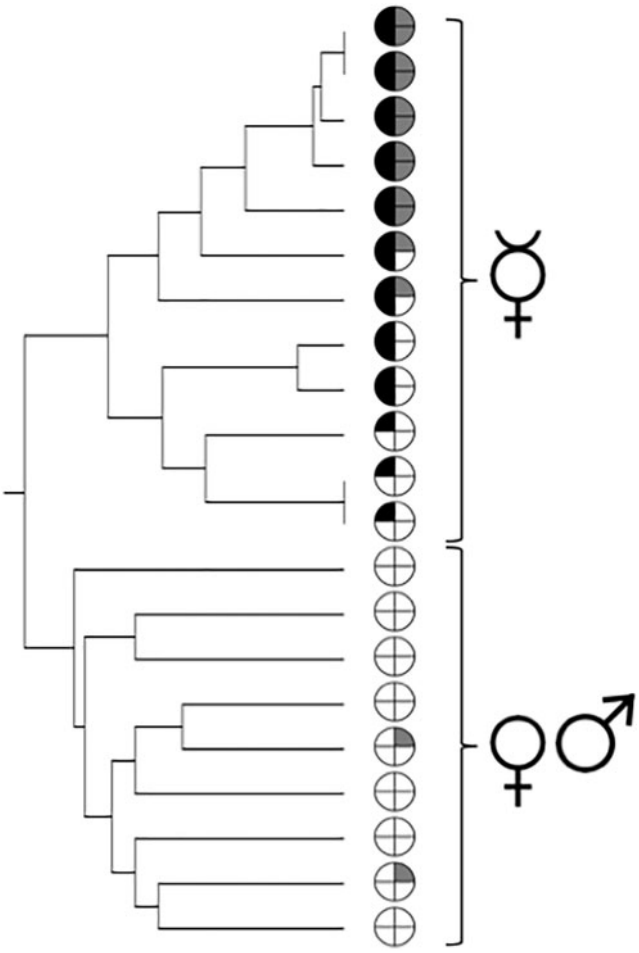

FIG. 4.—Phylogenetic relationship among twelve asexual and nine sexual lineages of *Leptopilina clavipes*, based on microsatellite markers; data from (Kraaijeveld et al. 2011). Pie charts indicate the presence (either black or dark grey)/absence (white) of putative trait-loss variants (left: two genes enriched in testes, right: two genes enriched in accessory glands).

genes. The list of candidate genes we identified will provide an excellent starting point for unraveling the genomics of trait decay in this and similar systems.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We acknowledge Yavuz Ariyurek, Henk Buermans, Emile de Meijer and Kristiaan van der Gaag for help with components of the sequencing work. We thank Pauline Ng and Swarnaseetha Adusumalli for help with the SIFT analysis. Peter Neleman helped with the bioinformatics. JW, ON, TZ, MP acknowledge Dr Alexander Donath for help installing software on the HPC cluster of the Zoological Research Museum Alexander Koenig in Bonn. The research by JW, ON, TZ, MP was supported by the Leibniz Graduate School "Genomic Biodiversity Research". JE and KK were supported by a Vici grant from the Netherlands Organization for Scientific Research. KK was supported by a Veni grant from the Netherlands Organization for Scientific Research.

Literature Cited

Allen AK, Spradling AC. 2008. The Sf1-related nuclear hormone receptor Hr39 regulates Drosophila female reproductive tract development and function. Development 135:311–321.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Baker DA, et al. 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*. BMC Genomics 12:296.

Brückner D. 1978. Why are there inbreeding effects in haplo-diploid systems? Evolution (N Y) 32:456–458.

Chan YF, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327:302–305.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34:D363–D368.

Cheng J, Randall A, Baldi P. 2006. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30:31–37.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet. 39:715–720.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6:1–13.

Cui J, Pan Y-H, Zhang Y, Jones G, Zhang S. 2011. Progressive pseudogenization: vitamin C synthesis and its loss in bats. Mol Biol Evol. 28:1025–1031.

Dakovic N, et al. 2014. The loss of adipokine genes in the chicken genome and implications for insulin metabolism. Mol Biol Evol. 10:2637–2646.

Davies NJ, Tauber E. 2015. WaspAtlas: a Nasonia vitripennis gene database and analysis platform. Database 2015:bav103.

Drouin G, Godin J-R, Pagé B. 2011. The genetics of vitamin C loss in vertebrates. Curr Genomics 12:371–378.

Ellers J, Kiers ET, Currie CR, McDonald BR, Visser B. 2012. Ecological interactions drive evolutionary loss of traits. Ecol Lett. 15:1071–1082.

Feng P, Zheng J, Rossiter SJ, Wang D, Zhao H. 2014. Massive losses of taste receptor genes in toothed and baleen whales. Genome Biol Evol. 6:1254–1265.

Finn RD, et al. 2015. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31:371–373.

Hare EE, Johnston JS. 2011. Genome size determination using flow cytometry of propidium iodide-stained nuclei. In: Orgogozo V, Rockman MV, editors. Molecular methods for evolutionary genetics.. Totowa (NJ): Humana Press. p. 3–12.

Henter HJ. 2003. Inbreeding depression and haplodiploidy: experimental measures in a parasitoid and comparisons across diploid and haplodiploid insect taxa. Evolution (N Y) 57:1793–1803.

Hiller M, et al. 2012. A 'forward genomics' approach links genotype to phenotype using independent phenotypic losses among related species. Cell Rep. 2:817–823.
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.
Jones P, et al. 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30:1236–1240.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
Kraaijeveld K, et al. 2012. Transposon proliferation in an asexual parasitoid. Mol Ecol. 21:3898–3906.
Kraaijeveld K, Franco P, De Knijff P, Stouthamer R, Van Alphen JJM. 2011. Clonal genetic variation in a Wolbachia-infected asexual wasp: horizontal transmission or historical sex?. Mol Ecol. 3644–3652.
Kraaijeveld K, Franco P, Reumer BM, van Alphen JJM. 2009. Effects of parthenogenesis and geographic isolation on female sexual traits in a parasitoid wasp. Evolution 63:3085–3096.
Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet. 4:237.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
Letunic I, Doerks T, Bork P. 2009. SMART 6: Recent updates and new developments. Nucleic Acids Res. 37:229–232.
Liu B, et al. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv 1308.2012.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494–6506.
Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. Science (80-.) 252:1162–1164.
Macarthur DG, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science (80-.) 205:823–828.
McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
McLean CY, et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471:216–219.
Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41:D377–D386.
Myers EW, et al. 2000. A whole-genome assembly of Drosophila. Science 287:2196–2204.
Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res. 11:863–874.
Pannebakker BA, et al. 2005. Sexual functionality of *Leptopilina clavipes* (Hymenoptera: Figitidae) after reversing Wolbachia-induced parthenogenesis. J Evol Biol. 18:1019–1028.
Pannebakker BA, Beukeboom LW, van Alphen JJM, Brakefield PM, Zwaan BJ. 2004a. The genetic basis of male fertility in relation to haplodiploid reproduction in *Leptopilina clavipes* (Hymenoptera: Figitidae). Genetics 168:341–349.
Pannebakker BA, Pijnacker LP, Zwaan BJ, Beukeboom LW. 2004b. Cytology of Wolbachia-induced parthenogenesis in *Leptopilina clavipes* (Hymenoptera: Figitidae). Genome 47:299–303.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.
Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37:289–297.
Pedruzzi I, et al. 2013. HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 41:584–589.
Prokupek A, et al. 2008. An evolutionary expressed sequence tag analysis of *Drosophila spermatheca* genes. Evolution 62:2936–2947.
Schnakenberg SL, Matias WR, Siegal ML. 2011. Sperm-storage defects and live birth in Drosophila females lacking spermathecal secretory cells. PLoS Biol. 9:e1001192.
Servant F. 2002. ProDom: automated clustering of homologous domains. Brief Bioinform. 3:246–251.
Sigrist CJA, et al. 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41:1–4.
Sillitoe I, et al. 2015. CATH: Comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43:D376–D381.
Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30:1228–1235.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:ii215–ii225.
Suen G, et al. 2011. The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet. 7:e1002007.
Tien NSH, Sabelis MW, Egas M. 2015. Inbreeding depression and purging in a haplodiploid: gender-related effects. Heredity (Edinb) 114: 327–332.
Tortajada AM, Carmona MJ, Serra M. 2009. Does haplodiploidy purge inbreeding depression in rotifer populations? PLoS One 4:e8195.
van der Kooi CJ, Schwander T. 2014. On the fate of sexual traits under asexuality. Biol Rev Camb Philos Soc. 89:805–819.
Wilson D, Madera M, Vogel C, Chothia C, Gough J. 2007. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res. 35:308–313.
Wu CH, et al. 2004. PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. 32:D112–D114.
Wyder S, Kriventseva EV, Schröder R, Kadowaki T, Zdobnov EM. 2007. Quantification of ortholog losses in insects and vertebrates. Genome Biol. 8:R242.

Associate editor: Sarah Schaack

Platyzoan Paraphyly Based on Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia

Torsten H. Struck,*[,1,2] Alexandra R. Wey-Fabrizius,[3] Anja Golombek,[1] Lars Hering,[4] Anne Weigert,[5] Christoph Bleidorn,[5] Sabrina Klebow,[3] Nataliia Iakovenko,[6,7] Bernhard Hausdorf,[8] Malte Petersen,[1] Patrick Kück,[1] Holger Herlyn,[9] and Thomas Hankeln[3]

[1]Zoological Research Museum Alexander Koenig, Bonn, Germany

[2]University of Osnabrück, FB05 Biology/Chemistry, AG Zoology, Osnabrück, Germany

[3]Institute of Molecular Genetics, Biosafety Research and Consulting, Johannes Gutenberg University, Mainz, Germany

[4]Animal Evolution and Development, Institute of Biology II, University of Leipzig, Leipzig, Germany

[5]Molecular Evolution and Systematics of Animals, Institute of Biology, University of Leipzig, Leipzig, Germany

[6]Department of Biology and Ecology, Ostravian University in Ostrava, Ostrava, Czech Republic

[7]Department of Invertebrate Fauna and Systematics, Schmalhausen Institute of Zoology NAS of Ukraine, Kyiv, Ukraine

[8]Zoological Museum, University of Hamburg, Hamburg, Germany

[9]Institute of Anthropology, Johannes Gutenberg University, Mainz, Germany

***Corresponding author:** E-mail: torsten.struck.zfmk@uni-bonn.de.

Associate editor: Gregory Wray

Abstract

Based on molecular data three major clades have been recognized within Bilateria: Deuterostomia, Ecdysozoa, and Spiralia. Within Spiralia, small-sized and simply organized animals such as flatworms, gastrotrichs, and gnathostomulids have recently been grouped together as Platyzoa. However, the representation of putative platyzoans was low in the respective molecular phylogenetic studies, in terms of both, taxon number and sequence data. Furthermore, increased substitution rates in platyzoan taxa raised the possibility that monophyletic Platyzoa represents an artifact due to long-branch attraction. In order to overcome such problems, we employed a phylogenomic approach, thereby substantially increasing 1) the number of sampled species within Platyzoa and 2) species-specific sequence coverage in data sets of up to 82,162 amino acid positions. Using established and new measures (long-branch score), we disentangled phylogenetic signal from misleading effects such as long-branch attraction. In doing so, our phylogenomic analyses did not recover a monophyletic origin of platyzoan taxa that, instead, appeared paraphyletic with respect to the other spiralians. Platyhelminthes and Gastrotricha formed a monophylum, which we name Rouphozoa. To the exclusion of Gnathifera, Rouphozoa and all other spiralians represent a monophyletic group, which we name Platytrochozoa. Platyzoan paraphyly suggests that the last common ancestor of Spiralia was a simple-bodied organism lacking coelomic cavities, segmentation, and complex brain structures, and that more complex animals such as annelids evolved from such a simply organized ancestor. This conclusion contradicts alternative evolutionary scenarios proposing an annelid-like ancestor of Bilateria and Spiralia and several independent events of secondary reduction.

Article

Introduction

Molecular data have profoundly changed the view of the bilaterian tree of life by recognizing three major clades: Deuterostomia, Ecdysozoa, and Spiralia (Halanych 2004; Edgecombe et al. 2011). The term Spiralia is occasionally used as a synonym for Lophotrochozoa (Halanych 2004). However, the term Lophotrochozoa is actually reserved for all descendants of the last common ancestor of Annelida, Mollusca, and the three lophophorate taxa (Halanych 2004), whereas the more comprehensive taxon Spiralia includes all animals with spiral cleavage and, hence, also Platyhelminthes (Edgecombe et al. 2011). Herein, we use Spiralia in the terms of the more inclusive definition.

Previous results of the molecular phylogenetic analyses initiated a still on-going debate about the evolution of complexity in Bilateria. It was proposed that the last common ancestor of Deuterostomia, Ecdysozoa and Spiralia had a segmented and coelomate body organization resembling that of an annelid, and that morphologically more simply organized taxa such as nematodes or flatworms (Platyhelminthes) evolved by secondary reductions (Brinkman and Philippe 2008; De Robertis 2008; Couso 2009; Tomer et al. 2010; Chesebro et al. 2013). This is in stark contrast to the traditional "acoeloid–planuloid" hypothesis favoring evolution of Bilateria from a simple body organization toward more complex forms with a last common ancestor resembling a flatworm without segmentation and coelomic cavities (Hyman 1951; Halanych 2004; Hejnol et al. 2009). Unraveling the phylogenetic relationships within Bilateria is crucial to resolve this controversy (Halanych 2004; Edgecombe et al. 2011).

While recent phylogenomic studies recovered most of the relations of the major branches within Deuterostomia and

Ecdysozoa, the internal phylogeny of Spiralia is still unclear (Edgecombe et al. 2011). Indeed, spiralian animals exhibit a wide variety and plasticity in development and morphology including body organization (Nielsen 2012) which gave rise to the distinction of two major taxa: Lophotrochozoa and Platyzoa (Halanych 2004; Edgecombe et al. 2011). As mentioned above, Lophotrochozoa comprises at least annelids (ringed worms), lophophorates, and mollusks (Halanych 2004) and hence animals with a more complex morphology. In contrast, Platyzoa subsumes more simple appearing taxa such as flatworms, hairy backs (Gastrotricha), wheel animals (classical Rotifera), thorny-headed worms (Acanthocephala), and jaw worms (Gnathostomulida) (Cavalier-Smith 1998). Although some authors regard Platyzoa as sister to Lophotrochozoa (Edgecombe et al. 2011), others place Platyzoa within Lophotrochozoa, thus rendering Spiralia synonymous with Lophotrochozoa (Halanych 2004). Importantly, unique morphological autapomorphies supporting the monophyly of Platyzoa are lacking (Giribet 2008) and phylogenetic analyses of nuclear and mitochondrial data failed to resolve the question as well (Paps et al. 2009a, 2009b; Bernt et al. 2013). Nevertheless, there seems to be a tendency for a weakly supported monophylum Platyzoa as long as larger data sets were analyzed (Halanych 2004; Hausdorf et al. 2007; Struck and Fisse 2008; Hejnol et al. 2009; Paps et al. 2009a; Witek et al. 2009). However, across all these analyses placement of platyzoan taxa appeared unstable, probably due to low data and taxa coverage (Edgecombe et al. 2011). Moreover, parallel evolution of character states on long branches (also known as LBA) might also have confounded these analyses (Edgecombe et al. 2011). In summary, monophyly of Platyzoa and the phylogenetic positions of the platyzoan taxa within Spiralia are still contentious although their positions have major implications for bilaterian evolution. In particular, monophyly of Platyzoa and a placement within Lophotrochozoa would be in line with the theory of a more complex ancestry (Brinkman and Philippe 2008), whereas paraphyletic Platyzoa with respect to Lophotrochozoa would support the "acoeloid–planuloid" hypothesis.

Results and Discussion

To address the major outstanding issues of bilaterian phylogeny with respect to spiralian and more specifically platyzoan relationships, we applied a phylogenomic approach, generating transcriptome sequence data for 10 putative platyzoan and two nemertean species using second-generation sequencing technology and a modified RNA amplification method, which allowed the generation of sequencing libraries from as few as 10 specimens of microscopic species of Gnathostomulida, Gastrotricha, and classical Rotifera (supplementary table S1, Supplementary Material online). These data were complemented with transcriptomic or genomic data of 53 other spiralian and ecdysozoan species, including additional representatives of Platyzoa (supplementary table S2, Supplementary Material online). Hereby, the taxon coverage of Platyzoa increased 3.5-fold and for individual platyzoan taxa such as Syndermata (wheel animals and thorny-headed worms) and Gastrotricha even 5-fold in comparison to previous large-scale analyses of spiralian relationships (Dunn et al. 2008; Hejnol et al. 2009). After orthology assignment (Ebersberger et al. 2009), the data were further screened for sequence redundancy (Kvist and Siddall 2013), potentially paralogous sequences (Struck 2013) and contamination (Struck 2013) resulting in a pruning of about 7% of sequence data (supplementary tables S3–S8, Supplementary Material online).

Brute-Force Approach: More Taxa and Data

Phylogenetic reconstructions based on the largest data sets d01 with 82,162 amino acid positions and 38.3% sequence coverage (supplementary table S9, Supplementary Material online) recovered monophyly of both Platyzoa and Lophotrochozoa with strong bootstrap support (BS) of 99 for both (fig. 1). Within Platyzoa, monophyly of Platyhelminthes, of Syndermata, as well as of Gnathostomulida was maximally supported, whereas monophyly of Gastrotricha was not recovered. The chaetonotid gastrotrich *Lepidodermella squamata* appeared as sister to Platyhelminthes (BS 46), whereas the macrodasyidan gastrotrichs formed a monophylum (BS 74) as sister to all other platyzoan taxa (BS 68). Finally, Gnathostomulida was sister to Syndermata (BS 61) consistent with the Gnathifera hypothesis (Ahlrichs 1997; Herlyn and Ehlers 1997).

To study the influence of unstable taxa, leaf stability analyses were performed. With a leaf stability index of 0.876, the gastrotrich *Lep. squamata* was the most unstable species within the sampled platyzoans, followed by the two gnathostomulid species (0.941) and the macrodasyidan gastrotrich *Dactylopodola baltica* (0.969) (fig. 2 and supplementary table S10, Supplementary Material online). Excluding these four platyzoan taxa from data set d01 and conducting a new phylogenetic reconstruction did not influence the remaining topology, but led to an increased BS value of 99 for a clade uniting Platyhelminthes and Syndermata and decreased values for the monophyly of both Platyzoa and Lophotrochozoa (BS 82 and 86, table 1). Thus, the four unstable taxa showed some influence on support for the phylogenetic placement of other platyzoan taxa. Therefore, we excluded these four taxa from the following analyses, which addressed the potential role of LBA on platyzoan phylogeny in more detail.

LBA Accounts for Monophyly of Platyzoa

Monophyletic Platyzoa as sister to Lophotrochozoa gained strong support in the analyses described above. However, thorough inspection of the topology (fig. 1) revealed considerable branch length heterogeneity, with long branches in the analyzed platyzoan lineages and rather short branches in lophotrochozoan and ecdysozoan lineages. Hence, the observed strong support for monophyletic Platyzoa might originate from artificial rather than phylogenetic signal (Bergsten 2005; Edgecombe et al. 2011; Kück et al. 2012). For the tree derived from data set d01 and shown in figure 1, the LB scores showed a bimodal distribution with a minimum between the

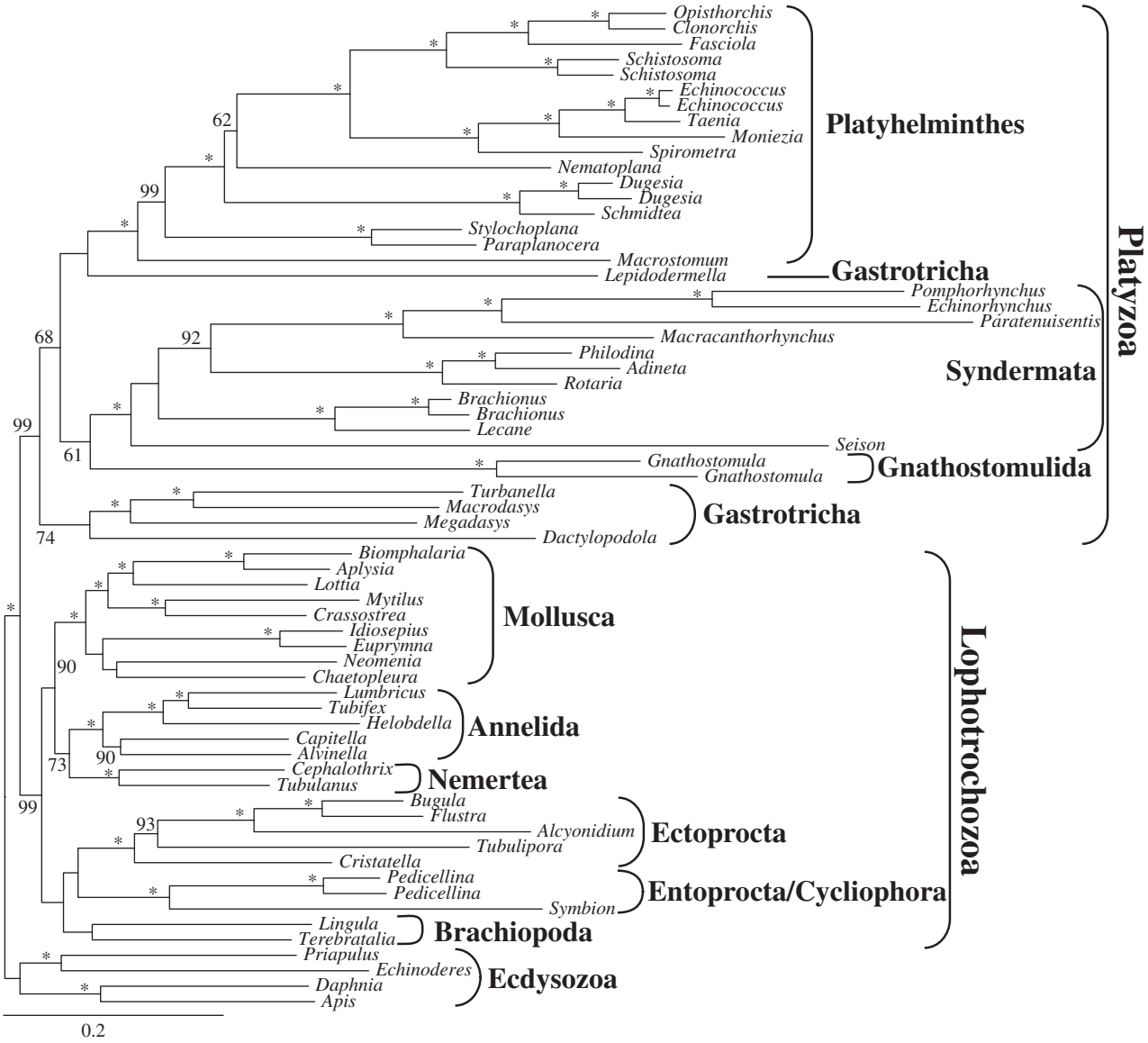

FIG. 1. Maximum-likelihood (ML) tree obtained by analysis of data set d01 with 65 taxa and 82,162 amino acid positions. Only BS values ≥50 are shown at the branches. *Maximal support of 100. Higher taxonomic units are indicated.

two highest optima at an LB score value of 0 (fig. 2*B*). Putative platyzoan species had generally higher LB score values than lophotrochozoan and ecdysozoan species (fig. 2 and supplementary table S11, Supplementary Material online). Only the LB scores inferred for *Stylochoplana* and *Paraplanocera* within Platyhelminthes, the two *Brachionus* species and *Lecane* in Syndermata, and *Megadasys* and *Macrodasys* in Gastrotricha approximated those of most lophotrochozoans and ecdysozoans. On the other hand, *Symbion* (Cycliophora), *Alcyonidium*, and *Tubulipora* (Ectoprocta) showed values >0, resembling those of most of the platyzoan species sampled (fig. 2).

To assess the effect of long branches on tree reconstruction, all species with LB scores above 0 were excluded from data set d01 (82,162 positions). Interestingly, monophyly of Platyzoa was no longer recovered (fig. 3). Gastrotricha now emerged as sister to Platyhelminthes (BS 96; table 1), and this clade was sister to monophyletic Lophotrochozoa (BS 95; table 1), whereas Syndermata was sister to all other spiralian taxa. Thus, exclusion of long-branched species had a tremendous effect on the analyses rendering a strongly supported monophyly of Platyzoa with BS values >95 into a paraphyletic assemblage, in which the clade consisting of Gastrotricha + Platyhelminthes and Lophotrochozoa obtained strong support with a BS value of 95.

Biases Causing Monophyletic Platyzoa

To gain further insights into the issue of mono- versus paraphyletic Platyzoa, we analyzed the data with respect to the different properties of individual genes. In detail, we studied the effect of gene-specific proportions of hydrophobic amino acids and missing data, base composition and branch length heterogeneity, and evolutionary rates on tree reconstruction. A common procedure is to choose one of these properties as

the most influential one either based a priori on literature or a posteriori on the obtained results (e.g., Brinkman and Philippe 2008; Simmons 2012b; Nesnidal et al. 2013; Nosenko et al. 2013; Roure et al. 2013; Salichos and Rokas 2013). Herein, we used another procedure based on the variability exhibited in the data itself prior to analyses of alternative data sets reflecting different degrees of data reduction. According to the principal component analysis (Alexe et al. 2008), the first principal component explained 31.0% of the variance between the different genes. It was mainly derived from the proportion of missing data and base composition heterogeneity with eigenvectors pointing into opposite directions (supplementary fig. S3 and table S12, Supplementary Material online). Branch length heterogeneity and evolutionary rate were the largest factors in the second component, which explained 26.8% of the variance. Correlation analyses showed that in our case evolutionary rate, which is often used as a proxy for branch length heterogeneity (Brinkman and Philippe 2008), did not correlate with actual measurements of branch length heterogeneity ($R^2 = 0.0324$ and 0.0635; supplementary fig. S4, Supplementary Material online).

As we wanted to test for LBA, we used the direct measurement of branch length heterogeneity instead of evolutionary rate. Thus, we generated data sets with either different degrees of missing data (d02–d06), proportion of low base composition heterogeneity (d07), or low branch length heterogeneity (d08) as well as genes being part of the 70% or 95% confidence intervals of the first two principal components (d09 and d10) (supplementary tables S9 and S13, Supplementary Material online). Based on the results of the principal component analysis, we present in detail the results of three data sets d07 (low base composition heterogeneity), d08 (low branch length heterogeneity), and d02. The latter combines a low degree of missing data with a high number of positions.

Analyses of these three data sets excluding the four abovementioned unstable platyzoan taxa consistently resulted in paraphyletic Platyzoa (fig. 4 and table 1) as observed before when excluding long-branched taxa from the large data set d01. Once more, Platyhelminthes was sister to Gastrotricha (BS 76, 84, and 71, Rouphozoa in table 1) and Lophotrochozoa was recovered as a monophyletic group (BS 98, 47, and 95). The clade of Gastrotricha/Platyhelminthes was sister to Lophotrochozoa (BS 72, 86, and 75, table 1) and Syndermata was sister to the all other spiralian taxa again. Thus, either by increasing the coverage (d02) or decreasing base composition or branch length heterogeneity (d07 and d08) paraphyletic Platyzoa was recovered (table 1), as a clade comprising Gastrotricha, Platyhelminthes and Lophotrochozoa gained strong branch support exceeding values of 70.

Additional exclusion of long-branched species (figs. 2 and 3) reproduced paraphyly of Platyzoa in all analyses, even with maximum BS in some analyses. Again Platyhelminthes was sister to Gastrotricha (BS 93, 97, and 54, fig. 5, Rouphozoa in table 1) and both were more closely related to the lophotrochozoan taxa than to Syndermata (BS 100, 100, and 50, fig. 5, Platyzoa Para. in table 1).

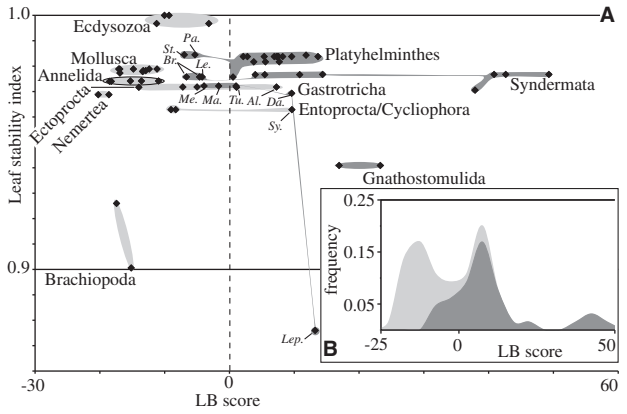

FIG. 2. Leaf stability indices and LB scores based on the ML analysis of data set d01 with 65 taxa and 82,162 amino acid positions. (*A*) Plot of leaf stability indices against LB scores. (*B*) Distribution of LB scores. The dashed line in *A* indicates LB score = 0. Shades distinguish ecdysozoan and lophotrochozoan species (light gray) from putative platyzoan species (dark gray). Some species are also labeled: Pa, *Paraplanocera*; St, *Stylochoplana*; Br, *Brachionus*; Le, *Lecane*; Me, *Megadasys*; Ma, *Macrodasys*; Da, *Dactylopodola*; Lep, *Lepidodermella*; Sy, *Symbion*; Tu, *Tubulipora*; Al, *Alcyonidium*.

Table 1. Bootstrap Support (BS) for Monophyly and Paraphyly of Platyzoa as well as Monophyly of Rouphozoa.

Data Set	Excl. Taxa	# Pos.	# Taxa	Platyzoa		Rouphozoa
				Mono.	Para.	Mono.
d01 (all data)	**None**	**82,162**	**65**	**99[a]**	**0**	**3**
	Unstable	**82,162**	**61**	**82[b]**	**1**	**1**
	LB	**82,162**	**34**	**3**	**95[a]**	**96[a]**
d02 (high coverage)	**Unstable**	**36,513**	**61**	**3**	**86[b]**	**84[b]**
	LB	**36,513**	**34**	**0**	**100[a]**	**93[b]**
d07 (low base frequency heterogeneity)	**Unstable**	**37,907**	**61**	**19**	**75[b]**	**71[b]**
	LB	**37,907**	**34**	**0**	**100[a]**	**97[a]**
d08 (low branch length heterogeneity)	**Unstable**	**29,133**	**61**	**18**	**72[b]**	**76[b]**
	LB	**29,133**	**34**	**10**	**50**	**54**

Excl., excluded; # pos., number of positions; # taxa, number of taxa; LB, long-branched taxa.

[a]Support values are part of the 95% confidence set.

[b]Support values are part of the 70% confidence set.

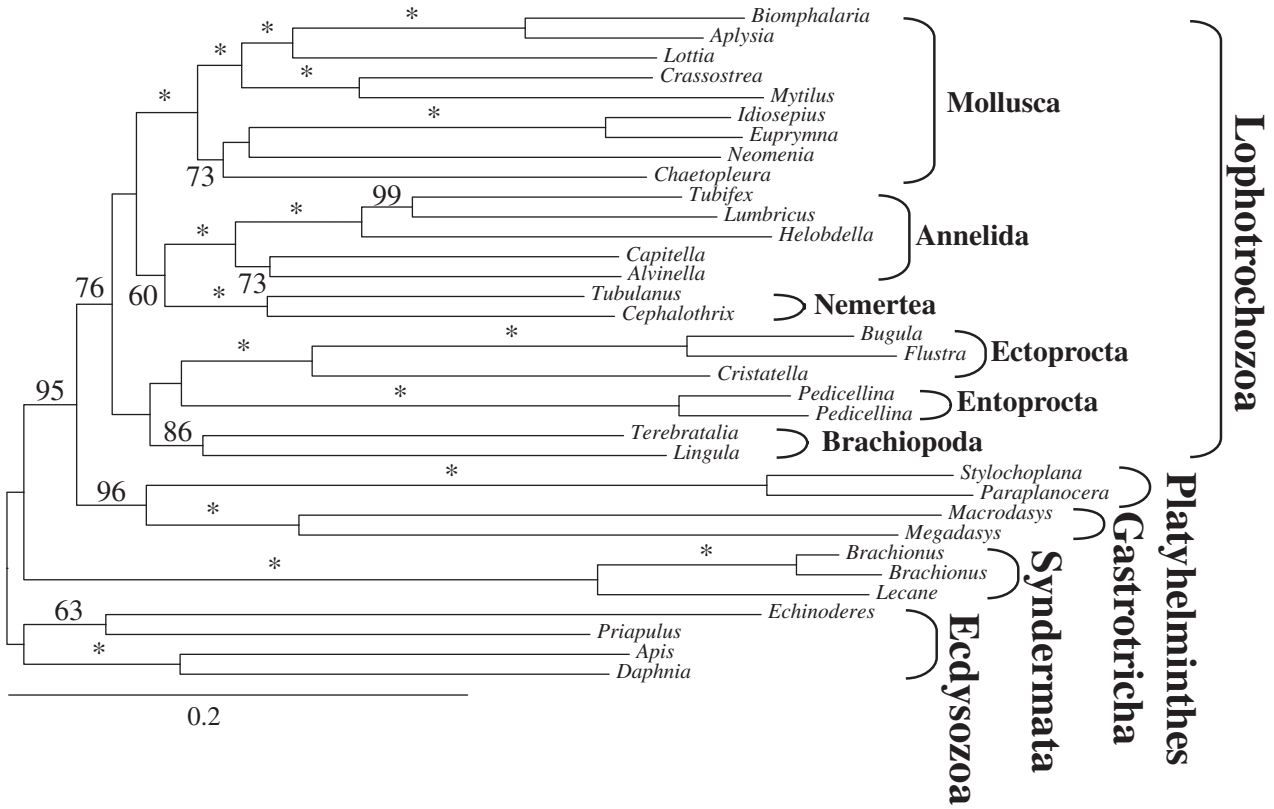

FIG. 3. ML tree obtained by analysis of data set d01 with 34 taxa and 82,162 amino acid positions. All taxa exceeding LB scores >0 in tree of figure 1 were excluded. Only BS ≥50 are shown at the branches. *Maximal support of 100. Higher taxonomic units are indicated.

Moreover, comparing the trees without long-branched species (figs. 3 and 5) to the one with all species (fig. 1) shows that now similar branch lengths lead to the "platyzoan" and lophotrochozoan species (figs. 3 and 5). Additionally, the standard deviation of the species-specific LB scores for the trees shown in figures 3 and 5 are 10.3 and 9.3, respectively, and, hence, lower than the standard deviation of 15.4 for the tree of figure 1. This means that the latter exhibits much stronger branch length heterogeneity across all taxa than the former two. Similarly, the standard deviations for the classical tip-to-root distances are lower for the trees of figures 3 and 5 with 0.054 and 0.143 than for the tree of figure 1 with 0.202.

We also used a Bayesian approach with the GTR + CAT model, as this is known to be more robust toward LBA than classical ML models such as LG (Lartillot et al. 2007). Due to computational time restrictions and high memory requirements, we were not able to use the large data set d01 (82,162 positions). Instead, we chose data set d02 (low to medium-low degree of missing data; 36,513 positions; 46.1% coverage; supplementary table S9, Supplementary Material online) as the principal component analysis indicated coverage as the most influential property in the first component. Importantly, the Bayesian approach did not recover monophyletic Platyzoa, but instead a clade including Gastrotricha + Platyhelminthes and monophyletic Lophotrochozoa (posterior probability [PP] = 1.00, fig. 6) and again Gnathostomulida + Syndermata was sister to this clade (PP = 1.00, fig.6).

Thus, combining Bayesian and maximum-likelihood analyses with different data and taxa exclusion strategies could not recover monophyletic Platyzoa in contrast to analyses using only large numbers of data (figs. 1, 3–6 and table 1). Considering all 10 data sets (i.e., d01–d10), BS for monophyletic Platyzoa substantially increased with additional amino acid positions (dark gray line in fig. 7A), whereas support for paraphyly decreased (black line in fig. 7A). In contrast, support for monophyly of Lophotrochozoa was not strongly affected by the number of positions analyzed (light gray line in fig. 7A). It is a well-known phenomenon of LBA that it is positively misleading; that is, with increasing numbers of positions the artificial group is more robustly recovered (Felsenstein 1978; Huelsenbeck 1997; Bergsten 2005). On the other hand, excluding long-branched species from analyses did not lead to such correlations. In particular, support for the monophyly of Platyzoa remained low irrespective of the number of alignment positions (dark gray line in fig. 7B).

Additionally, we determined for each data set the number of single-gene trees supporting monophyly or paraphyly of Platyzoa. Across all data sets the percentage of single genes supporting platyzoan paraphyly ranged from 8.6% to 11.7% and, thus, was higher than the percentage supporting monophyletic Platyzoa, ranging from 0.5% to 3.2% (table 2). Interestingly, decreasing the degree of missing data (i.e., d01–d06) and, hence, increasing the number of taxa per gene, the ratio of the percentage of trees supporting paraphyly relative to the percentage of trees supporting monophyly strongly increased (black line in fig. 7C).

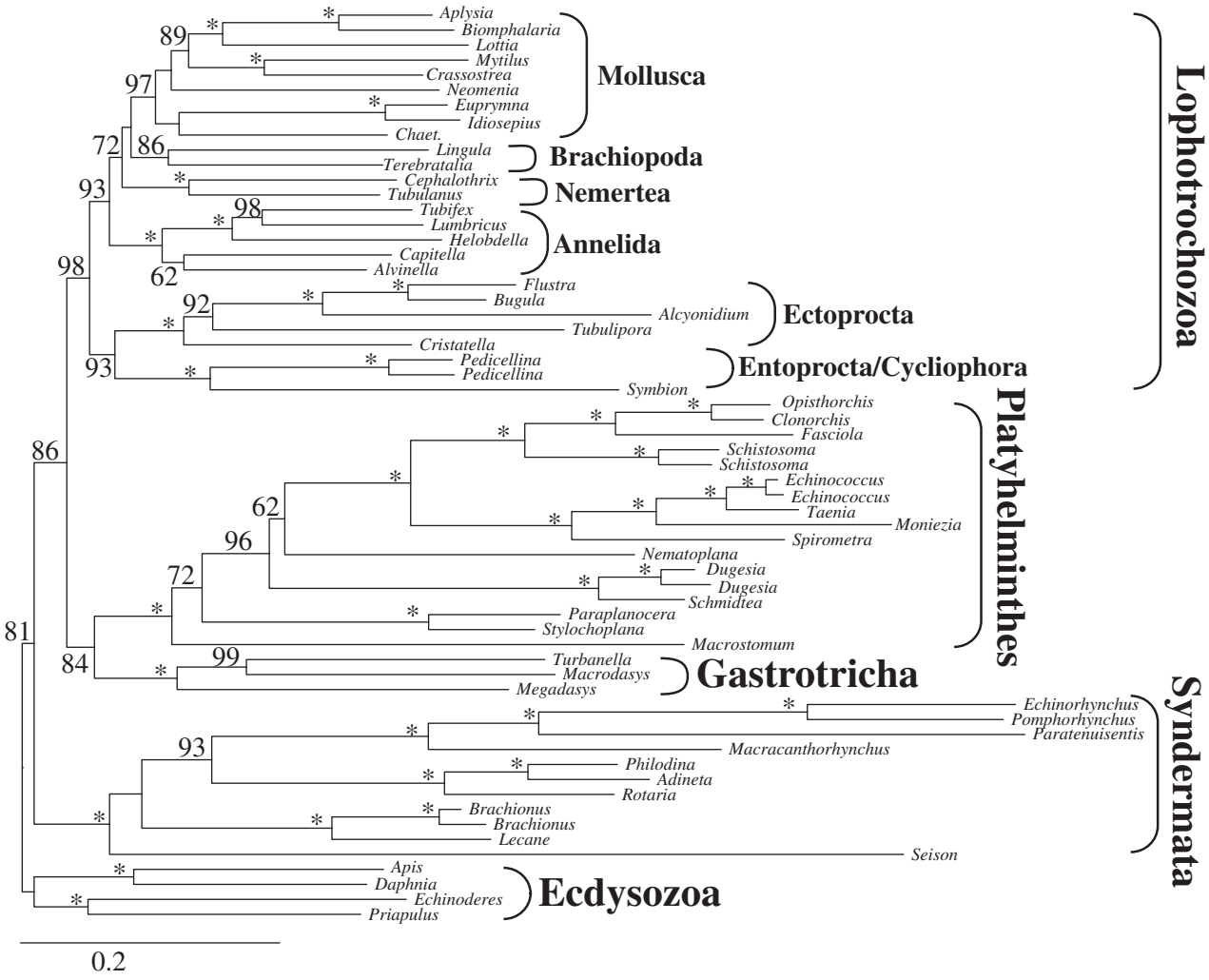

FIG. 4. ML tree obtained by analysis of data set d02 with 61 taxa and 36,513 amino acid positions. Only partitions with low to medium up to low degrees of missing data were included and the four unstable taxa (*Lepidodermella squamata*, *Dactylopodola baltica*, and the two Gnathostomulida species) were excluded. Only BS ≥ 50 are shown at the branches. *Maximal support of 100. Higher taxonomic units are indicated.

Directly addressing biases in the data such as base or branch length heterogeneity did not have such an effect on the ratio. In the case of LBA only strategies as used herein, which are able to attenuate its misleading effect by excluding either biased data or species or by increasing taxon coverage per gene, can reveal whether or not an assembly of long-branched taxa is artificially grouped together (Bergsten 2005). In conclusion, our analyses support platyzoan paraphyly, whereas recovery of monophyletic "Platyzoa" is most probably due to LBA.

Position of Gnathostomulida

In addition to LBA, the inference of a stable topology was hampered by the inclusion of Gnathostomulida and the two gastrotrichs *Lepidodermella* and *Dactylopodola*. In order to elucidate the phylogenetic position of Gnathostomulida within Spiralia, we reincluded the two formerly excluded gnathostomulid species into different data sets. Importantly, their inclusion did not alter the topology with respect to platyzoan paraphyly in any tree reconstruction (e.g., cf. figs. 4 and 8). Analysis of data set d07 excluding unstable taxa except Gnathostomulida (i.e., *Lepidodermella* and *Dactylopodola*) and of data set d02 excluding all long-branched taxa recovered Gnathostomulida as part of a clade with Gastrotricha and Platyhelminthes (table 3). However, all other analyses placed Gnathostomulida as sister to Syndermata with BS values of up to 91, even though overall BS remained low (fig. 8 and table 3). Moreover, the Bayesian analysis also recovered a sister group-relationship of Gnathostomulida and Syndermata with strong support (PP = 0.98, fig. 6). This position of Gnathostomulida as sister to Syndermata is consistent with the Gnathifera hypothesis (Ahlrichs 1997; Herlyn and Ehlers 1997). Monophyly of Gnathifera has also been found in previous studies based on ribosomal protein data (Witek et al. 2009; Hausdorf et al. 2010) and is also strongly supported by the likely homology of gnathostomulidan jaws and rotiferan trophi (Rieger and Tyler 1995; Haszprunar 1996; Ahlrichs 1997; Herlyn and Ehlers 1997; Jenner 2004a). For a thorough analysis of the phylogenetic relations within Syndermata and

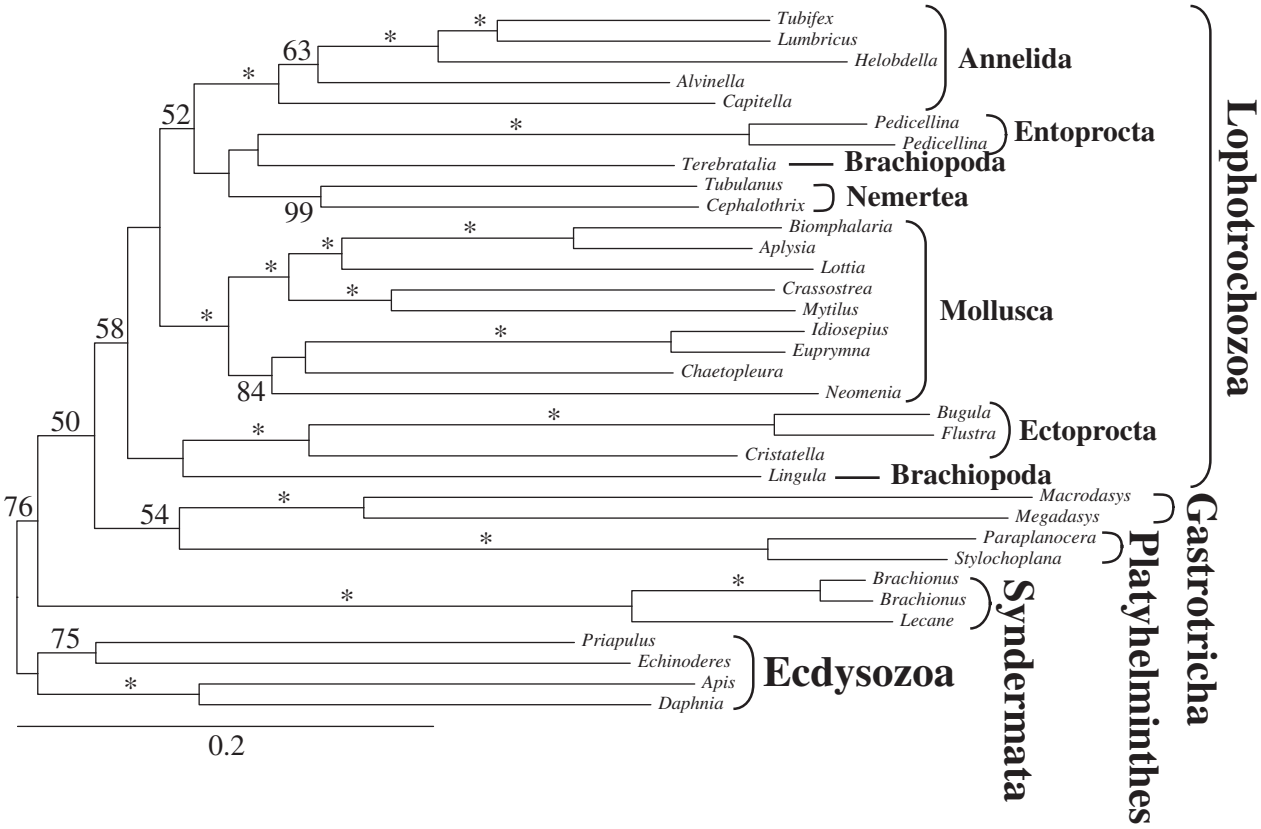

FIG. 5. ML tree obtained by analysis of data set d08 with 34 taxa and 29,133 amino acid positions. Only partitions with low degrees of branch length heterogeneity were included and all taxa exceeding LB scores >0 in tree of figure 1 were excluded. Only BS ≥50 are shown at the branches. *Maximal support of 100. Higher taxonomic units are indicated.

the implication for their evolution, we refer to a recent transcriptome-based study (Wey-Fabrizius et al. 2014).

A Novel View on Spiralian Phylogeny

In summary, our analyses support the monophyly of Lophotrochozoa and of a clade combining Gastrotricha and Platyhelminthes. Gnathifera is sister to a clade comprising the aforementioned taxa (fig. 9). No morphological apomorphy is known to date supporting either a monophyletic origin of Platyhelminthes and Gastrotricha or of Platyhelminthes, Gastrotricha, and Lophotrochozoa (Jenner 2004a; Rothe and Schmidt-Rhaesa 2009) and, hence, could be used for naming these two clades. However, whereas most of the other spiralian taxa exhibit additional structures for food gathering in their ground pattern (e.g., palps in annelids, proboscis in nemerteans, filter feeding apparatuses in lophophorates, entoprocts, and cycliophorans, as well as jaw-like elements in rotifers, gnathostomulids, and mollusks), gastrotrichs and most flatworm species ingest food without such extrastructures, just by dilating their rather simple pharynx. The respective pharynx simplex is part of the ground pattern of Platyhelminthes and enables the swallowing of prey by either sucking action or engulfment (Doe 1981). Gastrotricha possess a Y-shaped or inverted Y-shaped sucking pharynx (Kieneke et al. 2008). Although gathering food by sucking is not necessarily an autapomorphy of these two taxa, this common characteristic can nonetheless be utilized for naming the clade. We therefore suggest the name Rouphozoa (derived from the Greek word *rouphao* for ingesting by sucking) to define the last common ancestor of Platyhelminthes and Gastrotricha and all its descendants. The clade of Rouphozoa + Lophotrochozoa can be named Platytrochozoa, reflecting that it comprises Platyhelminthes and taxa with a trochophore larva and all extant descendants of the last common ancestor of Platyhelminthes and Lophotrochozoa. Spiralia then comprises Gnathifera (Syndermata + Gnathostomulida) and Platytrochozoa.

Implications for bilaterian evolution

The paraphyly of Platyzoa with respect to Lophotrochozoa is more in line with the traditional "acoeloid–planuloid" hypothesis than with the scenario of a last common ancestor of Deuterostomia, Ecdysozoa, and Spiralia with a segmented and coelomate body organization resembling an annelid. Within Spiralia the non-coelomate, small-sized taxa successively branch off first (fig. 9). Both Gnathostomulida and Gastrotricha comprise small interstitial organisms with an acoelomate body organization and <4 or 2 mm of length, respectively (Nielsen 2012). Within Syndermata only the highly modified, parasitic Acanthocephala are larger than a few millimeters and all exhibit a pseudocoelomate organization (Herlyn and Röhrig 2003; Nielsen 2012). Similarly, in Platyhelminthes, the ancestral condition is also a small-sized,

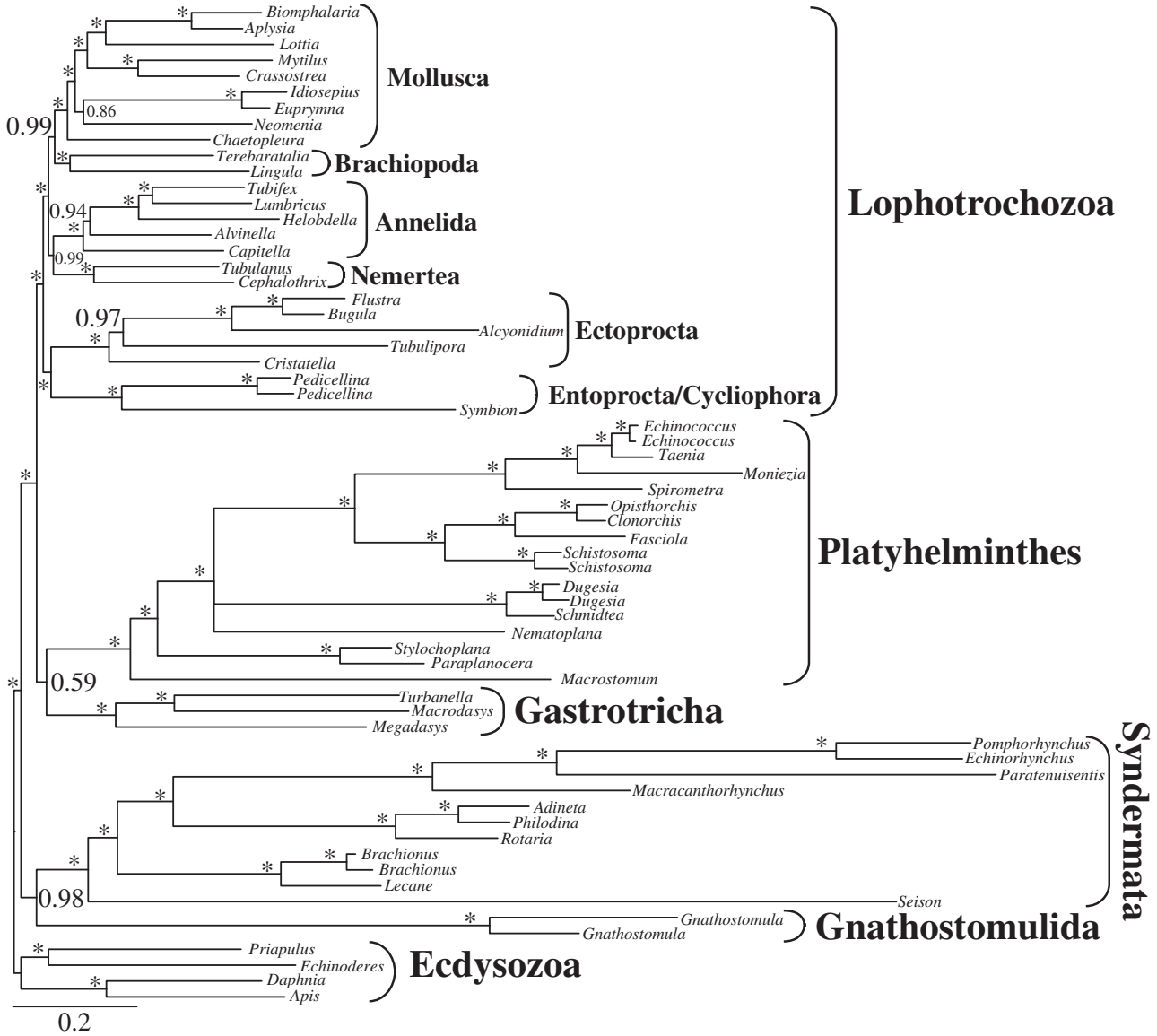

Fig. 6. BI tree obtained by analysis of data set d02 with 63 taxa and 36,513 amino acid positions. Only partitions with low to medium up to low degrees of missing data were included and only the two unstable gastrotrich taxa *Lepidodermella squamata* and *Dactylopodola baltica* were excluded. Only PPs $\geq$ 0.50 are shown at the branches. *Maximal support of 1.00. Higher taxonomic units are indicated.

acoelomate organization as seen today in Catenulida and Macrostomorpha, which are <5 mm in length (Nielsen 2012). Within Spiralia, animals with a coelomate body organization are, according to our analyses, only found in Lophotrochozoa (fig. 9). Thus, it is epistemologically more parsimonious to assume that the last common ancestor of Spiralia was an animal lacking a coelomic body cavity. Although many relationships within Lophotrochozoa are still unresolved in our study and warrant further investigations, our analyses suggest that within Spiralia coelomic cavities with a lining epithelium might have originated at the earliest in the stem lineage of Lophotrochozoa. Additionally, recent investigations of development and formation of coelomic cavities using a comparative anatomical approach revealed considerable differences between Annelida and Panarthropoda already in the earliest steps of coelomogenesis (for review, see Koch et al. 2014). Hence, segmental coeloms in annelids and arthropods are not necessarily homologous structures (Koch et al. 2014). In addition, the developmental origins of coelomic cavities in deuterostomes differ from those in lophotrochozoans and panarthropods (Nielsen 2012). Considering these differences and our results, it is more probable that coelomic cavities evolved independently within the major bilaterian clades Deuterostomia, Ecdysozoa, and Spiralia. Clearly, further analyses of the underlying genetic regulatory networks in coelom formation across a wide variety of coelomate and non-coelomate taxa are necessary to substantiate or reject this conclusion.

The position of coelomate Chaetognatha within Bilateria is also of interest in this aspect, but still enigmatic based on both molecular and morphological data. Deuterostome as well as protostome affinities including a sister group relationship to

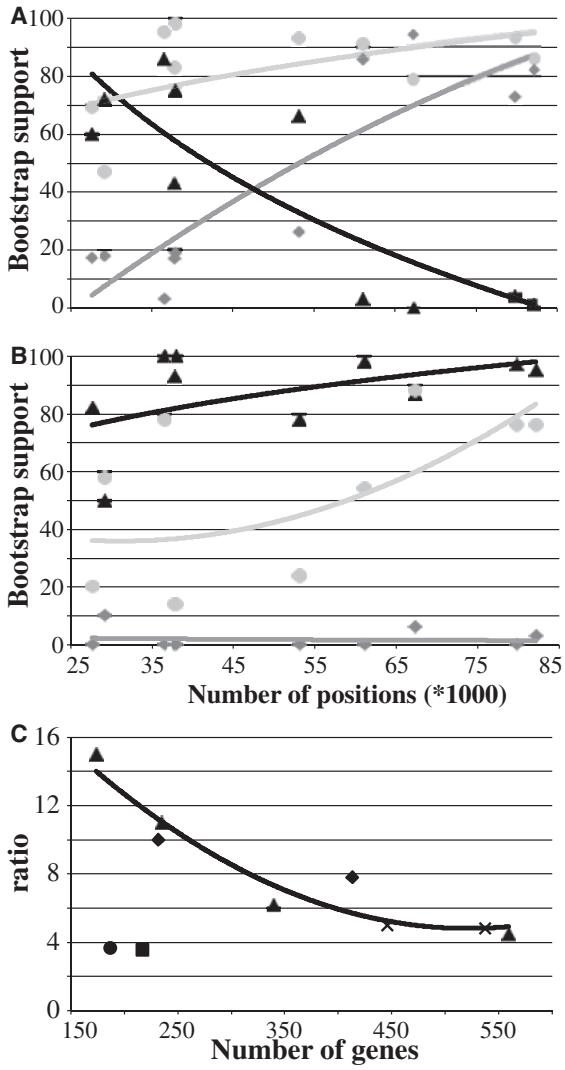

Fig. 7. BS and the ratio of single-genes supporting paraphyly over monophyly relative to the number of alignment positions or genes. (*A* and *B*) BS for monophyly and paraphyly of Platyzoa as well as monophyly of Lophotrochozoa relative to the number of positions. (*A*) Analyses based on 61 taxa, from which the four unstable taxa (*Lepidodermella squamata*, *Dactylopodola baltica*, and the two Gnathostomulida species) were excluded. (*B*) Analyses based on 34 taxa, from which all taxa exceeding LB scores >0 in tree of figure 1 were excluded. Light gray = monophyly of Lophotrochzoa, dark gray = monophyly of Platyzoa, black = paraphyly of Platyzoa. Best-fitting trend lines generated by Excel are also shown in the same colors. (*C*) Ratio of the percentage of single-gene trees supporting paraphyly of Platyzoa to the percentage of single-gene trees supporting monophyly of Platyzoa relative to the number of genes. Diamonds = data sets d02 and d03 with reduced missing data; triangles = data sets d01, d04–d06 generated using MARE; circle = data set d07 with reduced base heterogeneity; square = data set d08 with reduced branch length heterogeneity; crosses = data sets d09 and d10 based on confidence intervals of PCA. The best-fitting trend line generated by Excel for the data sets d01–d06 with decreasing degrees of missing data is shown in black.

Spiralia have been proposed (Marletaz et al. 2006; Matus et al. 2006; Dunn et al. 2008; Perez et al. 2014). Moreover, Chaetognatha possess a unique type of coelom formation, heterocoely, which exhibits no strong similarities to the other types of coelom formation (Kapp 2000; Perez et al. 2014) and, hence, might be indicative of a convergent evolution of coelomic cavities in Chaetognatha. However, ultrastructural studies of coelom formation are lacking at the moment (Perez et al. 2014).

The alternative scenario whereupon evolution progressed from complex to simple in Bilateria is mainly based on similarities in segmentation in vertebrates, arthropods, and annelids (De Robertis 2008; Couso 2009; Chesebro et al. 2013). However, in our analyses, Annelida was always deeply nested within Lophotrochozoa. Thus, similar to the evolution of coelomic cavities, a segmented ancestry of Spiralia would imply several independent losses of this organization, which we regard as less parsimonious. Moreover, Annelida and Arthropoda exhibit high plasticity in segmentation and, on the other hand, other spiralian and ecdysozoan taxa exhibit varying degrees of repetitive organization in organ systems. This includes Kinorhyncha, Monoplacophora and Polyplacophora, Eucestoda and other platyhelminths, some nematodes and nematomorphs, and a nemertean (Hannibal and Patel 2013; Struck 2012). In addition, segmentation is mostly restricted to tissue derived from the ectoderm in arthropods, from the mesoderm in vertebrates, and from both germ layers in annelids (Nielsen 2012). A possible explanation for similarities in segment formation including developmental pathways like the *notch* oscillation could be that these gene regulatory networks have been co-opted from ancestral networks involved in the organization of repetitive organ systems (Davidson and Erwin 2006; Chipman 2010). However, this hypothesis cannot be conclusively proven due to a current lack of data on developmental gene pathways in taxa with such repetitive organ systems (Chesebro et al. 2013). Nonetheless, the spiralian phylogeny derived herein provides additional support for the hypothesis that segmentation evolved independently within Deuterostomia, Ecdysozoa, and Spiralia.

Support for a complex bilaterian ancestor also arose from the observation of neuronal structures called mushroom bodies that were consistently present in arthropods and some annelids, as well as similar gene expression patterns noted in these bodies and in the vertebrate pallium (Heuer et al. 2010; Tomer et al. 2010). However, within annelids, mushroom bodies occur exclusively in five families of the subgroup Errantia, which are all characterized by a high vagility (Heuer et al. 2010; Struck et al. 2011), while they are not known for any other annelid or spiralian taxa (Rothe and Schmidt-Rhaesa 2009; Heuer et al. 2010; Nielsen 2012; Loesel 2014). Thus, if such distinct higher brain centers are taken as an ancestral condition of a complex last common spiralian ancestor (Heuer et al. 2010), several losses within Spiralia, including even several ones within Annelida, have to be assumed. On the other hand, the gastrotrich nervous system consists of a brain with a solid arch-like dorsal commissure with laterally positioned cell somata and a fine ventral commissure as well as a pair of longitudinal, lateroventral nerve cords joining posteriorly (Rothe and Schmidt-Rhaesa 2009). This organization is similar to the organization of the nervous system of Acoelomorpha. Hence, in comparison to

Table 2. Percentage of Single-Genes Supporting Monophyly or Paraphyly of Platyzoa.

		Degree of Missing Data					Heterogeneity		PCA	
Data Set	d01	d02	d03	d04	d05	d06	d07	d08	d09	d10
# Genes	559	232	413	340	235	174	217	187	446	537
% Mono.	2.1	0.9	1.2	1.5	0.9	0.6	3.2	3.2	2.0	2.0
% Para.	9.7	8.6	9.4	9.1	9.4	8.6	11.5	11.8	10.1	9.9
% Lack	88.2	90.5	89.3	89.4	89.8	90.8	85.3	85.0	87.9	88.1
Para./Mono.	4.5	10	7.8	6.2	11	15	3.6	3.7	5	4.8

Genes, number of genes in data set; % Mono., percentage of single-gene trees supporting monophyly of Platyzoa; % Para., percentage of single-gene trees supporting paraphyly of Platyzoa; % Lack, percentage of single-gene trees lacking resolution regarding this question; Para./Mono., ratio of the percentage of single-gene trees supporting paraphyly of Platyzoa to the percentage of single-gene trees supporting monophyly of Platyzoa; PCA, principal component analysis.

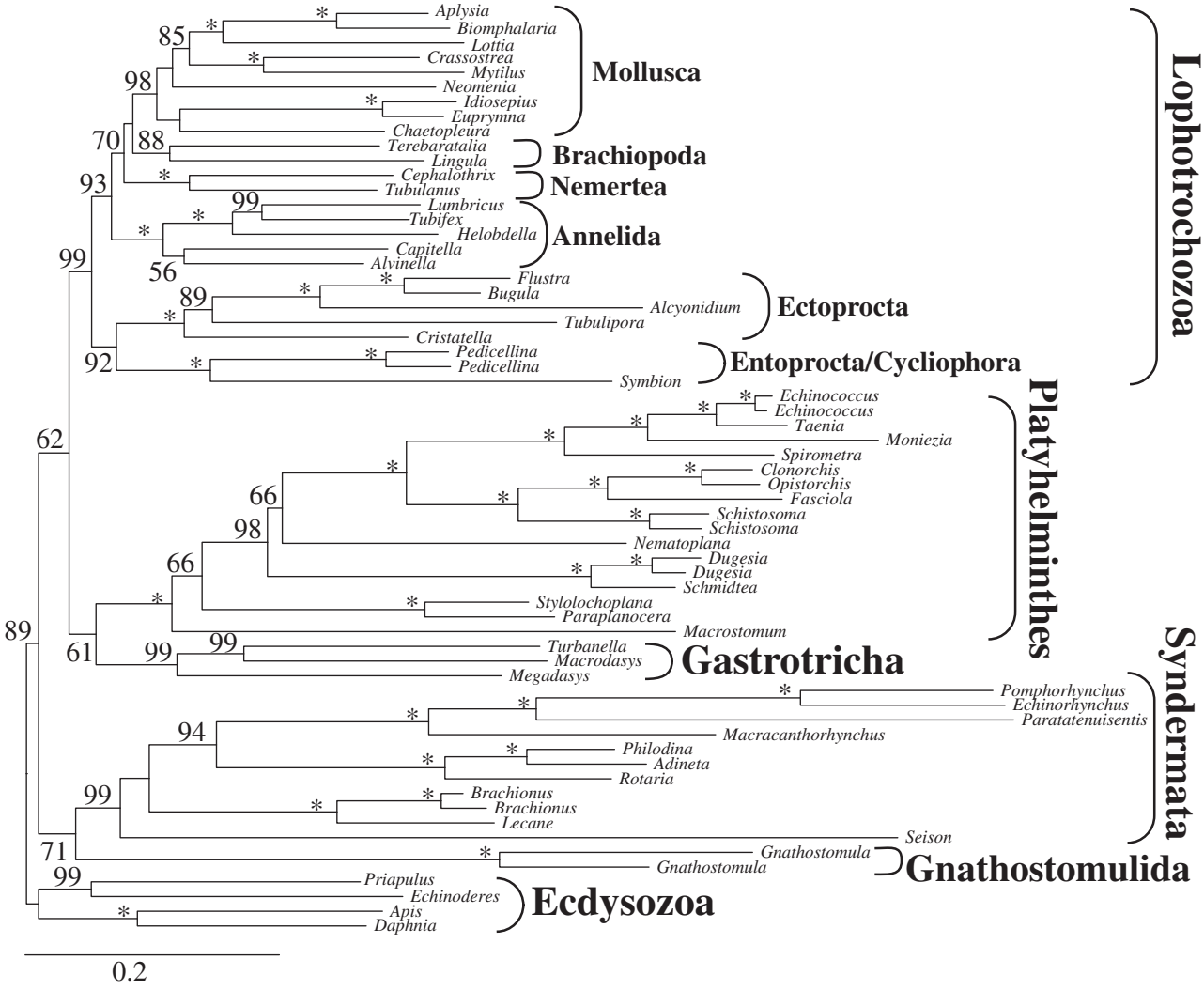

FIG. 8. ML tree obtained by analysis of data set d02 with 63 taxa and 36,513 amino acid positions. Only partitions with low to medium up to low degrees of missing data were included and only the two unstable gastrotrich taxa *Lepidodermella squamata* and *Dactylopodola baltica* were excluded. Only BS ≥50 are shown at the branches. *Maximal support of 100. Higher taxonomic units are indicated.

the net-like plexus without a cerebral ganglion in non-bilaterian animals, both Gastrotricha and Acoelomorpha express a certain degree of condensation at the anterior end to form a more or less condensed commissural brain, but to a lesser degree than other bilaterian taxa (Rothe and Schmidt-Rhaesa 2009). Thus, Gastrotricha might still exhibit the ancestral bilaterian condition indicative that also the last common ancestor of Spiralia showed that characteristic. Moreover, also, for example, platyhelminths, syndermatans, gnathostomulids, or entoprocts show anterior condensations of the central nervous system, but not to the same degree as in elaborate brains, which can be found in some mollusks or annelids (Northcutt 2012; Loesel 2014). Such a condensation is in general agreement with a small-sized, noncoelomate

Table 3. BS for Monophyly of Gnathifera.

Data Set	Excl. Taxa	# Taxa	Gnathifera
d01 (all data)	**None**	**65**	**61**
	Unstable	**63**	**91**[a]
	LB	**36**	**67**
d02 (high coverage)	**Unstable**	**63**	**71**[a]
	LB	**36**	**10**[b]
d07 (low base frequency heterogeneity)	**Unstable**	**63**	**48**[b]
	LB	**36**	**86**[a]
d08 (low branch length heterogeneity)	**Unstable**	**63**	**24**
	LB	**36**	**12**

Excl., excluded (same as in table 1 except for Gnathostomulida); # Taxa., number of taxa; LB, long-branched taxa.

[a]Support values are part of the 70% confidence set.

[b]Gnathostomulida not placed as sister to Syndermata in the ML tree, but in a clade with Gastrotricha and Platyhelminthes.

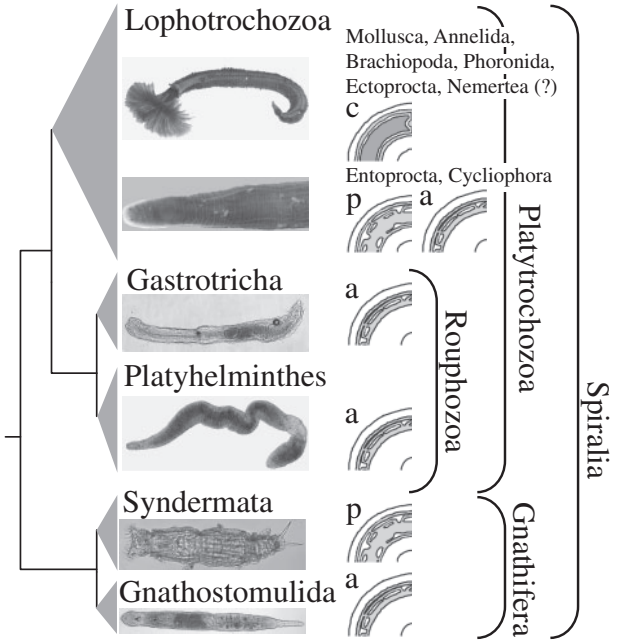

Fig. 9. Proposed phylogeny of Spiralia. Higher taxonomic units and names are given. Drawings depict the acoelomate (=a), pseudocoelomate (=p), and coelomate (=c) body organization. Picture of *Rotaria neptunoida* (Syndermata) was courtesy of Michael Plewka. (?) means that it is still discussed if the lateral vessels of the nemertean circulatory system are homologous to coelomic cavities of other lophotrochozoan taxa (Turbeville 1986).

ancestor for Spiralia showing no complex body organization. On the other hand, the observed similarities in the expression profiles of mushroom bodies of arthropods and annelids as well as the vertebrate pallium support the view that the evolution of more complex brain centers occurred early on in Bilateria (Heuer et al. 2010; Tomer et al. 2010). However, all three organs are part of the olfaction system. Analyses of these expression profiles in the brains of other bilaterian taxa are lacking in the moment. Hence, instead of being indicative of elaborative morphological structures, the observed similar expression profiles could be part of ancestral gene regulatory networks involved in the integration of chemosensory input in clusters of cells of more simply organized brains. However, developmental biological studies of the olfaction system of other bilaterian taxa such as Gastrotricha or Platyhelminthes are required to substantiate either hypothesis.

In conclusion, paraphyly of "Platyzoa" with respect to Lophotrochozoa and the spiralian phylogeny presented herein provide support for the view that the last common ancestor of Spiralia was an organism without coelomic cavity, segmentation, and elaborate brain structures, which probably inhabited the marine interstitial realm. This implies that evolution in Bilateria progressed most likely from a simple ancestor to more complex descendants independently within the three major bilaterian clades. However, we cannot rule out that miniaturization or a progenetic origin of the discussed taxa lead to loss of their morphological complexity. Several such examples are known from annelids and arthropods as in these cases it was more parsimonious to assume secondary simplification than convergent evolution (Jenner 2004b; Bleidorn 2007). However, the above discussion also shows that besides a robust phylogeny of Spiralia and Bilateria developmental biological studies of gene regulatory networks and expression profiles beyond the few standard model organisms are necessary to understand the evolution of Spiralia.

Material and Methods

Data Generation

Supplementary table S1, Supplementary Material online, lists species (four gastrotrich, two flatworms, two wheel animals, one acanthocephalan, one gnathostomulid, as well as two nemertean species) collected for this study. As deeply sequenced transcriptome libraries were lacking for nemerteans, we additionally constructed them for representatives of this taxon. Upon collection, samples were either snap-frozen at −80 °C or stored in RNAlater. Total RNA was isolated using the NucleoSpin RNA XS Kit (Macherey-Nagel) for *Rotaria rotatoria* and *Lecane inermis* (both Syndermata, classical Rotifera); the peqGOLD MicroSpin Total RNA kit (peqlab) for *Gnathostomula paradoxa* (Gnathostomulida), *Megadasys* sp., *Macrodasys* sp., *Dac. baltica*, and *Lep. squamata* (all Gastrotricha); or the peqGOLD Total RNA kit (peqlab) for *Tubulanus polymorphus*, *Cephalothrix linearis* (both Nemertea), *Nematoplana coelogynoporoides* and *Stylochoplana maculata* (both Platyhelminthes), and *Macracanthorhynchus hirudinaceus* (Syndermata, Acanthocephala).

For all species, except the nemerteans, total RNA was reverse-transcribed to double-stranded cDNA with the MINT UNIVERSAL cDNA synthesis kit (Evrogen) to produce amplified cDNA libraries. For *R. rotatoria*, Gnathostomulida and Gastrotricha a modified amplification protocol, which included an in-vitro transcription step, had been used. For this protocol, the cDNA synthesis was modified to contain 1 mM T7-PlugOligo (5′-C AATT GTAA TAC GAC TCA CTA TAGG GAGAACGGGGG-3′) comprising a T7 promotor sequence instead of 1 mM PlugOligo-3 M in combination with CDS-3 M adapter for the first strand synthesis and 0.1 mM

T7-primer (5′-AATT GTAA TAC GAC TCA CTA TAGG-3′) plus 0.1 mM M1-primer instead of 0.2 mM M1-primer for the second strand synthesis. Amplified cDNA was purified using the peqGOLD Cycle-Pure Kit (peqlab), digested with *SfiI* and size-fractioned using CHROMA SPIN-1000 (Clontech). Purified cDNA was vacuum-concentrated to 15.5 µl and 13 µl was used for the generation of mRNA by in vitro transcription (over night; 37 °C) employing T7 RNA polymerase (reaction conditions: 40 µl with 0.075 mM of each NTP, 1 u/µl RNase inhibitor, 0.5 mM DTT, and 5 u/µl T7 RNA polymerase [Invitrogen]). Messenger RNA was purified using peqGOLD Total RNA kit (peqlab).

The amplified cDNA libraries prepared from platyhelminths and *Lec. inermis* were sequenced by GENterprise GmbH (Mainz) or the Max Planck Institute for Molecular Genetics (Berlin) by 454 pyrosequencing using standard protocols. Illumina sequencing libraries for Nemertea, Gnathostomulida, and Gastrotricha were prepared with double indices following the protocol described by Meyer and Kircher (2010) and Kircher et al. (2011) starting either with totalRNA (Nemertea) or amplified mRNA (Gnathostomulida and Gastrotricha) as described by Hering et al. (2012). The libraries were sequenced at the Max Planck Institute of Evolutionary Anthropology (Leipzig), using an Illumina Genome Analyzer IIx (GAIIx) with 76 cycles paired end. Total RNA of *M. hirudinaceus* and amplified mRNA of *R. rotatoria* were sequenced using an Illumina HiSeq 2000 (100 bp paired end) at the Institute of Molecular Genetics, Johannes Gutenberg University (Mainz). The sequencing library of *M. hirudinaceus* was additionally run on an Illumina MiSeq machine (150 bp paired end) by GENterprise GmbH (Mainz). Publically available transcriptomes (ESTs and RNA-Seq) and genomic data from 49 spiralian species complemented these data (supplementary table S2, Supplementary Material online). For the choice of outgroup taxa, different considerations have to be taken into account given that platyzoan taxa are eventually affected by LBA. First of all, the outgroup taxa should not introduce additional long branches themselves (Bergsten 2005). Hence, distantly related outgroup taxa should be avoided as well as outgroups exhibiting increased substitution rates (Milinkovitch et al. 1996; Philippe et al. 2011). Therefore, we used only representatives of Ecdysozoa, the sister group of Spiralia, and did not consider nematodes and nematomorphs, which are known to possess long branches themselves. Moreover, more than a single outgroup taxon should be used and the diversity of outgroup taxa should be reflected (Milinkovitch et al. 1996; Bergsten 2005). Thus, we chose representative species of priapulids, kinorhynchs, and pancrustaceans as it has been previously shown that three to four outgroup taxa are sufficient to resolve difficult phylogenies when one also takes into account the computational limitations of phylogenomic studies (Rota-Stabelli and Telford 2008). Finally, the properties of the outgroup taxa sequence data should be similar to the ones of the ingroup taxa (Rota-Stabelli and Telford 2008) and in the case of LBA being more similar to short-branched ingroup taxa than to the long-branched ones. The LB scores show that the chosen ecdsyozoan species are similar to the short-branched spiralian taxa (fig. 2). For other properties, such as proportion of missing data and especially base composition heterogeneity, ecdysozoan taxa are similar to the ingroup taxa (supplementary fig. S5, Supplementary Material online).

Data Assembly

Processing of *M. hirudinaceus* and *R. rotatoria* data was performed using the FastX toolkit and included trimming of (I) 12 bp at the 5′-end, (II) adapter sequences, and (III) low-quality bases (cutoff 25). Reads longer than 20 bp after trimming were sorted into intact pairs and singletons using a custom perl script and were subsequently assembled using the CLC Genomics Workbench 5.5 (CLC Bio).

For the GAIIx databases were called with IBIS 1.1.2 (Kircher et al. 2009), adaptor and primer sequences removed and reads with low complexity as well as mispaired indices discarded. Raw data of all libraries were trimmed, discarding all reads with more than 5 bases below a quality score of 15. For 454 pyrosequencing data, sequences were thinned and quality filtered as implemented by Roche. In contrast to those data that were retrieved from the NCBI nr database (i.e., *Moniezia expansa*) as well as the genomic data present in the lophotrochozoan core ortholog set of HaMStR (i.e., *Schistosoma mansoni*, *Lottia gigantea*, *Helobdella robusta*, *Capitella teleta*, and *Apis mellifera*), the other data were further trimmed, quality-filtered and assembled as described in either Hausdorf et al. (2007) or in Riesgo et al. (2012) using the CLC Genomics Workbench with 0.05 as the limit for thinning and the scaffolding option in the assembly.

Sets of orthologous genes were determined using a profile hidden Markov model-based, reciprocal hit triangulation search using a modified version of HaMStR version 8 (Ebersberger et al. 2009) (called HaMStRad and the modified files are available at https://github.com/mptrsen/HaMStRad, last accessed April 24, 2014). As a core set we used the Lophotrochozoa set of 1,253 genes derived from the Inparanoid database (http://inparanoid51.sbc.su.se, last accessed April 24, 2014) for the primer-taxa *Cap. teleta*, *H. robusta*, *Lo. gigantea*, *S. mansoni*, *Daphnia pulex*, *Ap. mellifera*, and *Caenorhabditis elegans*. Modifications of HaMStR included the usage of Exonerate (Slater and Birney 2005) instead of Genewise (Birney et al. 2004) to provide frameshift-corrected, corresponding nucleotide sequences. We used the representative option with all primer taxa, the relaxed option and a cutoff e value of e^{-05}. Using the representative option might result in the assignment of the same sequence into different sets of orthologous genes. Such redundantly assigned sequences were removed using custom perl scripts, and the responsible bug in HaMStR fixed for future analyses. Each set of orthologous genes was individually aligned using MAFFT-Linsi (Katoh et al. 2005) followed by the determination of questionably aligned positions with AliScore (Kück et al. 2010) and masking with AliCut using default parameters. The 1,253 genes were concatenated into a super-matrix using FASconCAT (Kück and Meusemann 2010) and the super-matrix was reduced based on the

phylogenetic signal in a gene by assessing the tree-likeness by quartet-mapping using extended geometry mapping as implemented in MARE (Meusemann et al. 2010). We excluded the species of the core ortholog set *S. mansoni*, *Lo. gigantea*, *H. robusta*, *Cap. teleta*, *Dap. pulex*, and *Ap. mellifera* prior to matrix reduction and used a *d* value of 0.5 generating the large data set d01 (supplementary fig. S6, Supplementary Material online).

Paralogy and Contamination Screening

The 559 genes present in data set d01 were further screened for paralogous sequences and contamination within single-gene data sets. For this purpose, a screening based on bootstrap maximum-likelihood (ML) analyses of the individual genes (Philippe et al. 2011; Struck 2013) was conducted using TreSpEx (www.annelida.de, last accessed April 24, 2014). Initially, ML analyses were conducted for the unmasked individual genes (supplementary fig. S6, Supplementary Material online). All bipartitions supported by a bootstrap value $\geq$95 were extracted from the resulting topologies. As a first step all bipartitions congruent with clades for which independent a priori evidence of monophyly exist were masked for the following steps (Struck 2013). The columns "group" and "subgroup" in supplementary table S2, Supplementary Material online, as well as genera with more than one representative indicate these a priori clades. To be conservative, only sequences of bipartitions that exhibited a conflict with these a priori clades were pruned (supplementary tables S4 and S5, Supplementary Material online). A conflict in this case meant that species of an a priori clade as well as other species were present in both groups of the bipartition. For example, Platyhelminthes was such an a priori clade and, if in a bipartition platyhelminth as well as other spiralian and/or ecdysozoan species were present in both clades of the bipartition, this was regarded as a strong conflict. Thus, there was a strong conflict in these cases regarding the monophyly of a clade with a priori independent evidence of monophyly. At the group level all, but one clade fulfilled this criterion, that is, showed strong conflicts. The single exception was a clade comprising only all gnathiferan species in that data set eventually reflecting true phylogenetic signal. Previous studies have shown that such a pattern is characteristic for phylogenies of paralogous sequences reflecting the gene tree rather than the species tree (Rodríguez-Ezpeleta et al. 2007; Philippe et al. 2009, 2011; Struck 2013). However, other sources of artificial signal like shared missing data, compositional biases, contamination, or LBA (Bergsten 2005; Lemmon et al. 2009; Simmons and Freudenstein 2011; Simmons 2012a, 2012b; Struck 2013) can also result in such a pattern. In any case, potentially strong misleading signal with significant BS in single gene analyses has been masked by this procedure.

The paralogy screening was followed by a screening procedure for contamination in the libraries of our study. Therefore, the 18 S rRNA sequence of *Lineus bilineatus* (DQ279932) was blasted against each assembled library (supplementary fig. S6, Supplementary Material online) using BlastN and a cutoff value of e^{-20}. All detected contigs were then blasted against the NCBI nr database using BlastN. If the best hit represented a species from a different supra-specific taxon with the traditional rank of a phylum than the query species, this was taken as an indication of possible contamination (supplementary table S6, Supplementary Material online). For example, for some of the contigs of the *Alvinella pompejana* (Annelida) library, blast searches resulted in best hits linking the query sequence to the nematod *Tripylella* sp., the arthropod *Ptinus fur*, or an uncultured acaulosporan fungus. To prune eventually contaminated sequences from the sets of 559 genes, reference databases were specifically generated for each affected species based on the blast results against the NCBI database. For the *Alvinella* example, a reference database consisted of the non-redundant proteome information retrieved from the genomes of *Ap. mellifera* (Arthropoda), *Cae. elegans* (Nematoda), *Schizosaccharomyces cerevisiae* (Fungi), and the transcriptome of *Dap. pulex* (Arthropoda) as negative references as well as from the genomes of *Cap. teleta*, *H. robusta* (Annelida), *Lo. gigantea* (Mollusca), and *Schmidtea mediterranea* (Platyhelminthes) as positive references. Each of the 559 genes present for that species (e.g., *Al. pompejana*) was blasted against this species-specific reference database. Three pruning strategies were tested: a sequence was pruned when (I) the best hit was a negative reference sequence, (II) the best hit was a negative reference sequence and in addition the *E* value was at least one order of a magnitude better than that of the best hit for a positive reference, or (III) the best hit was a negative reference sequence and in addition the *E* value was at least four orders better than that of the best hit for a positive reference. As ML analyses of the data set d01 with 65 taxa and 82,162 amino acid positions using the three different pruning strategies resulted in no significant differences of the topologies inferred, we chose the most conservative first pruning strategy for subsequent analyses. Custom Perl scripts were written for all these steps.

Phylogenetic Analyses

The most appropriate substitution model was LG + I + Γ as determined using the ProteinModelSelection script for RAxML (Stamatakis 2006). Before the time-consuming Bayesian Inference (BI), we conducted a series of ML analyses as part of the sensitivity analyses and screening procedures (see supplementary fig. S6, Supplementary Material online). In total, 1,129 ML analyses were conducted with RAxML 7.3 (Stamatakis 2006) using 300 and 100 bootstrap replicate searches for concatenated and individual gene data sets, respectively. The bootstrap searches were followed by a search of the best tree. Preliminary analyses using the automatic bootstopping option (Pattengale et al. 2009) (-# autoMRE) in RAxML obtained a maximum of 240 bootstrap replicates for different tested concatenated data sets and, hence, we used 300 replicates for all analyses for reasons of comparability. Moreover, these preliminary analyses showed that a bootstrap search followed by a best tree search always found a tree with an equal or better likelihood score than independent

searches for the best tree using 100 replicate searches starting from randomized maximum-parsimony trees.

For the BI analysis, we used PhyloBayes MPI 1.4f (Lartillot and Philippe 2004; Lartillot et al. 2013) using the GTR + CAT model and the data set d02 generated by excluding genes with high degrees of missing data (see sensitivity analyses below). For the analysis, four chains ran in parallel for 13,669 cycles on average (ranging from 12,164 to 14,217). Convergence of likelihood values, alpha parameter, and tree length of the four chains was assessed using Tracer v1.5 (http://tree.bio.ed.ac.uk/software/tracer, last accessed April 24, 2014). Upon convergence the average standard deviation of split frequencies was <0.1 with a value of 0.055. The first 6,000 cycles of each chain were discarded as burnin and the majority rule consensus tree containing the posterior probabilities was calculated from the remaining trees of the chain with the best average likelihood score sampling every second tree.

Sensitivity Analyses

Leaf stability indices of species were determined using Phyutility (Smith and Dunn 2008) and the bootstrap trees of ML analyses of the data set d01 comprising all species sampled. To assess the branch length heterogeneity, we used the herein newly developed LB score using TreSpEx (www.annelida.de, last accessed April 24, 2014), which we also used to calculate classical tip-to-root distances. Taxa were excluded from the data sets d01–d10 in accordance with these results and the phylogenetic reconstructions repeated.

To objectively assess the branch length heterogeneity in a tree, we developed a new tree-based measurement, which we call the LB score. The score utilizes patristic distances (PDs), that is, the distance between two taxa based on the connecting branches, and is based on the mean pairwise PD of a taxon i to all other taxa in the tree relative to the average pairwise PD over all taxa (a):

$$\mathrm{LB}_i = \left(\frac{\overline{\mathrm{PD}_i}}{\overline{\mathrm{PD}_a}} - 1 \right) * 100.$$

In specific, the score measures for each taxon the percentage deviation from the average and is independent of the root of the tree. The latter is also the reason for not using the traditional tip-to-root distance (Bergsten 2005). When using tip-to-root distances, the recognition of long-branched taxa heavily depends on the root of the tree. For example, in the reconstruction of the individual gene with the ID 111427 in our analyses below the ecdysozoan outgroup species are not monophyletic. Whereas tip-to-root distances based on an *Apis*-rooted tree and LB scores indicate the same taxa as long-branched, rooting the tree with either *Echinoderes* or *Priapulus* some of these species would be indicated as short-branched (supplementary fig. S1, Supplementary Material online). Given the automatic process pipelines in phylogenomic analyses due to the vast amount of genes detection of long-branched taxa should be robust against changes in the root of the tree. Moreover, in the search for the best tree in phylogenetic reconstructions only unrooted trees are used and rooting is an a posteriori procedure. Thus, notwithstanding that outgroup species might be long-branched the artificial grouping of species due to LBA in phylogenetic reconstructions is not directly due to the root by itself (Bergsten 2005). Hence, detection of LBA should be independent of the root. Fortunately, either using the large data set d01 in our analyses below or the 559 individual genes of this data set LB scores and tip-to-root distances are highly and positively correlated with a R^2 value of 0.91543 or an average R^2 of 0.85684, respectively (supplementary fig. S2, Supplementary Material online).

For data partitioning, we analyzed the 559 genes of the data set d01 generated with the MARE setting "all taxa included" and a d value of 0.5. We determined both alignment- and tree-based properties. Using BaCoCa (Kück and Struck 2014), the proportion of hydrophobic and polar amino acids, the proportion of missing data as well as the compositional heterogeneity as measured by the RCFV values (Zhong et al. 2011) were determined from the pruned and masked alignments across all species in each gene (supplementary fig. S6, Supplementary Material online). ML trees from these alignments were used to determine the evolutionary rate for each gene, calculated as the average pairwise PD between two species in the tree, as well as the mean of the upper quartile of LB scores (i.e., the upper 25% of all LB scores) and the standard deviation of all LB scores as measurements of branch length heterogeneity with the aid of TreSpEx (www.annelida.de). Correlation studies of these properties were conducted in Excel (supplementary figs. S2 and S4, Supplementary Material online). Principal component analyses were conducted in R with scaled values (supplementary fig. S3, Supplementary Material online). The determination of the branch length heterogeneity within a gene was based on either the mean of the upper quartile of LB scores or the standard deviation of all LB scores within a gene. However, both approaches led to a strong linear correlation ($R^2 = 0.8363$, supplementary fig. S4, Supplementary Material online) and, thus, we used solely the standard deviation of LB scores as a measure of branch length heterogeneity in the principal component analysis. Similarly, the proportions of hydrophobic and polar amino acids were also strongly correlated ($R^2 = 0.6481$, supplementary fig. S4, Supplementary Material online) and, hence, we excluded the proportion of polar amino acids.

Genes with either high degrees of missing data or high base composition heterogeneity were excluded based on the results of heat map analyses in combination with hierarchical clustering without scaling the values in *R* (Bapteste et al. 2005; Susko et al. 2006). Four clusters of proportion of missing data were found ranging from low to high degrees of missing data (supplementary fig. S7, Supplementary Material online). From data set d01 genes belonging to the groups with medium-high to high degrees of missing data were excluded to generate data set d02 characterized by only low degrees of missing data. We also generated a data set d03, where we excluded only high degrees of missing data from data set d01. Alternatively, the data set d01 was condensed using MARE

(Meusemann et al. 2010) with *d* values of 1.0, 1.5, or 2.0 instead of 0.5 used above resulting in the data sets d04, d05, and d06, respectively.

For compositional heterogeneity, the heatmap revealed three clusters with low, medium, and high compositional heterogeneity (supplementary fig. S8, Supplementary Material online). Only genes, which were part of the cluster with low compositional heterogeneity, were kept for data set d07. The tree-based property branch length heterogeneity was ranked and divided into three equal parts. To generate data set d08 only the genes from the third with the lowest heterogeneity values were not excluded (supplementary fig. S6 and table S9, Supplementary Material online). Moreover, we excluded all genes from data set d01, which were not part of the 70% or 95% confidence interval of the first two principal components (supplementary fig. S3, Supplementary Material online) resulting in data sets d09 and d10, respectively. Finally, for each data set we determined the number of single-gene trees, which found a monophyletic or paraphyletic Platyzoa using custom perl scripts. For the latter at least one platyzoan taxon (i.e., Platyhelminthes, Gastrotricha, Gnathostomulida, or Syndermata) had to be placed more closely to the outgroup than at least one other platyzoan taxon.

Supplementary Material

Supplementary tables S1–S13 and figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the priority program "Deep Metazoan Phylogeny" of the Deutsche Forschungsgemeinschaft DFG-STR 683/5-2 and DFG-STR 683/8-1 to T.H.S., DFG-Ha2103/4 to T.H., DFG-HA 2763/5 to B.H., and DFG-BL 787/5-1 to C.B. T.H. wishes to thank the Center for Computational Sciences Mainz (CSM) for additional financial support. T.H.S. also acknowledges the support of the Regional Computing Centre of Cologne (RRZK) by, among others, providing access to the HPC cluster CHEOPS for the parallel PhyloBayes analyses. The authors gratefully acknowledge Edyta Fialkowska (Institute of Environmental Sciences, Jagiellonian University, Poland) for providing *Lecane* specimens, László Sugár (Faculty of Animal Science, Kaposvár University, Hungary) for collecting and providing *Macracanthorhynchus* specimens, Sanja Ramljak (Institute of Clinical Research and Development, Mainz, Germany) for helping with collecting *Seison* specimens and Michael Plewka (http://plingfactory.de/) for the picture of *Rotaria neptunoida*. They also thank Birgit Nickel and Matthias Meyer (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany) for their assistance in sequencing using Illumina GAIIx and Steffen Rapp (Institute of Molecular Genetics, Johannes Gutenberg University, Mainz) for operating the Illumina HiSeq 2000. Sequence data have been deposited in the NCBI short read archive. All other data and trees used in this study are available at DataDryad (doi:10.5061/dryad.n435p).

References

Ahlrichs WH. 1997. Epidermal ultrastructure of *Seison nebaliae* and *Seison annulatus*, and a comparison of epidermal structures within Gnathifera. *Zoomorphology* 117:41–48.

Alexe G, Vijaya Satya R, Seiler M, Platt D, Bhanot T, Hui S, Tanaka M, Levine AJ, Bhanot G. 2008. PCA and clustering reveal alternate mtDNA phylogeny of N and M clades. *J Mol Evol.* 67:465–487.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol.* 5:33.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21: 163–193.

Bernt M, Bleidorn C, Braband A, Dambach J, Donath A, Fritzsch G, Golombek A, Hadrys H, Jühling F, Meusemann K, et al. 2013. A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. *Mol Phylogenet Evol.* 69:352–364.

Birney E, Clamp M, Durbin R. 2004. Genewise and genomewise. *Genome Res.* 14:988–995.

Bleidorn C. 2007. The role of character loss in phylogenetic reconstruction as exemplified for the Annelida. *J Zool Syst Evol Res.* 45:299–307.

Brinkman H, Philippe H. 2008. Animal phylogeny and large-scale sequencing: progress and pitfalls. *J Syst Evol.* 46:274–286.

Cavalier-Smith T. 1998. A revised six-kingdom system of life. *Biol Rev.* 73: 203–266.

Chesebro JE, Pueyo JI, Couso JP. 2013. Interplay between a *Wnt*-dependent organiser and the *Notch* segmentation clock regulates posterior development in *Periplaneta americana*. *Biol Open.* 2: 227–237.

Chipman AD. 2010. Parallel evolution of segmentation by co-option of ancestral gene regulatory networks. *Bioessays* 32:60–70.

Couso JP. 2009. Segmentation, metamerism and the Cambrian explosion. *Int J Dev Biol.* 53:1305–1316.

Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. *Science* 311:796–800.

De Robertis EM. 2008. The molecular ancestry of segmentation mechanisms. *Proc Natl Acad Sci U S A.* 105:16411–16412.

Doe D. 1981. Comparative ultrastructure of the pharynx simplex in turbellaria. *Zoomorphology* 97:133–193.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–750.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol.* 9: 157.

Edgecombe G, Giribet G, Dunn C, Hejnol A, Kristensen R, Neves R, Rouse G, Worsaae K, Sørensen M. 2011. Higher-level metazoan relationships: recent progress and remaining questions. *Org Divers Evol.* 11:151–172.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool.* 27:401–410.

Giribet G. 2008. Assembling the lophotrochozoan (=spiralian) tree of life. *Philos Trans R Soc Lond B Biol Sci.* 363:1513–1522.

Halanych KM. 2004. The new view of animal phylogeny. *Annu Rev Ecol Evol Syst.* 35:229–256.

Hannibal R, Patel N. 2013. What is a segment? *Evo Devo* 4:35.

Haszprunar G. 1996. Plathelminthes and Plathelminthomorpha—paraphyletic taxa. *J Zool Syst Evol Res.* 34:41–48.

Hausdorf B, Helmkampf M, Meyer A, Witek A, Herlyn H, Bruchhaus I, Hankeln T, Struck TH, Lieb B. 2007. Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta. *Mol Biol Evol.* 24:2723–2729.

Hausdorf B, Helmkampf M, Nesnidal MP, Bruchhaus I. 2010. Phylogenetic relationships within the lophophorate lineages

(Ectoprocta, Brachiopoda, and Phoronida). *Mol Phylogenet Evol.* 55: 1121–1127.
Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc Lond B Biol Sci.* 276:4261–4270.
Hering L, Henze MJ, Kohler M, Kelber A, Bleidorn C, Leschke M, Nickel B, Meyer M, Kircher M, Sunnucks P, et al. 2012. Opsins in Onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods. *Mol Biol Evol.* 29: 3451–3458.
Herlyn H, Ehlers U. 1997. Ultrastructure and function of the pharynx of *Gnathostomula paradoxa* (Gnathostomulida). *Zoomorphology* 117:135.
Herlyn H, Röhrig H. 2003. Ultrastructure and overall organization of ligament sac, uterine bell, uterus and vagina in *Paratenuisentis ambiguus* (Acanthocephala, Eoacanthocephala)—the character evolution within the Acanthocephala. *Acta Zool.* 84:239–247.
Heuer C, Muller C, Todt C, Loesel R. 2010. Comparative neuroanatomy suggests repeated reduction of neuroarchitectural complexity in Annelida. *Front Zool.* 7:13.
Huelsenbeck JP. 1997. Is the Felsenstein zone a fly trap? *Syst Biol.* 46: 69–74.
Hyman LH. 1951. The invertebrates. Vol. 2. Platyhelminthes and Rhynchocoela: the Acoelomate bilateria. New York: McGraw-Hill.
Jenner RA. 2004a. Towards a phylogeny of the Metazoa: evaluating alternative phylogenetic positions of Platyhelminthes, Nemertea, and Gnathostomulida, with a critical reappraisal of cladistic characters. *Cont Zool.* 73:3–163.
Jenner RA. 2004b. When molecules and morphology clash: reconciling conflicting phylogenies of the Metazoa by considering secondary character loss. *Evol Dev.* 6:372–8.
Kapp H. 2000. The unique embryology of Chaetognatha. *Zool Anz.* 239: 263–266.
Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–518.
Kieneke A, Riemann O, Ahlrichs WH. 2008. Novel implications for the basal internal relationships of Gastrotricha revealed by an analysis of morphological characters. *Zool Scr.* 37:429–460.
Kircher M, Sawyer S, Meyer M. 2011. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 2011:1–8.
Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* 10:R83.
Koch M, Quast B, Bartolomaeus T. 2014. Coeloms and nephridia in annelids and arthropods. In: Wägele JW, Bartolomaeus T, editors. Deep metazoan phylogeny: the backbone of the tree of life—new insights from analyses of molecules, morphology, and theory of data analysis. Berlin (Germany): De Gruyter. p. 173–284.
Kück P, Mayer C, Wägele J-W, Misof B. 2012. Long branch effects distort Maximum Likelihood phylogenies in simulations despite selection of the correct model. *PLoS One* 7:e36593.
Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. *Mol Phylogenet Evol.* 56:1115–1118.
Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool.* 7:10.
Kück P, Struck TH. 2014. BaCoCa—a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Mol Phylogenet Evol.* 70:94–98.
Kvist S, Siddall ME. 2013. Phylogenomics of Annelida revisited: a cladistic approach using genome-wide expressed sequence tag data mining and examining the effects of missing data. *Cladistics* 29:435–448.
Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol.* 7:S4.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.
Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol.* 62:611–615.
Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by Maximum Likelihood and Bayesian Inference. *Syst Biol.* 58:130–145.
Loesel R. 2014. Brain complexity in protostomes. In: Wägele JW, Bartolomaeus T, editors. Deep metazoan phylogeny: the backbone of the tree of life—new insights from analyses of molecules, morphology, and theory of data analysis. Berlin (Germany): De Gruyter. p. 79–91.
Marletaz F, Martin E, Perez Y, Papillon D, Caubit X, Lowe CJ, Freeman B, Fasano L, Dossat C, Wincker P. 2006. Chaetognath phylogenomics: a protostome with deuterostome-like development. *Curr Biol.* 16: R577–R578.
Matus DQ, Copley RR, Dunn CW, Hejnol A, Eccleston H, Halanych KM, Martindale MQ, Telford MJ. 2006. Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Curr Biol.* 16: R575–R576.
Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walzl M, Pass G, Breuers S, et al. 2010. A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol.* 27: 2451–2464.
Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc.* 2010:pdb.prot5448.
Milinkovitch MC, LeDuc RG, Adachi J, Farnir F, Georges M, Hasegawa M. 1996. Effects of character weighting and species sampling on phylogeny reconstruction: a case study based on DNA sequence data in cetaceans. *Genetics* 144:1817–1833.
Nesnidal M, Helmkampf M, Meyer A, Witek A, Bruchhaus I, Ebersberger I, Hankeln T, Lieb B, Struck TH, Hausdorf B. 2013. New phylogenomic data support the monophyly of Lophophorata and an Ectoproct–Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. *BMC Evol Biol.* 13:253.
Nielsen C. 2012. Animal evolution—interrelationships of the living phyla. New York: Oxford University Press, Inc.
Northcutt RG. 2012. Evolution of centralized nervous systems: two schools of evolutionary thought. *Proc Natl Acad Sci U S A.* 109: 10626–10633.
Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WEG, Nickel M, Schierwater B, et al. 2013. Deep metazoan phylogeny: when different genes tell different stories. *Mol Phylogenet Evol.* 67:223–233.
Paps J, Baguna J, Riutort M. 2009a. Bilaterian phylogeny: a broad sampling of 13 nuclear genes provides a new Lophotrochozoa phylogeny and supports a paraphyletic basal Acoelomorpha. *Mol Biol Evol.* 26: 2397–2406.
Paps J, Baguñà J, Riutort M. 2009b. Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. *Proc R Soc B Biol Sci.* 276:1245–1254.
Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. 2009. How many bootstrap replicates are necessary? In: Batzoglou S, editor. RECOMB 2009, LNCS 5541. Berlin (Germany): Springer-Verlag. p. 184–200.
Perez Y, Müller CHG, Harzsch S. 2014. The Chaetognatha: an anarchistic taxon between Protostomia and Deuterostomia. In: Wägele JW, Bartolomaeus T, editors. Deep metazoan phylogeny: the backbone of the tree of life—new insights from analyses of molecules, morphology, and theory of data analysis. Berlin (Germany): De Gruyter. p. 49–77.
Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9:3.
Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, et al. 2009.

Phylogenomics revives traditional views on deep animal relationships. *Curr Biol.* 19:706–712.
Rieger RM, Tyler S. 1995. Sister-group relationship of Gnathostomulida and Rotifera-Acanthocephala. *Invertebr Biol.* 114:186–188.
Riesgo A, Andrade SC, Sharma P, Novo M, Perez-Porro A, Vahtera V, Gonzalez V, Kawauchi G, Giribet G. 2012. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Front Zool.* 9:33.
Rodríguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Curr Biol.* 17:1420–1425.
Rota-Stabelli O, Telford MJ. 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Mol Phylogenet Evol.* 48:103–111.
Rothe B, Schmidt-Rhaesa A. 2009. Architecture of the nervous system in two *Dactylopodola* species (Gastrotricha, Macrodasyida). *Zoomorphology* 128:227–246.
Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol.* 30:197–214.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.
Simmons MP. 2012a. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics* 28:208–222.
Simmons MP. 2012b. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. *Mol Phylogenet Evol.* 62:472–484.
Simmons MP, Freudenstein JV. 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. *Mol Phylogenet Evol.* 61: 177–191.
Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24:715–716.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
Struck TH. 2013. The impact of paralogy on phylogenomic studies—a case study on annelid relationships. *PLoS One* 8:e62892.
Struck TH. 2012. Phylogeny of Annelida. Handbook of Zoology Online [Internet]. Berlin (Germany): DeGruyter; [cited 2014 Apr 24]. Available from: http://www.degruyter.com/view/Zoology/bp_029147-6_1.
Struck TH, Fisse F. 2008. Phylogenetic position of Nemertea derived from phylogenomic data. *Mol Biol Evol.* 25:728–736.
Struck TH, Paul C, Hill N, Hartmann S, Hösel C, Kube M, Lieb B, Meyer A, Tiedemann R, Purschke G, et al. 2011. Phylogenomic analyses unravel annelid evolution. *Nature* 471:95–98.
Susko E, Leigh J, Doolittle WF, Bapteste E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol Biol Evol.* 23:1019–1030.
Tomer R, Denes AS, Tessmar-Raible K, Arendt D. 2010. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell* 142:800–809.
Turbeville JM. 1986. An ultrastructural analysis of coelomogenesis in the hoplonemertine *Prosorhochmus americanus* and the polychaete *Magelona* sp. *J Morphol.* 187:51–60.
Wey-Fabrizius AR, Herlyn H, Rieger B, Rosenkranz D, Witek A, Welch DBM, Ebersberger I, Hankeln T. 2014. Transcriptome data reveal syndermatan relationships and suggest the evolution of endoparasitism in Acanthocephala via an epizoic stage. *PLoS One* 9:e88618.
Witek A, Herlyn H, Ebersberger I, Mark Welch DB, Hankeln T. 2009. Support for the monophyletic origin of Gnathifera from phylogenomics. *Mol Phylogenet Evol.* 53:1037–1041.
Zhong M, Hansen B, Nesnidal MP, Golombek A, Halanych KM, Struck TH. 2011. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis for terebelliform annelids. *BMC Evol Biol.* 11: 369.

Peters et al. BMC Evolutionary Biology 2014, 14:52
http://www.biomedcentral.com/1471-2148/14/52

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data

Ralph S Peters[1†], Karen Meusemann[2,3†], Malte Petersen[2], Christoph Mayer[2], Jeanne Wilbrandt[2], Tanja Ziesmann[2], Alexander Donath[2], Karl M Kjer[4], Ulrike Aspöck[5,6], Horst Aspöck[7], Andre Aberer[8], Alexandros Stamatakis[8,9], Frank Friedrich[10], Frank Hünefeld[11], Oliver Niehuis[2], Rolf G Beutel[11] and Bernhard Misof[2*]

Abstract

Background: Despite considerable progress in systematics, a comprehensive scenario of the evolution of phenotypic characters in the mega-diverse Holometabola based on a solid phylogenetic hypothesis was still missing. We addressed this issue by *de novo* sequencing transcriptome libraries of representatives of all orders of holometabolan insects (13 species in total) and by using a previously published extensive morphological dataset. We tested competing phylogenetic hypotheses by analyzing various specifically designed sets of amino acid sequence data, using maximum likelihood (ML) based tree inference and Four-cluster Likelihood Mapping (FcLM). By maximum parsimony-based mapping of the morphological data on the phylogenetic relationships we traced evolutionary transformations at the phenotypic level and reconstructed the groundplan of Holometabola and of selected subgroups.

Results: In our analysis of the amino acid sequence data of 1,343 single-copy orthologous genes, Hymenoptera are placed as sister group to all remaining holometabolan orders, *i.e.*, to a clade Aparaglossata, comprising two monophyletic subunits Mecopterida (Amphiesmenoptera + Antliophora) and Neuropteroidea (Neuropterida + Coleopterida). The monophyly of Coleopterida (Coleoptera and Strepsiptera) remains ambiguous in the analyses of the transcriptome data, but appears likely based on the morphological data. Highly supported relationships within Neuropterida and Antliophora are Raphidioptera + (Neuroptera + monophyletic Megaloptera), and Diptera + (Siphonaptera + Mecoptera). ML tree inference and FcLM yielded largely congruent results. However, FcLM, which was applied here for the first time to large phylogenomic supermatrices, displayed additional signal in the datasets that was not identified in the ML trees.

Conclusions: Our phylogenetic results imply that an orthognathous larva belongs to the groundplan of Holometabola, with compound eyes and well-developed thoracic legs, externally feeding on plants or fungi. Ancestral larvae of Aparaglossata were prognathous, equipped with single larval eyes (stemmata), and possibly agile and predacious. Ancestral holometabolan adults likely resembled in their morphology the groundplan of adult neopteran insects. Within Aparaglossata, the adult's flight apparatus and ovipositor underwent strong modifications. We show that the combination of well-resolved phylogenies obtained by phylogenomic analyses and well-documented extensive morphological datasets is an appropriate basis for reconstructing complex morphological transformations and for the inference of evolutionary histories.

* Correspondence: b.misof.zfmk@uni-bonn.de
†Equal contributors
[2]Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Molekulare Biodiversitätsforschung (zmb), Adenauerallee 160, 53113 Bonn, Germany
Full list of author information is available at the end of the article

Background

Holometabola (or Endopterygota) are, given their evolutionary age, by far the most species-rich subgroup of insects (Hexapoda) and comprise more than 60% of all described metazoan species [1]. Within the Holometabola, the mega-diverse orders Coleoptera (beetles), Diptera (midges, mosquitos, and flies), Lepidoptera (moths and butterflies), and Hymenoptera (sawflies, bees, wasps, and ants) comprise together almost 800,000 species [2] and therefore more than 95% of the total species diversity of the entire lineage. The smaller orders are Neuroptera (lacewings), Megaloptera (alderflies and dobsonflies), Raphidioptera (snakeflies), Trichoptera (caddisflies), Mecoptera (scorpionflies and relatives), and Siphonaptera (fleas). Complete metamorphosis, which is characterized by the presence of a more or less inactive and non-feeding pupal stage between a feeding larva and a reproducing adult, is the most striking difference between Holometabola and other hexapods. Whereas the monophyly of Holometabola and of all its orders (with few exceptions, see below) has been consistently recovered (*e.g.*, [1,3]), the interordinal relationships are still insufficiently resolved. This impedes our understanding of the ancestral holometabolan morphology and life history and the modifications that occurred during the subsequent diversification of this highly successful lineage.

A reliable reconstruction of evolutionary transformations within Holometabola requires a well-founded hypothesis of the phylogenetic relationships of the major included groups. The first comprehensive reconstruction of holometabolan phylogenetic relationships was presented by Hennig [4], although a substantial contribution had already been made earlier by Hinton [5]. Alternative concepts to Hennig's proposal were presented by Rasnitsyn and Quicke [6] and Kukalová-Peck and Lawrence [7], with the main difference that Hymenoptera were not placed as sister group of Mecopterida (Diptera, Siphonaptera, and Mecoptera (= Antliophora), and Lepidoptera and Trichoptera (= Amphiesmenoptera)) (as in [4] and, *e.g.*, [1,8,9]), but as the first diverging extant holometabolan insect order. A distinctly different view was presented by Wheeler and colleagues [10] (see also [11,12]): they discussed a sister group relationship between Hymenoptera and Mecopterida (as in Hennig's concept), a sister group relationship between Strepsiptera and Diptera (Halteria), and paraphyletic Mecoptera, with the mecopteran Boreidae as sister group of Siphonaptera. Based on entirely new molecular and morphological datasets, Wiegmann et al. [13], McKenna and Farrell [14], and Beutel et al. [15] (see also [16]) congruently revived the view that Hymenoptera are sistergroup of all remaining Holometabola; Strepsiptera were recovered as closely related to Coleoptera, and Mecoptera were found monophyletic. Recently, these hypotheses gained additional support by a phylogenetic analysis of nucleotide sequence data from whole genome sequencing projects [17]. However, several interordinal relationships within Holometabola remained elusive. Despite remarkable progress, the genomic depth of published molecular sequence data, which potentially offers a plethora of phylogenetically informative characters, is still very low: large-scale transcriptome or genome data have been only available for representatives of less than half of all recognized holometabolan orders, with most studies so far dealing with model species. Consequently, the aim of our study was to present the first reconstruction of holometabolan relationships based on transcriptomic data of representatives of all currently recognized orders.

In this study, we address the following phylogenetic questions:

1. Are Hymenoptera the sister group of Mecopterida (Antliophora and Amphiesmenoptera) or of all other holometabolan insect lineages (*e.g.*, [4] *versus* [13])?
2. Are Neuropteroidea (Neuropterida, Coleoptera, and Strepsiptera) monophyletic? Neuropteroidea were found monophyletic by Wiegmann et al. [13] but not found by Wheeler et al. [10], Kukalová-Peck and Lawrence [7], and Beutel et al. [15].
3. Are Megaloptera monophyletic? and 4. Are Neuroptera and Megaloptera sister groups? Proposed relationships of the groups of Neuropterida (Megaloptera, Neuroptera, and Raphidioptera) are incongruent, and nearly all possible topological arrangements concerning this problem have been published over the last years (see, *e.g.*, [1,3,15,18-21]).
5. Are Coleopterida (Coleoptera and Strepsiptera) monophyletic? The whole genome-based analyses by Niehuis et al. [17] inferred Strepsiptera as sister group of Coleoptera, but did not include representatives of Neuropterida.
6. Are Mecopterida monophyletic? This group was neither found monophyletic by Kukalová-Peck and Lawrence [7] nor by some of the analyses in Beutel et al. [15], but was monophyletic in Wiegmann et al. [13], though not well supported.
7. What are the phylogenetic relationships within Antliophora? Contradicting phylogenetic relationships among Diptera, Mecoptera, and Siphonaptera have been published, and the monophyly of Mecoptera has been questioned (see above, and [8,10,13,15]).

In order to address the above questions, we generated transcriptomic data of at least one representative of each holometabolan order. For transcriptome sequencing, we selected species mostly characterized by plesiomorphic morphological character conditions and representing

taxa that presumably diverged early in the evolutionary history of each group (see [15]). In our molecular phylogenetic analyses, we used specific decisive datasets for each of our phylogenetic questions. Following the arguments put forth by Dell'Ampio et al. [22], a dataset is deemed "decisive" if information of each gene is available from each taxonomic group of interest and thus can contribute to resolving the relationships among these groups. In addition to maximum likelihood (ML) based tree inference, we applied Four-cluster Likelihood Mapping (FcLM) [23] to study potential incongruent signal in our datasets that might not be revealed by a phylogenetic multi-species tree.

We mapped a comprehensive set of morphological data [15] on the transcriptome-based phylogeny, and addressed the following issues regarding the evolutionary history of Holometabola:

- Major morphological features of the ancestral larva and the ancestral adult of Holometabola (groundplan) (*e.g.*, larval eyes, legs, prognathous *versus* orthognathous head; adult prognathous *versus* orthognathous head, size of pterothoracic segments, eyes)
- Ancestral larval and adult life habits of Holometabola (*e.g.*, diet, phytophagy/fungivory *versus* carnivory)
- Major transformations of larval and adult characters within Holometabola (*e.g.,* flight apparatus transformations: shift of segment and wing size, wing coupling mechanisms; modifications of oviposition strategy)
- Ancestral mode of ontogenetic development of Holometabola (*e.g.*, pupal characters)

In summary, we aimed to trace evolutionary changes of phenotypic features and to reconstruct groundplans for Holometabola and well-established clades within the Holometabola tree. An evolutionary history based on a solid phylogenetic background represents an important step toward a better understanding of the unparalleled diversification of this exceptional group of organisms.

Results and discussion

The phylogeny of Holometabola

We analyzed a total of 1,343 1:1 orthologous genes (*i.e.*, groups of orthologous sequences, also called ortholog groups (OGs)) and, by including also published data, data from a total of 88 species (Table 1). The seven specifically designed decisive datasets that we analyzed to address our seven phylogenetic questions each consisted of a subset of taxa and genes from the complete dataset, except for dataset 1 which is identical to the complete dataset. The seven questions, the taxonomic groups that we selected as relevant for answering the questions, and the numbers of species and OGs for each dataset are shown in Table 2. For each dataset we performed 1) ML tree reconstruction, and 2) Four-cluster Likelihood Mapping (FcLM) (see Table 3). Results are summarized in Figure 1 (see Additional file 1: Figures S1-S7 for presence and absence of genes in the datasets, Additional file 2: Figures S8-S15 for the full phylogenetic trees, and Additional file 3: Figures S17-S25 for the full results of the FcLM).

The analysis of dataset 1 yielded Hymenoptera as sister group to all remaining holometabolan orders in both ML tree reconstruction and FcLM (Table 3, Figure 1). This relationship had already been recovered in several multiple gene studies (*e.g.*, [13,14]), and based on whole

Table 1 Holometabola species, for which data were newly sequenced

Order	Family	Species	No. of contigs	No. of OGs
Hymenoptera	Xyelidae	*Xyela alpigena* (Strobl, 1895)	9,931	471
Raphidioptera	Raphidiidae	*Raphidia ariadne* Aspöck & Aspöck, 1964	29,636	983
Neuroptera	Nevrorthidae	*Nevrorthus apatelios* Aspöck, Aspöck & Hölzel, 1977	17,673	695
Megaloptera	Sialidae	*Sialis lutaria* (Linnaeus, 1758)	14,200	801
Megaloptera	Corydalidae	Corydalinae sp.	60,455	1,109
Coleoptera	Cupedidae	*Priacma serrata* (Leconte, 1861)	18,808	868
Coleoptera	Carabidae	*Carabus granulatus* (Linnaeus, 1758)	55,582	1,159
Strepsiptera	Mengenillidae	*Mengenilla moldrzyki* Pohl et al., 2012	60,642	999
Lepidoptera	Micropterigidae	*Micropterix calthella* (Linné, 1761)	137,093	969
Trichoptera	Philopotamidae	*Philopotamus ludificatus* McLachlan, 1878	24,628	914
Diptera	Tipulidae	*Tipula maxima* Poda, 1761	24,724	938
Siphonaptera	Pulicidae	*Archaeopsylla erinacei* (Bouché, 1835)	35,270	1,191
Mecoptera	Nannochoristidae	*Nannochorista philpotti* (Tillyard, 1917)	44,935	1,212

Shown are taxonomic classification, number of contigs after assembly (only contigs longer than 200 bp after removal of suspicious sequences are considered, according to the NCBI guidelines (VecScreen)), and number of assigned single-copy orthologous genes in the complete dataset (after redundancy and outlier check, see Methods section).

Table 2 The seven datasets, designed to address seven phylogenetic questions

Dataset	Addressed phylogenetic question	Covered subgroups/FcLM clusters (4 clusters per analysis)	No. of species	No. of OGs	Alignment length (aa)	Coverage [%] all species	Coverage [%] addressed groups
Dataset 1 (complete dataset)	Position of Hymenoptera?	1) Hymenoptera	88	1,343	662,107	61.1	100
		2) outgroup taxa					
		3) Mecopterida					
		4) Neuropteroidea					
Dataset 2	Are Neuropteroidea monophyletic?	1) Neuropterida	71	1,303	643,051	65.0	100
		2) Mecopterida					
		3) Coleopterida					
		4) Hymenoptera					
Dataset 3	Are Megaloptera monophyletic?	1) Raphidioptera	4	358	174,065	100	100
		2) Corydalidae					
		3) Sialidae					
		4) Neuroptera					
Dataset 4	Are Neuroptera and Megaloptera sister groups?	1) Raphidioptera	71	540	242,820	72.9	100
		2) Megaloptera					
		3) Neuroptera					
		4) remaining holometabolans					
Dataset 5	Are Coleopterida monophyletic?	1) Neuropterida	71	972	505,528	66.2	100
		2) Strepsiptera					
		3) Coleoptera					
		4) remaining holometabolans					
Dataset 6a	a) Are Mecopterida monophyletic? or	a) 1) Antliophora	71	1,343	662,107	64.3	100
		2) Amphiesmenoptera					
Dataset 6b	b) Are Antliophora + Coleopterida monophyletic?	3) Neuropteroidea					
		4) remaining holometabolans					
		b) 1) Antliophora					
		2) Amphiesmenoptera					
		3) Coleopterida					
		4) remaining holometabolans					
Dataset 7	Relationships within Antliophora?	1) Diptera	71	1,101	557,276	66.5	100
		2) Siphonaptera					
		3) Mecoptera					
		4) remaining holometabolans					

For each dataset, we selected four taxonomic groups (clusters), assigned species to one of the groups, and extracted only those ortholog groups (OGs) that contained a sequence of at least one representative of each group. All species that were not assigned to either of the groups were excluded. Coverage [%] all species: Coverage of the dataset in terms of presence of OGs considering all species. Coverage [%] addressed groups: Coverage of the dataset in terms of presence of OGs considering the four groups defined for each dataset, which is, by definition, 100%.

genome data but a limited taxon sampling [17]. Previously published analyses of morphological data yielded contradictory results, such as for instance Hymenoptera + Mecopterida in Beutel and Gorb [9] *versus* Hymenoptera + remaining holometabolan orders in Beutel et al. [15]. Potential problems of topological artifacts in these analyses that are caused by convergent reductions in many morphological character systems were discussed in detail by Friedrich and Beutel [24] and Beutel et al. [15]. The placement of Hymenoptera as sister group to all remaining holometabolan orders implies that presumptive synapomorphies of Hymenoptera and Mecopterida (*e.g.*, single claw of larvae, sclerotized sitophore plate of adults; see [1]) are in fact homoplasies.

Our analyses of dataset 2 yielded monophyletic Neuropteroidea (*i.e.*, a clade comprising Neuropterida, Coleoptera,

Table 3 FcLM Results

Dataset	Possible unambiguous topologies	No. of drawn quartets	Support T1 [%] 1,2 | 3,4	Support T2 [%] 1,3 | 2,4	Support T3 [%] 1,4 | 2,3
Dataset 1 (complete dataset)	**T1: Hymenoptera,outgroup taxa | Mecopterida, Neuropteroidea**	142,800	**83**	8	8
	T2: Hymenoptera, Mecopterida | outgroup taxa, Neuropteroidea				
	T3: Hymenoptera, Neuropteroidea | outgroup taxa, Mecopterida				
Dataset 2	T1: Neuropterida, Mecopterida | Coleopterida, Hymenoptera	20,160	8	**80**	11
	T2: Neuropterida, Coleopterida | Mecopterida, Hymenoptera				
	T3: Neuropterida, Hymenoptera | Mecopterida, Coleopterida				
Dataset 3	T1: Raphidioptera, Corydalidae | Sialidae, Neuroptera	1	0	0	**100**
	T2: Raphidioptera, Sialidae | Corydalidae, Neuroptera				
	T3: Raphidioptera, Neuroptera | Corydalidae, Sialidae				
Dataset 4	T1: Raphidioptera, Megaloptera | Neuroptera, remaining holometabolans	134	25	1	**72**
	T2: Raphidioptera, Neuroptera | Megaloptera, remaining holometabolans				
	T3: Raphidioptera, remaining holometabolans | Megaloptera, Neuroptera				
Dataset 5	T1: Neuropterida,Strepsiptera | Coleoptera,remaining holometabolans	1,220	6 (8)	**55 (53)**	38 (38)
	T2: Neuropterida, Coleoptera | Strepsiptera,remaining holometabolans				
	T3: Neuropterida,remaining holometabolans | Strepsiptera, Coleoptera				
Dataset 6a	**T1: Antliophora, Amphiesmenoptera | Coleopterida, remaining holometabolans**	80,640	**80**	14	5
	T2: Antliophora, Coleopterida | Amphiesmenoptera, remaining holometabolans				
	T3: Antliophora, remaining holometabolans | Amphiesmenoptera, Coleopterida				
Dataset 6b	**T1: Antliophora, Amphiesmenoptera | Coleopterida, remaining holometabolans**	57,600	**79**	15	5
	T2: Antliophora, Coleopterida | Amphiesmenoptera, remaining holometabolans				
	T3: Antliophora, remaining holometabolans | Amphiesmenoptera, Coleopterida				
Dataset 7	T1: Diptera, Siphonaptera | Mecoptera, remaining holometabolans	1,034	0	0	**100**
	T2: Diptera, Mecoptera | Siphonaptera, remaining holometabolans				
	T3: Diptera, remaining holometabolans | Siphonaptera, Mecoptera				

For the four groups (clusters) that were selected for each of the seven datasets, three unambiguous topologies are possible (see Additional file 4, Chapter 3, and Additional file 3: Figure S16). For details which species are included in the groups for each dataset see Additional file 12. The number of drawn quartets is the product of the numbers of species in each group. In bold print: Topology that gained the highest support (support [%]: relative amount of quartets which show predominant support for either T_1, T_2 or T_3). Results of partitioned analyses of dataset 5 in parentheses.

and Strepsiptera) with maximal support in the ML tree reconstruction and strong support in the FcLM (Table 3, Figure 1). Neuropteroidea was not supported as a clade in Beutel et al. [15], but was found monophyletic in many previous studies [1,8,9,13,14,25], even though in most cases with weak or without support.

We did not find any signal for paraphyletic Megaloptera as discussed by Beutel et al. [15] and Winterton et al. [26] (dataset 3, Table 3, Figure 1). Within Neuropterida, our ML analyses maximally supported a sister group relationship between Raphidioptera and Neuroptera + Megaloptera, which was also supported by more than 2/3 of all quartets in the FcLM (dataset 4, Table 3, Figure 1). Phylogenetic relationships among neuropterid orders have been discussed controversially with two alternative hypotheses: Raphidioptera + Megaloptera being

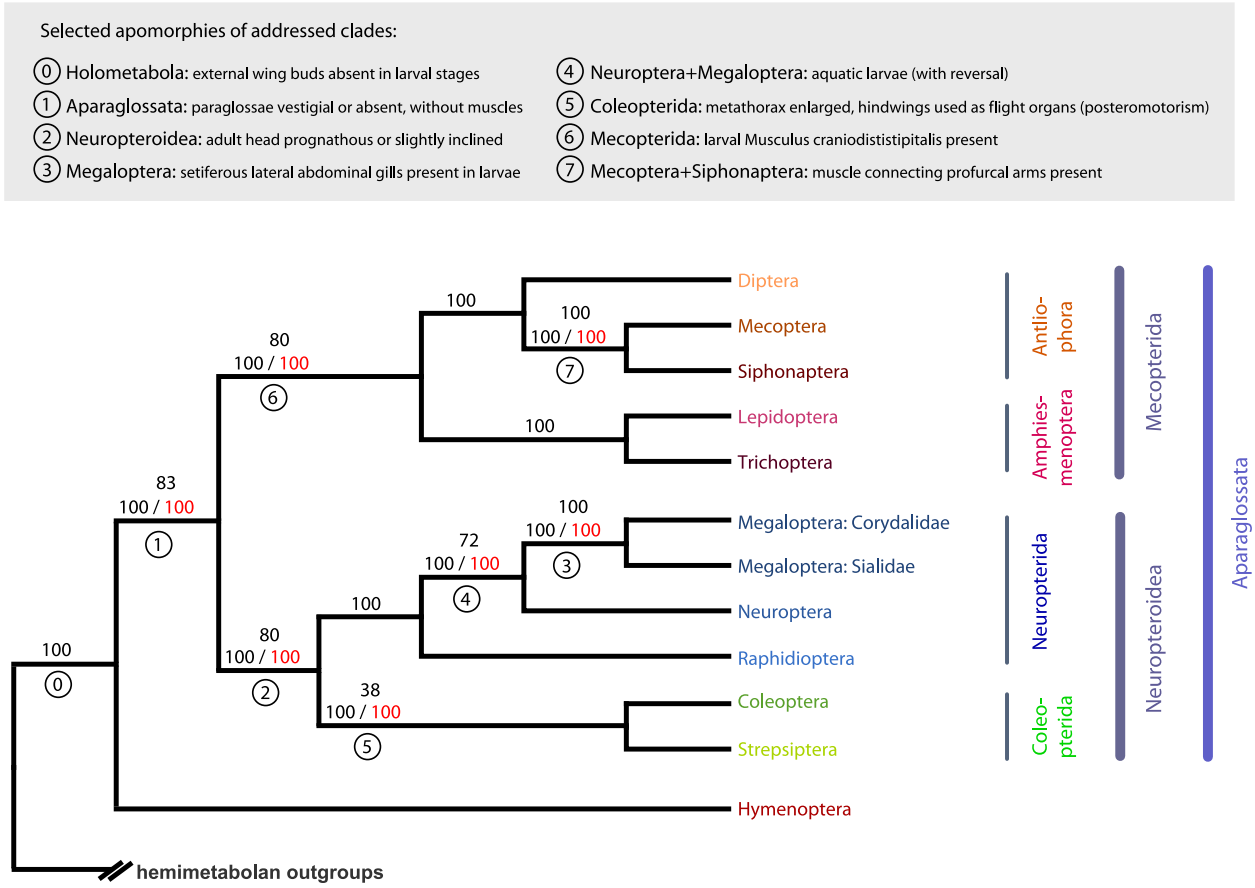

Figure 1 Combined and simplified cladogramm of holometabolan insect relationships, with selected autapomorphies for the clades addressed in this study. The topology is taken from the ML tree inferred from dataset 1 (*i.e.*, the complete datamatrix). (1) Bootstrap support (BS) (bottom, black) is derived from 72 bootstrap replicates (MRE-based bootstopping criterion) of dataset 1. (2) BS values for the specific phylogenetic relationship (bottom, red) are derived from ML tree inferences from the seven specific decisive datasets 1 to 7. (3) relative support [%] values for the specific phylogenetic relationship (top) are derived from the Four-cluster Likelihood Mapping (FcLM) with the seven specific decisive datasets. Apomorphies are selected from the full lists of reconstructed groundplan characters (see Additional file 4, Chapter 5).

monophyletic (*e.g.*, [8,9,13-15,27]) or Neuroptera + Megaloptera being monophyletic (*e.g.*, [18,19,25,28,29]). Our results strongly support the latter hypothesis.

Analysis of dataset 5 yielded ambiguous results with respect to a possible clade comprising Coleoptera and Strepsiptera (Coleopterida) (Table 3, Figure 1). Resolving this longstanding problem is difficult due to the extremely modified morphology (*e.g.*, [30]) and the distinctly derived genomic features [17,31] of the endoparasitic Strepsiptera ("the Strepsiptera problem", [1]; "insects from outer space", [32]). In most recent contributions, evidence was found for monophyletic Coleopterida (*e.g.*, [13-15,17]). However, the studies based on molecular data remained ambiguous in their results. Coleopterida were not supported by all datasets analyzed by McKenna and Farrell [14]. The results of Wiegmann et al. [13] were based on a relatively small set of genes and showed only weak support for this clade. Niehuis et al. [17] analyzed whole genome nucleotide sequences of holometabolous insects and found well-supported Coleopterida but the taxon sampling did not include any neuropterid orders. In our study, Coleopterida is supported in the ML tree (with maximal bootstrap support), but not in the FcLM analyses (Table 3, Figure 1). In the ML tree, Strepsiptera are placed within Coleoptera (like in some of the trees of McKenna and Farrell [14]), however, with poorly supported relationships (Additional file 2: Figure S12). We further analyzed whether the incongruence between ML tree reconstruction and FcLM analyses vanished considering partitioned ML and FcLM analyses using different models on different partitions. Partitioned analyses might reduce potential model misspecifications and might yield congruent topologies. However, the incongruence between ML and FcLM analyses did not disappear (Table 3, Additional files 2 and 3). This implies that model misspecifications due to unpartitioned analyses are not the source of incongruence (see also [22] and discussion therein). Apparently, the data and analytical procedures of our study did not yield an unambiguous solution of the

question whether or not Coleopterida is a monophyletic group. However, evidence from morphology clearly suggests monophyletic Coleopterida (see also [17]) as the most plausible result.

In order to test the monophyly of Mecopterida, a clade comprising Amphiesmenoptera (Lepidoptera + Trichoptera) and Antliophora (Diptera + Siphonaptera + Mecoptera), we analyzed two versions of dataset 6 to account for two possible hypotheses (dataset 6a, b; Tables 2 and 3). Both analyses recovered monophyletic Mecopterida with strong support (Table 3, Figure 1). Monophyletic Mecopterida, as proposed by Hinton [5] under the name Panorpoidea (or panorpoid complex), was not well supported in Kjer et al. [25] and Wiegmann et al. [13], and only supported in the Bayesian analyses of morphological characters in Beutel et al. [15]. Niehuis et al. [17] found tentative support for this clade based on whole genome data but the incomplete taxon sampling – genomes of Neuropterida, Trichoptera, Siphonaptera, and Mecoptera have not been sequenced yet – diminished the decisiveness of this dataset concerning the question of monophyletic Mecopterida.

Our analyses clearly corroborated the monophyly of Amphiesmenoptera (Trichoptera + Lepidoptera) (Figure 1). However, we did not test this hypothesis with a specifically designed dataset because it has never been seriously disputed [1].

Within Antliophora, which showed maximal bootstrap support in the ML tree, we found a sister group relationship of Mecoptera and Siphonaptera, also with maximal bootstrap support and with maximal support in the FcLM (dataset 7, Table 2, Figure 1). This result corroborates views put forward by Beutel and Gorb [8], McKenna and Farrell [14], and Wiegmann et al. [13], though the clade Mecoptera + Siphonaptera was not well supported in the latter study. A sister group relationship between Diptera and Siphonaptera as retrieved in Beutel et al. ([15], see discussion therein) is highly unlikely based on our analyses.

With this study, we do not contribute to the question whether Mecoptera are a monophyletic group as only one species, *Nannochorista philpotti,* was part of our taxon sampling. However, morphological data [15] and analyses of nine nuclear genes [14] strongly suggest that Mecoptera indeed form a monophyletic group.

In summary, we inferred a solid phylogenetic backbone of Holometabola, with three maximally supported mega-diverse clades Hymenoptera, Neuropteroidea, and Mecopterida, with approximately 135,000, 370,000, and 300,000 described species, respectively. For the well-defined unit comprising Neuropteroidea and Mecopterida we suggest the name Aparaglossata (Figure 1). The name refers to the loss of the paraglossae, one of the most conspicuous apomorphies of the group (see below and Table 4).

Our compilation of molecular sequence datasets and our design of the phylogenetic analysis exhibit some major differences compared to earlier studies on the phylogeny of Holometabola. Specifically, i) we used a massive amount of data generated with Illumina Next Generation Sequencing (Table 1). ii) We ensured decisiveness of our datasets by specifically designing datasets for each of our seven research questions (Table 2) (see [22]). Decisiveness means that all genes included in a dataset are covered by at least one representative of all taxonomic groups that are relevant for the specific phylogenetic relationship under study. Accordingly, each dataset has a coverage of 100% in terms of presence of genes, with respect to the relevant taxonomic groups. By ensuring decisiveness, we alleviate the potentially misleading effects of missing data. Missing data can lead to inference of highly supported but wrong topologies (see [22]). iii) We performed FcLM [23] for each of our seven datasets (Table 3). We re-implemented FcLM in RAxML to cope with these large-scale data matrices and complemented the method by newly-written scripts that map respective results into 2D simplex graphs. Bootstrap support in phylogenetic trees alone is of limited conclusiveness in analyses of very large datasets [22,34]. FcLM is a method to identify possible support for alternative topologies in a dataset, *i.e.*, a method to display incongruent signal that might not be observable in phylogenetic trees. This study is the first to apply FcLM to large phylogenomic supermatrices. Finally, iv) we checked all datasets for rogue taxa. Rogue taxa are taxa that assume multiple phylogenetic positions in a set of bootstrap trees. They decrease resolution and/or support, for example, when building bootstrap consensus trees. Removing rogues may produce a more informative bootstrap consensus tree [35,36] (see Additional file 4, Chapter 4). All our datasets were free of rogues.

With a compilation of datasets as presented here (*i.e.*, by extracting the maximum number of genes that can contribute to resolving the phylogenetic relationship in question) we also ensured that inferred topologies were not based on an arbitrary selection of genes with respect to their inherent phylogenetic signal. Dell'Ampio et al. [22] showed that the selection of genes – if not driven by considerations concerning decisiveness of a dataset – can generate topologically different trees that may nonetheless all exhibit high support. Furthermore, Simon et al. [37,38] showed that genes involved in different biological pathways can support different topologies for a specific phylogenetic relationship. It can therefore be concluded that phylogenetic trees inferred from studying only a set of few to several genes are easily biased and thus might not reflect the correct species tree. While the currently best approach to address this problem is to include the maximum feasible amount of potentially informative data, we will have to further disentangle the

Table 4 Selection of groundplan characters and apomorphies of Holometabola and of those holometabolan subgroups whose phylogenetic relationships were addressed in this study and whose monophyly was confirmed

Taxon	Characters
Holometabola	•* Larval head orthognathous
	•* Larval compound eyes simplified but present
	•* Ocelli absent in larvae
	•* Larval tentorium X-shaped
	•* Retractile larval abdominal prolegs absent
	• Larval cerci absent (possible reversal in Strepsiptera [homology uncertain])
	•* Adult head orthognathous
	• Meso- and metasternum invaginated
	• Meso- and metacoxae closely adjacent medially
	• Appearance of fully developed compound eyes including external apparatus in the pupal stage (reversal in Strepsiptera)
	• External wing buds absent in larval stages (partial reversal in Strepsiptera)
Aparaglossata (Holometabola excluding Hymenoptera)	• Larval head prognathous
	• Well-developed larval stemmata
	• Larval tentorium H-shaped
	• Paraglossae vestigial or absent, without muscles
	• Ventral sclerites of segment VIII (gonocoxae and gonapophyses) indistinct (reversals within Neuropterida)
Neuropteroidea § (Neuropterida and Coleopterida)	• Adult head prognathous or slightly inclined (reversal in Neuroptera)
Megaloptera §	• Sensorium on antepenultimate larval antennomere
	• Larval salivary duct strongly narrowed, without recognizable lumen
	• Setiferous lateral abdominal gills present in larvae
Neuroptera + Megaloptera	• Mesothoracic prealare present (also in Amphiesmenoptera)
	• Muscular connection between metafurcal arm and epimeral apophysis
	• Aquatic larvae (with reversal)
Coleopterida (Coleoptera and Strepsiptera)	• Antenna with 9 flagellomeres or less
	• Pronotum and propleuron partly or completely connected (also in Diptera)
	• Metathorax enlarged, hind wings used as flight organs (posteromotorism)
	• Membranous area between mesoscutellum and mesopostnotum present
Mecopterida (Antliophora and Amphiesmenoptera)	• Larval dorsal tentorial arm strongly reduced or absent
	• Less than 3 larval antennomeres (reversal to 3 in some groups)
	• Larval galea and lacinia extensively or completely fused (also missing as separate structures in Neuroptera and Strepsiptera)
	• Larval Musculus craniodististipitalis present
Siphonaptera + Mecoptera §	• Muscle connecting profurcal arms (Musculus profurca-spinalis) present
	• Acanthae of proventriculus close-set, prominently elongated

Plesiomorphic groundplan characters are marked with an asterisk *. For a full list and for apomorphies found for additional subgroups see Additional file 4, Chapter 5. Characters apply to adults if not mentioned otherwise. For groups marked with § behind taxon name, no selection but rather all obtained apomorphies are listed. Groundplan characters and apomorphies were inferred from the morphological datamatrix of Beutel et al. [15] and the interordinal topology of the ML tree of dataset 1 by formal character mapping in Mesquite [33].

contributing factors of topological incongruences in datasets (see also [22]).

Phylogenetic studies exclusively based on morphology (*e.g.*, [15,24]) also yielded problematic groupings in some cases. The authors addressed and discussed apparent artifacts that were mainly caused by parallel reductions in character complexes (*e.g.*, the flight apparatus). However, the problems turned out as intractable given the data and analytical procedures at hand [15]. With our molecular datasets we were able to provide reliable solutions for most interordinal phylogenetic relationships within Holometabola (Figure 1, and above). For tracing

evolutionary changes on the phenotypic level we used the most extensive morphological dataset presently available, including 356 characters of representatives of all holometabolan orders and of carefully selected outgroup taxa [15]. The characters were mapped onto the transcriptome-based phylogeny in a formal approach (see Methods section for details). This allowed us to trace and re-interpret evolutionary changes of numerous characters and to conduct parsimony-based groundplan reconstructions for all clades of the tree (see "The evolution within Holometabola" below).

The evolution within Holometabola

Larvae and development

Our phylogenetic results suggest that the ancestral larva of Holometabola was terrestrial, orthognathous, equipped with moderately simplified but distinctly developed compound eyes, and well developed thoracic legs. Abdominal prolegs and cerci were absent (Figure 2). The muscle system was generally well developed. Distinct simplifications of the antennae and labial endite lobes and associated muscles are larval autapomorphies of Holometabola. The orthognathous head in the groundplan suggests that the earliest holometabolan larvae were feeding externally on plant material or fungi and not burrowing in substrate or penetrating narrow crevices (*e.g.*, under bark).

The ancestral aparaglossatan larva was likely prognathous and equipped with stemmata. Whether these larvae were of the agile campodeid type, like the larvae of many beetles (*e.g.*, Adephaga, Myxophaga [*partim*], Staphylinoidea), Strepsiptera (first instar), Neuropterida, and some groups of Trichoptera (*e.g.*, Rhyacophilidae), remains unclear. It is conceivable that this larval type is an apomorphic condition characterizing Neuropteroidea, with parallel evolution in Trichoptera. Prognathism is often linked with carnivorous feeding habits (Neuropterida, Adephaga, and some polyphagan subgroups), but can also be related with penetrating narrow crevices or burrowing in substrates, as it is the case in the wood-associated larvae of Archostemata (Coleoptera), but also in early lepidopteran lineages (*e.g.*, [1]). Thus, it is unclear whether or not the ancestral aparaglossatan larvae were predaceous. Larvae of Mecopterida display some simplifications (tentorium and antennal segments), and a distinct trend towards reductions characterizes antliophoran larvae, especially those of Siphonaptera and Diptera. Both have entirely lost their thoracic legs (distinctly shortened in Mecoptera) and are characterized by simplifications of cephalic structures, especially of the muscle system [39]. This reflects the widespread larval life history in Antliophora, with larvae living in the upper soil layer, leaf litter, moist substrates, or different water bodies, feeding mainly on soft substrates or small particles. The important question whether ancestral antliophoran larvae were terrestrial (Lepidoptera, Mecoptera, Siphonaptera, Mecoptera excl. Nannochoristidae, Diptera *partim*) or aquatic (Trichoptera, Nannochoristidae, Diptera *partim*) remains ambiguous.

Our phylogenetic results clearly indicate that a typical holometabolous development with larvae completely lacking external wing buds ("endopterygote insects") and also lacking cerci belongs to the groundplan of Holometabola (see also [1,17]). The conditions characterizing strepsipteran primary larvae (abdominal segment XI and cerci present) and secondary larvae (external wing buds recognizable as external convexities) are apparently the result of reversals, like the early appearance of the prospective compound eyes (see [17]). Largely immobilized pupae with immobilized mandibles (pupa adectica) have almost certainly evolved several times independently. It appears likely that a mobile pupa with movable mandibles as it is characteristic for Raphidioptera is ancestral for Holometabola even though this is not confirmed by a formal character analysis.

Adults and egg deposition

The ancestral holometabolan adult apparently differed only slightly from the neopteran groundplan (Neoptera:

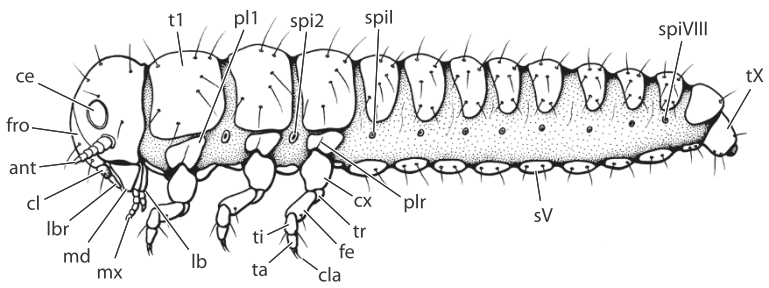

Figure 2 Illustration of reconstructed groundplan larva of Holometabola. The putative groundplan larva was orthognathous, and equipped with simplified but distinctly developed compound eyes, and well developed thoracic legs. Abdominal prolegs and cerci were absent. For a list of larval and adult groundplan characters of Holometabola, see Table 4. ce: compound eye. fro: frons. ant: antenna. cl: clypeus. lbr: labrum. md: mandible. mx: maxille. lb: labium. t1: tergite of first thoracic segment. pl1: pleurite of first thoracic segment. spi2: spiracle of second thoracic segment. plr: pleural ridge. cx: coxa. tr: trochanter. fe: femur. tib: tibia. ta: tarsus. cla: claw. spil: spiracle of first abdominal segment. sV: sternite of fifth abdominal segment. spiVIII: spiracle of eighth abdominal segment. tX: tergite of tenth abdominal segment.

all winged insects except Odonata and Ephemeroptera). Cephalic structures, the entire muscle system, the flight apparatus, and abdominal structures appear largely unmodified [15,23,39,40]. The most profound apomorphies in adult holometabolan insects are related to the invagination of the pterothoracic sternites (*e.g.*, closely adjacent meso- and metacoxae) [24]. Our data do not lead to a reliable assessment of ancestral feeding habits of holometabolan adults, but it is apparent that feeding in the adult stage played a minor role compared to feeding in the larval stages. Exceptions to this rule are for instance predaceous beetles (*e.g.*, Dytiscidae and Carabidae) with a very rapid postembryonic development and long-lived adults.

Distinct morphological character transformations characterize the rise of Aparaglossata: the reduction of the labial endite lobes (paraglossae), including muscles, the distinct modification of the orthopteroid ovipositor, and possibly the reduced number of Malpighian tubules (also in Acercaria (true bugs, psocopterans, lice, and relatives)) [15,41]. Our results do not allow for an unambiguous reconstruction of the ancestral condition of the flight apparatus for Holometabola and Aparaglossata. It appears plausible that approximately equally sized pterothoracic segments (as in Neuropterida, early lepidopteran lineages, and Mecoptera) are plesiomorphic for Aparaglossata, but the reconstruction of the ancestral state of this character in the formal analysis remained ambiguous. As pointed out above, the question whether or not Coleopterida is a monophylic group is not completely settled. However, it appears plausible to assume that posteromotorism evolved only once in a common ancestor of Strepsiptera and Coleoptera, with a suite of related features, such as the size reduction of the mesothorax, a distinct reduction of the mesothoracic muscle system [42], and an increased size of the metathorax. A distinct anteromotorism as it is present in Hymenoptera, Trichoptera, "higher" Lepidoptera, and Diptera is possibly ancestral in Holometabola, but it is conceivable that this condition has evolved (secondarily?, see above) several times independently (*e.g.*, almost equally sized pterothoracic segments in non-glossatan Lepidoptera).

Wing coupling mechanisms have apparently evolved independently in Hymenoptera (hamuli as an autapomorphy of the order, see Additional file 4, Chapter 5), Trichoptera, Lepidoptera, and some families of Neuroptera (different mechanisms occur in these orders).

The primary mode of egg deposition in Holometabola was very likely endophytic, as it can be assumed for the groundplan of Hymenoptera ("Symphyta"). This mode of egg deposition is arguably maintained in the groundplan of Neuropteroidea. Raphidioptera have a modified, elongated ovipositor which they use to deposit eggs under bark or into ground litter. This resembles egg deposition as assumed for the groundplan of Holometabola and Hymenoptera; however, it might also be a derived character. The complete or nearly complete reduction of elements of the primary ovipositor is a characteristic of Mecopterida and obviously related with superficial egg-deposition or oviposition in soft substrates. Our results mostly confirm an evolutionary scenario for the female postabdomen and egg-deposition as outlined in detail in Hünefeld et al. [41].

Conclusions

Our transcriptome-based phylogenetic results allowed a reconstruction of transformations of morphological characters of larvae and adults. To summarize our findings, we show a hypothesized ancestral holometabolan larva in Figure 2, and a selection of adult and larval groundplan features in Table 4 (see Additional file 4, Chapter 5 for a full list). The ancestral state of the adult thorax remained ambiguous. Three main holometabolan types are shown in Figure 3 (and in Additional file 5 as 3D pdf). A selection of apomorphic features of the major subgroups of Holometabola whose phylogenetic origins have now been elucidated is presented in Table 4 (see Additional file 4, Chapter 5 for a full list).

For the first time in insect systematics a scenario for transformations on the phenotypic level is based on a strictly formal procedure, using a well-documented comprehensive morphological data-set in combination with analyses of phylogenomic data. Our combined approach may lead to a new level of reciprocal enlightenment between researchers with a main focus on morphology and molecular data, respectively, and eventually to new and well-founded insights into the evolution of Hexapoda and other groups of organisms.

Methods

Data acquisition

Our study included a total of 88 species: 71 holometabolan species, and 17 species belonging to different hemimetabolous lineages for outgroup comparison. Of these, we generated transcriptomic data *de novo* for 13 holometabolan species. From all remaining species, we used published transcriptomic data or the transcripts of the official gene set (OGS) if the genome of a species is already sequenced (see below).

The 13 holometabolan species (at least one representative of each order) with newly generated transcriptomic data are listed in Table 1 (for details see Additional file 6, Table S1). Extraction of RNA, cDNA library construction, library normalization, sequencing of 12.5 million paired end reads (~ 2.5 Gigabases raw reads per species) using the Illumina Technique (Hiseq 1000), and sequence processing (vector-clipping, trimming and soft-masking of raw reads, and assembly into contigs) were done by LGC

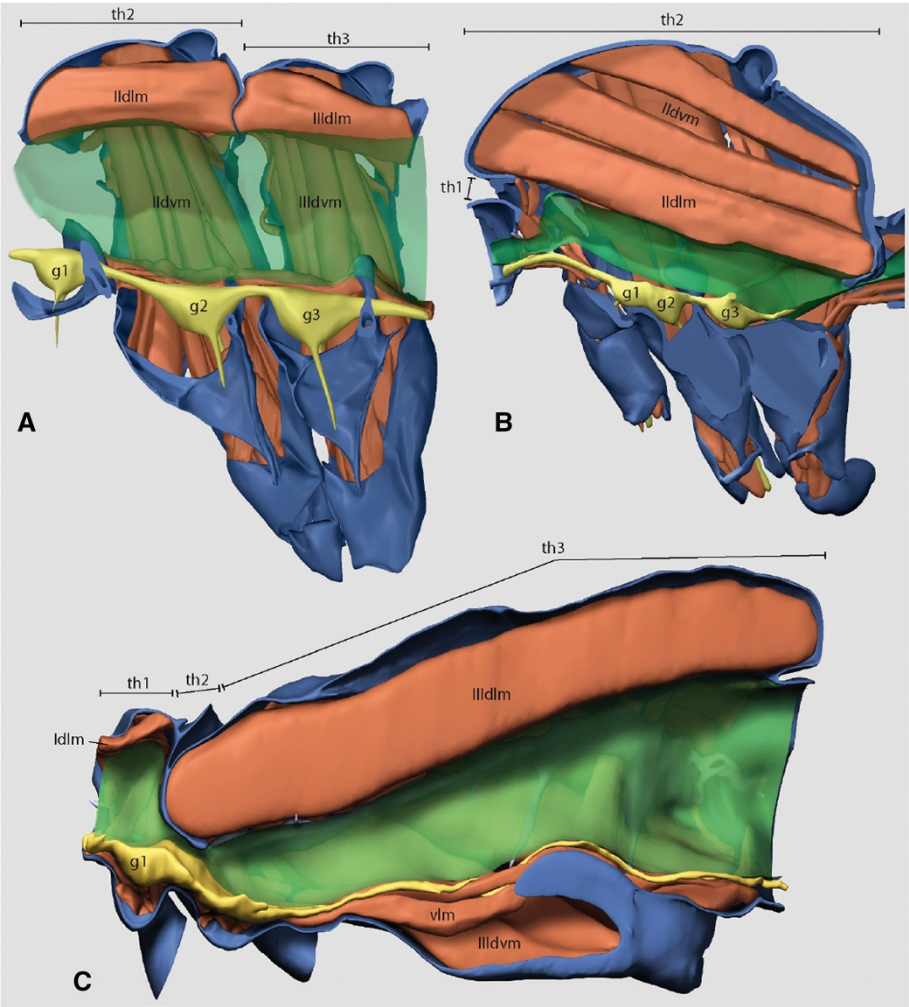

Figure 3 Three holometabolan adult thorax states. A) A thorax with approximately equally sized pterothoracic segments is possibly ancestral for Aparaglossata (Figure shows thorax of *Nannochorista neotropica* (Mecoptera, Nannochoristidae); prothorax not shown.). **B)** shows a thorax of taxa with anteromotorism, *i.e.*, flight with mainly the fore wings (*e.g.*, Hymenoptera, Trichoptera, "higher" Lepidoptera, and Diptera; figure shows *Ptychoptera* sp. (Diptera, Ptychopteridae)). This state is possibly ancestral for Holometabola. However, the reconstruction of the ancestral state of this character in the formal analysis remained ambiguous for Holometabola and Aparaglossata. **C)** shows a thorax of taxa with posteromotorism, *i.e.*, flight with the hind wings (Coleoptera and Strepsiptera; figure shows *Mengenilla moldrzyki* (Strepsiptera, Mengenillidae)). red: muscles. blue: sceleton. green: gut. yellow: nerves. Numerals refer to thoracic segments. th: thorax segment. g: ganglion. dlm: dorsal longitudinal muscle. dvm: dorso-ventral muscle. vlm: ventral longitudinal muscle (not visible in **A** and **B)**. A 3D version of this figure can be found as Additional file 5 (Click on image to activate animation).

Genomics, Berlin, Germany (see Additional file 4, Chapter 1, and Additional files 6 and 7: Tables S1 and S2 for details). All raw nucleotide sequence reads are deposited at the NCBI Sequence Read Archive (SRA). The corresponding nucleotide assemblies have been deposited at the NCBI's Transcriptome Sequences Database (TSA) (Umbrella project ID PRJNA176423). For further details and accession numbers, please refer to Additional file 4, Chapter 1, and Additional file 7: Table S2.

Nucleotide sequence assemblies of published transcriptome data were obtained from the Deep Metazoan Phylogeny (DMP) database (http://www.deep-phylogeny.org/), NCBI's Transcriptome Sequences Database (TSA) and from various web sources of species whose official gene set was available. We only used species with more than 3,000 available contigs (status: November 2012) (Additional file 8: Table S3).

Orthology assignment

We mapped the transcripts to a set of 1,343 ortholog groups (OGs), *i.e.*, a set of genes that have been identified as single-copy orthologs in 14 reference species (13 insects, 1 crustacean) in OrthoDB 4 (http://cegg.unige.ch/orthodb4/) (see Additional file 9: Table S4 for reference

species, and Additional file 10: Table S5 for included orthologs; for details on the design of the ortholog reference set see Additional file 4, Chapter 2). Orthology of transcripts was assigned using HaMStRad, a modified version of HaMStR v.8 [43] (see Additional file 4, Chapter 2 for details on modifications). The modified program files are available at https://github.com/mptrsen/HaMStRad (Status: March 2013). HaMStRad maps transcripts to a set of OGs using hidden Markov models and the best reciprocal hit criterion. We ran HaMStRad with the following settings: (i) the E-value cut-off for the pHMM search was 1e-5, (ii) the reciprocity criterion was considered fulfilled if the candidate OG was found as best hit in at least one of the 14 reference species during the reciprocal best hit search (RBH) (*–relaxed* option), (iii) in case of multiple transcripts being assigned to a given OG, the best set of non-overlapping transcripts was chosen while non-overlapping transcripts are automatically concatenated (*–representative* option). Transcripts that were assigned to more than one OG were removed from the dataset using Perl scripts (available upon request) (redundancy check). Furthermore, we removed terminal stop codons and masked internal stop codons with 'X'.

Multiple amino acid sequence alignment, refinement, and masking

We aligned all OGs separately at the amino acid level using MAFFT L-INS-i [44] v6.951. Then we checked for misaligned sequences (henceforth called "outliers") in multiple amino acid sequence alignments (MSAs) of all OGs. This check was done with Perl scripts (available upon request) applying the following procedure: first, the maximal alignment length of a given multiple amino acid sequence alignment was recorded. Then, mean, median, and quartiles of BLOSUM62 distances of the amino acid sequences of all reference species were calculated. After that, the BLOSUM62 distance of each transcript to the sequence of its closest reference taxon (*i.e.*, the reference taxon found as best reciprocal hit) was calculated. Subsequently, it was checked whether this distance was below or above a cut-off value of 2.25 times the distance of the upper quartile to the mean of the BLOSUM62 distances among the reference species. Transcripts with a minimal BLOSUM62 distance to a reference species above the cut-off were classified as outliers, and also sequences with less than 20 overlapping sites to the corresponding sequence of the reference species. All outliers were extracted from the respective MSAs. Each outlier amino acid sequence was separately aligned to only the aligned orthologous sequences of the reference species, using the "–add" option in MAFFT L-INS-i. The refined outlier amino acid sequences were reintegrated into the respective MSA using the alignment of the reference species as a backbone. The outlier check procedure as described above was repeated for each MSA. Sequences that were still classified as outliers were finally removed from the respective MSA (see Additional file 8: Table S3). Gap-only sites were also removed from the MSAs.

Ambiguously aligned sections were identified with a modified version of ALISCORE [45-47]; for modifications, see [47]). We applied the default sliding window size, the maximal number of pairwise comparisons (*–r* option) and a special EST data scoring (*–e* option). Identified ambiguously aligned sections were removed ("masked") from the MSAs with ALICUT v.2.0 ([48], http://www.museumkoenig.de/web/ZFMK_Mitarbeiter/KckPatrick/Software/AliCUT/Download/index.de.html) (see Additional file 11: Table S6).

Design of seven specific decisive datasets addressing particular phylogenetic relationships

We call a dataset phylogenetically decisive if all included OGs contain at least one sequence of a representative of each taxonomic group of interest. To compile decisive datasets, we selected four taxonomic groups of interest for each of our seven phylogenetic questions (Table 2 and Table 3). All species relevant for a specific question were assigned to one of the four groups (also called "clusters", see below; see also Additional file 12). The monophyly of each group of species is assumed. All OGs that contained at least one sequence of a representative of each group were extracted with Perl scripts (available upon request) and concatenated into seven supermatrices that constitute the seven decisive datasets. The taxa that are not relevant for answering the respective question were removed (see also Additional file 13: Table S7). The amount and distribution of missing data in each dataset was visualized with *mare* v. 0.1.2-rc ([49], http://mare.zfmk.de) (Additional file 1: Figures S1-S7).

Phylogenetic analyses

For each of the seven datasets, we performed phylogenetic tree reconstruction with the maximum likelihood (ML) optimality criterion and Four-cluster Likelihood Mapping (FcLM) at the amino acid level. We refrained from calculating the Relative Composition Variability (RCV, see [50]) among the sequences in a dataset to select an optimal data subset (*e.g.*, first, second, and third codon positions of nucleotide sequence dataset, and amino acid sequence dataset) because the statistics is not independent of sequence length, number of sequences, and frequency of symbols. This renders a comparison of RCV between datasets with a different number of symbols and different lengths inappropriate.

For maximum likelihood tree inference, the smaller and larger datasets were treated in slightly different ways because of RAM limitations. For analyzing our small datasets (datasets 3, 4, 5, 7), we conducted one tree-search per dataset to determine the best fitting model, using the *–AUTO* function implemented in RAxML-Light

[51] v. 1.0.9., under the GAMMA model of rate heterogeneity [52] using the median for the discrete GAMMA approximation. Then, ML trees for the small datasets were inferred applying the *–f a* command line option in RAxML [53], v.7.3.1, HYBRID [54,55] with the CAT model of rate heterogeneity [53], the best-scoring amino acid substitution matrix, and empirical amino acid frequencies (PROTCAT, bestMODEL, F option). The final tree-searches were conducted under the GAMMA model of rate heterogeneity, again using the median for the discrete approximation. For analyzing our larger datasets (1, 2, and 6), we used RAxML-Light v. 1.0.9 to determine the best-scoring protein substitution model and for subsequent tree inferences. Based on randomized topologies of starting trees, we conducted 50 tree-searches with the CAT model of rate heterogeneity (PROTCATAUTOF) and estimated the best-scoring model using empirical frequencies (+ F) for each tree-search. We subsequently estimated the best final GAMMA likelihood and additional parameters under the GAMMA model using the median for the discrete approximation. For all datasets, the best-scoring amino acid model was the LG model [56].

We assessed statistical support for each node from bootstrap replicates. Bootstrap analyses were performed with the rapid bootstrap algorithm [53], using bootstopping criteria ([57], command line option: –# autoMRE -B 0.01). For analyzing the small datasets, the search for the best tree and the bootstrap analyses were performed in one single step (*–f a* option). For analyzing the large datasets, bootstrap analyses were performed separately and the bootstrap support was plotted on the respective best tree.

All ML analyses were conducted on Linux clusters at the Cologne High Efficient Operating Platform for Science (CHEOPS), Regionales Rechenzentrum Köln (RRZK) (http://rrzk.uni-koeln.de/cheops.html).

After tree inference, we scrutinized our trees for rogue taxa ([36,58], see Additional file 4, Chapter 4).

Trees were edited with Treegraph 2.0 [59], and rooted with respective outgroups (see Additional file 2: Figures S8-S15). Supermatrices (*i.e.*, datasets) are deposited at labarchives repository, DOI10.6070/H4G73BMJ, https://mynotebook.labarchives.com/share/ubulin/MC4wfDIzNDAzLzAvVHJlZU5vZGUvMjA0NzAzNzkzMHwwLjA.

Four-cluster Likelihood Mapping (FcLM)

We used FcLM proposed by Strimmer and von Haeseler [23] as an alternative method for analyzing single phylogenetic splits. In each decisive dataset, all included species were binned into four clusters that correspond to the taxonomic groups that are relevant for the respective phylogenetic relationship (see above, Table 2, and Additional file 12). The phylogenetic relationships between these four clusters represent the phylogenetic question of interest. In one case (dataset 6), we defined two different sets of clusters because two phylogenetic hypotheses had to be tested. For each dataset, we calculated the log-likelihood values of all non-redundant quartets drawn from the predefined species groups ("clusters") (see Additional file 4, Chapter 3). We implemented this in RAxML (as of v. 7.3) to be able to handle large-scale datasets. Calculation of log-likelihood values was performed using the GAMMA model of rate heterogeneity and empirical base frequencies with RAxML 7.3.1 (PTHREADS) on the MESCA System of the HPC Linux Cluster CHEOPS, RRZK, University of Cologne. We developed an additional tool written in Perl to map the support values of the RAxML analyses for each quartet onto 2D simplex graphs (available upon request). Results from the analysis of all seven datasets were plotted on the main tree (Figure 1). For the final phylogenetic inference, we compared support inferred from FcLM with ML bootstrap support.

Additional partitioned ML tree and FcLM analyses of dataset 5

We repeated ML tree reconstruction and FcLM based on partitioned analyses for dataset 5 to identify possible sources for incongruence between results of tree reconstruction and FcLM in this specific case. For the partitioned ML tree reconstruction (with 972 partitions), we followed the procedure applied on the large datasets (see above), but using ExaML (version 4.1 [2013-06-19]) instead of RAxML-Light, with the PSR model of rate heterogeneity (equal to CAT in RAxML-Light). We subsequently estimated the optimal parameters and the log-likelihood using the GAMMA model of rate heterogeneity. We performed 50 tree searches and choose the one with the best log-likelihood as best tree (Additional file 2: Figure S15). For partitioned FcLM analysis, we used the respective best models for each partition, selected during the preceding ML tree search (*–AUTO* option in RAxML), as input (Additional file 14: Table S8). For calculating the log-likelihood support for each drawn quartet, we used again the GAMMA model of rate heterogeneity and empirical base frequencies in RAxML 7.7.2 (PTHREADS). Results were again mapped onto a 2D simplex graph (Additional file 3: Figure S25).

Reconstruction of character evolution and groundplans

Morphological characters of immatures and adults were mapped onto the reconstructed tree using Mesquite ([33], http://mesquiteproject.org). As input, we used the datamatrix of morphological characters published by Beutel et al. [15] and the interordinal topology of the transcriptome-based phylogeny inferred from dataset 1, which represents the complete molecular datamatrix (Figure 1). The taxon sampling at the species level is not

congruent between Beutel et al. [15] and the present study. However, all orders are covered in both studies, and only evolutionary transformations between orders or supraordinal taxa are considered here. To reconstruct the character evolution and groundplan features at each node, we used the "Trace Character History" option and performed maximum parsimony reconstructions of groundplans (select "Parsimony Ancestral States") for categorical characters under unordered states assumption.

Availability of supporting data

The datasets supporting the results of this article are available in the labarchives repository, DOI10.6070/H4G73BMJ, https://mynotebook.labarchives.com/share/ubulin/MC4wfDIzNDAzLzAvVHJlZU5vZGUvMjA0NzAzNzkzMHwwLjA.

Additional files

Additional file 1: Figures S1-S7. Presence and absence of genes in datasets 1 to 7. Files visualize the data matrices of datasets 1 to 7, in terms of gene coverage (Figure S1: dataset 1 to Figure S7: dataset 7). Grey dot: gene present. White dot: gene absent. The data matrices were visualized with *mare* [49].

Additional file 2: Figures S8-S15. Full phylogenetic trees, inferred from ML analyses of datasets 1 to 7. Files show full phylogenetic trees, inferred from maximum likelihood (ML) tree reconstructions of datasets 1 to 7 (Figure S8: dataset 1 to Figure S14: dataset 7; Figure S15: best tree of the additional partitioned analysis of dataset 5). Branches with <50% bootstrap support are shown as unresolved. Species for which new transcriptome data were generated in this study are in bold print. For details of phylogenetic tree reconstruction, see Methods section of main text.

Additional file 3: Figures S16-S25. Results of the Four-cluster Likelihood Mapping (FcLM) as 2D simplex graphs. Figure S16. Exemplary 2D simplex graph based on the Four-cluster Likelihood Mapping (FcLM). For explanations see Additional file 4, Chapter 3. Figures S17-S25. 2D simplex graphs showing results of the Four-cluster Likelihood Mapping (FcLM) of datasets 1 to 7 (Figure S17: dataset 1 to Figure S221: dataset 5; Figure S22 and S23: dataset 6a and 6b; Figure S24: dataset 7, Figure S25: additional partitioned analysis of dataset 5). Left: the support for each quartet is shown as a single dot mapped onto the 2D simplex graph. Right: proportion of quartets with predominant support for the respective topology is given. For details on methods, topologies T1, T2, and T3, and interpretation of results see Methods and Results section of the main text, Additional file 4, Chapter 3, and Figure S16.

Additional file 4: More details on methods and results. The text gives more detailed information on methods (generation of new transcriptome data and retrieval of published data, orthology assignment, and Four-cluster Likelihood Mapping), and provides additional results (rogue taxa, morphological analyses).

Additional file 5: Figure_3_3D. Figure 3 of main text as 3D pdf. Click on image to activate animation.

Additional file 6: Table S1. Species for which new transcriptome data were generated, with collecting and preservation information. This table gives all available metadata for the species for which new transcriptome data were generated in this study, including, for example, collecting information, species identifying person, sex and stage, preservation details.

Additional file 7: Table S2. Statistics of newly generated transcriptome data. This table gives statistics of the generated data, *e.g.,* number of raw reads, number of contigs after assembly, length of contigs, and accession numbers at NCBI GenBank. All data can be found at NCBI Umbrella BioProject ID: PRJNA176423 - Evolution of holometabolous insects; BioProject accession number: SRP015962. For details on linker clipping and quality trimming see Additional file 4, Chapter 1.

Additional file 8: Table S3. All species included in this study, including previously published data. Listed are sources for download of data, results of orthology assignment, and results of subsequent quality assessment steps (see Methods section of main text for details). Capitalized species: whole genome sequence and an official gene set are available. Species marked with an asterisk were used as reference species in the ortholog reference set, see Additional file 4, Chapter 2 for details.

Additional file 9: Table S4. Reference species used in the ortholog reference set. Table lists the species that were used during compilation of the ortholog reference set, see Additional file 4, Chapter 2 for details, and information on download source and date. *Daphnia pulex* was used as reference species but not included in the taxon sampling.

Additional file 10: Table S5. List of 1,343 ortholog groups (OGs) included in the ortholog reference set. Table lists all OGs analyzed in this study, with OG ID, Uniprot ID, and preliminary annotation. Annotation was retrieved from OrthoDB4, either using a consensus rule for OGs marked with an asterisk, or adopting the annotation of *Pediculus humanus;* 'x' indicates the complete removal of an annotation during the cleaning process (see Additional file 4, Chapter 2 for details).

Additional file 11: Table S6. Proportion of excluded ambiguously aligned sites (%) for each ortholog group. In each ortholog group, alignment sections which were evaluated as ambiguous with ALISCORE were excluded prior to compilation of datasets 1 to 7, subsequent ML tree reconstruction and FcLM (see Methods section of main text for details).

Additional file 12: Species groups selected for the design of decisive datasets. For design of our seven datasets, we selected four taxonomic groups each of which is relevant to address a phylogenetic relationship in question. Species were binned into these four groups. In this file, we list the species included in each group for each of our datasets.

Additional file 13: Table S7. Number of ortholog groups (OGs) per species and dataset. Table lists how many OGs are covered by each species in the seven datasets that were analyzed in this study.

Additional file 14: Table S8. Best scoring model of each partition in partitioned analyses of dataset 5. The table lists the selected model for each partition of dataset 5, using the AUTO option implemented in ExaML, applied in the additional partitioned analyses (ML tree reconstruction and FcLM).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The study was conceived by BM, KM, KMK, UA, HA, RSP and RGB. MP and KM compiled the ortholog reference set. MP wrote HaMStRad. JW, MP, TZ, ON, CM, and BM wrote all necessary Perl scripts. AD conducted VecScreen analyses and did the submission of data. KM conducted all molecular data analyses except alignment refinement (done by CM) and rogue taxon analyses (done by AJA). AS re-implemented and parallelized the likelihood calculations on quartets in RAxML. Analyses of morphological data were done by RGB and FF. The manuscript was written by RSP, KM, RGB and BM with useful comments and revisions from all other authors. All authors read and approved the final manuscript.

Acknowledgements

We thank Dominique Zimmermann, Michael A. Ivie, Kai Schütte, Ewald Altenhofer, Dieter Stüning, Douglas Craig, Alexander Blanke, Martin Kubiak, Hans Malicky, Hans Pohl, and Eva Buscher and her husband for collecting or help in collecting, and identifying species. We acknowledge Torsten Struck, ZFMK, for kindly providing a Perl script for checking multiple assignments of transcripts to different orthologs. We want to thank LGC Genomics Berlin, especially Berthold Fartmann, Victor Achter at Cologne High Efficient Operating Platform for Science (CHEOPS), Regionales Rechenzentrum Köln

(RRZK), Julia Schwarzer (EAWAG, Kastanienbaum), and Claudia Etzbauer (ZFMK), as well as two anonymous reviewers for their helpful comments. RGB was funded by the DFG grant BE 1789/8-1. RSP, KM and BM were funded by the DFG grant MI 649/10.

Author details
[1]Zoologisches Forschungsmuseum Alexander Koenig, Abteilung Arthropoda, Adenauerallee 160, 53113 Bonn, Germany. [2]Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Molekulare Biodiversitätsforschung (zmb), Adenauerallee 160, 53113 Bonn, Germany. [3]CSIRO Ecosystem Sciences, Australian National Insect Collection, Clunies Ross Street, Acton ACT 2601, Australia. [4]Rutgers University, Department of Ecology, Evolution and Natural Resources, New Brunswick, NJ 08901, USA. [5]Naturhistorisches Museum Wien, 2. Zool. Abteilung, Burgring 7, 1010 Vienna, Austria. [6]Department of Evolutionary Biology, University of Vienna, Althanstraße 14, 1090 Vienna, Austria. [7]Institut für Spezifische Prophylaxe und Tropenmedizin, Medizinische Parasitologie, Medizinische Universität Wien (MUW), Kinderspitalgasse 15, 1090 Vienna, Austria. [8]Heidelberg Institute for Theoretical Studies (HITS), Scientific Computing Group, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany. [9]Karlsruher Institut für Technologie, Fakultät für Informatik, Postfach 698076128 Karlsruhe, Germany. [10]Biozentrum Grindel und Zoologisches Museum Hamburg, Universität Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany. [11]Entomology Group, Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum, Friedrich-Schiller-Universität Jena, Erbertstraße. 1, 07743 Jena, Germany.

Received: 8 October 2013 Accepted: 4 March 2014
Published: 20 March 2014

References

1. Kristensen NP: **Phylogeny of endopterygote insects, the most successful lineage of living organisms.** *Eur J Entomol* 1999, **96:**237–254.
2. Grimaldi D, Engel MS: *Evolution of the Insects.* Cambridge: Cambridge University Press; 2005.
3. Beutel RG, Pohl H: **Endopterygote systematics – where do we stand and what is the goal (Hexapoda, Arthropoda)?** *Syst Entomol* 2006, **31:**202–219.
4. Hennig W: *Die Stammesgeschichte der Insekten.* Frankfurt a. M: Waldemar Kramer; 1969.
5. Hinton HE: **The phylogeny of the panorpoid orders.** *Ann Rev Entomol* 1958, **3:**181–206.
6. Rasnitsyn AP, Quicke DLJ: *The history of insects.* Dordrecht: Kluwer Publications; 2002.
7. Kukalová-Peck J, Lawrence JF: **Relationships among coleopteran suborders and major endoneopteran lineages: evidence from hind wing characters.** *Eur J Entomol* 2004, **101:**95–144.
8. Beutel RG, Gorb S: **Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny.** *J Zool Syst Evol Res* 2001, **39:**77–207.
9. Beutel RG, Gorb S: **A revised interpretation of the evolution of attachment structures in Hexapoda (Arthropoda), with special emphasis on Mantophasmatodea.** *Arthropod Syst Phyl* 2006, **64:**3–25.
10. Wheeler WC, Whiting M, Wheeler QD, Carpenter JM: **The phylogeny of extant hexapod orders.** *Cladistics* 2001, **17:**113–169.
11. Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC: **The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology.** *Syst Biol* 1997, **46:**1–68.
12. Whiting MF: **Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera.** *Zool Scripta* 2002, **31:**93–104.
13. Wiegmann BM, Trautwein MD, Kim JW, Cassel BK, Bertone MA, Winterton SL, Yeates DK: **Single-copy nuclear genes resolve the phylogeny of the holometabolous insects.** *BMC Biol* 2009, **7:**34.
14. McKenna DD, Farrell BD: **9-genes reinforce the phylogeny of Holometabola and yield alternate views on the phylogenetic placement of Strepsiptera.** *PLoS ONE* 2010, **5:**e11887.
15. Beutel RG, Friedrich F, Hörnschemeyer T, Pohl H, Hünefeld F, Beckmann F, Meier R, Misof B, Whiting MF, Vilhemsen L: **Morphological and molecular evidence converging upon a robust phylogeny of the megadiverse Holometabola.** *Cladistics* 2011, **26:**1–15.
16. Ishiwata K, Sasaki G, Ogawa J, Miyata T, Su ZH: **Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences.** *Mol Phylogenet Evol* 2011, **58:**169–180.
17. Niehuis O, Hartig G, Grath S, Pohl H, Lehmann J, Tafer H, Donath A, Krauss V, Eisenhardt C, Hertel J, Petersen M, Mayer C, Meusemann K, Peters RS, Stadler PF, Beutel RG, Bornberg-Bauer E, McKenna DD, Misof B: **Genomic and morphological evidence converge to resolve the enigma of Strepsiptera.** *Curr Biol* 2012, **22:**1309–1313.
18. Aspöck U: **Phylogeny of the Neuropterida (Insecta: Holometabola).** *Zool Scripta* 2002, **31:**51–55.
19. Aspöck U, Haring E, Aspöck H: **The phylogeny of the Neuropterida: long lasting and current controversies and challenges (Insecta: Endopterygota).** *Arthropod Syst Phyl* 2012, **70:**119–129.
20. Beutel RG, Zimmermann D, Krauß M, Randolf S, Wipfler B: **Head morphology of *Osmylus fulvicephalus* (Osmylidae, Neuroptera) and its phylogenetic implications.** *Org Divers Evol* 2012, **10:**311–329.
21. Trautwein MD, Wiegmann BM, Beutel R, Kjer K, Yeates DK: **Advances in insect phylogeny at the dawn of the postgenomic era.** *Ann Rev Entomol* 2012, **57:**449–468.
22. Dell'Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walzl MG, Minh BQ, von Haeseler A, Ebersberger I, Pass G, Misof B: **Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects.** *Mol Biol Evol* 2014, **31:**239–249.
23. Strimmer K, von Haeseler A: **Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment.** *Proc Natl Acad Sci USA* 1997, **94:**6815–6819.
24. Friedrich F, Beutel RG: **Good bye Halteria? The thoracic morphology of Endopterygota (Insecta) and its phylogenetic implications.** *Cladistics* 2010, **26:**579–612.
25. Kjer KM, Carle FL, Litman J, Ware J: **A molecular phylogeny of Hexapoda.** *Arthropod Syst Phyl* 2006, **64:**35–44.
26. Winterton SL, Hardy NB, Wiegmann BM: **On wings of lace: phylogeny and Bayesian divergence time estimates of Neuropterida (Insecta) based on morphological and molecular data.** *Syst Entomol* 2010, **35:**349–378.
27. Achtelig M: **Neuropteroidea. Revisionary notes.** In *Insect Phylogeny.* Edited by Hennig W. Chichester, New York, Brisbane, Toronto: John Wiley & Sons; 1981:286–300.
28. Aspöck U, Plant JD, Nemeschkal HL: **Cladistic analysis of Neuropteran and their systematic position within Neuropterida (Insecta: Holometabola: Neuropterida: Neuroptera).** *Syst Entomol* 2001, **26:**73–86.
29. Aspöck U, Aspöck H: **Phylogenetic relevance of the genital sclerites of Neuropterida (Insecta: Holometabola).** *Syst Entomol* 2008, **33:**97–127.
30. Pohl H, Beutel RG: **The phylogeny of Strepsiptera (Hexapoda).** *Cladistics* 2005, **21:**328–374.
31. Johnston JS, Ross LD, Beani L, Hughes DP, Kathirithamby J: **Tiny genomes and endoreduplication in Strepsiptera.** *Insect Mol Biol* 2004, **13:**581–585.
32. Proffitt F: **Twisted parasites from "outer space" perplex biologists.** *Science* 2005, **307:**343.
33. Maddison WP, Maddison DR: *Mesquite: a modular system for evolutionary analysis*; 2007 [http://mesquiteproject.org]
34. Salichos L, Rokas A: **Inferring ancient divergences requires genes with strong phylogenetic signals.** *Nature* 2013, **497:**327–331.
35. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53:**506–514.
36. Aberer AJ, Stamatakis A: **A simple and accurate method for rogue taxon identification.** In *Bioinformatics and Biomedicine (BIBM).* Atlanta: IEEE International Conference on Bioinformatics and Biomedicine; 2011:118–122.
37. Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence.** *Mol Biol Evol* 2009, **26:**2719–2730.
38. Simon S, Narechania A, DeSalle R, Hadrys H: **Insect phylogenomics: exploring the source of incongruence using new transcriptomic data.** *Genome Biol Evol* 2012, **4:**1295–1309.
39. Beutel RG, Kristensen N-P, Pohl H: **Resolving insect phylogeny: the significance of cephalic structures of the Nannomecoptera in understanding endopterygote relationships.** *Arthropod Struct Dev* 2009, **38:**427–460.
40. Beutel RG, Vilhelmsen LB: **Head anatomy of Xyelidae (Hexapoda: Hymenoptera) and phylogenetic implications.** *Org Div Evol* 2007, **7:**207–230.
41. Hünefeld F, Mißbach C, Beutel RG: **The morphology and evolution of the female postabdomen of Holometabola (Insecta).** *Arthropod Struct Dev* 2012, **41:**361–371.

42. Beutel RG, Haas F: **Phylogenetic relationships of the suborders of Coleoptera (Insecta).** *Cladistics* 2000, **16:**103–141.
43. Ebersberger I, Strauss S, von Haeseler A: **HaMStR: profile hidden Markov model based search for orthologs in ESTs.** *BMC Evol Biol* 2009, **9:**157.
44. Katoh K, Toh H: **Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework.** *BMC Bioinforma* 2008, **9:**212.
45. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion.** *Syst Biol* 2009, **58:**21–34.
46. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont B, Wägele JW, Misof B: **Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees.** *Front Zool* 2010, **7:**10.
47. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walzl M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW, Misof B: **A phylogenomic approach to resolve the arthropod tree of life.** *Mol Biol Evol* 2010, **27:**2451–2464.
48. Kück P: *ALICUT: a Perlscript which cuts ALISCORE identified RSS.* 20th edition. Bonn, Germany: Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK); 2009 [http://www.museumkoenig.de/web/ZFMK_Mitarbeiter/KckPatrick/Software/AliCUT/Download/index.de.html]
49. Meyer B, Misof B: *MARE: Matrix Reduction – A tool to select optimized data subsets from supermatrices for phylogenetic inference.* Adenauerallee 160, 53113 Bonn, Germany: Zentrum für molekulare Biodiversitätsforschung (zmb) am ZFMK; 2010 [http://mare.zfmk.de]
50. Phillips MJ, Penny D: **The root of the mammalian tree inferred from whole mitochondrial genomes.** *Mol Phylogenet Evol* 2003, **28:**171–185.
51. Stamatakis A, Aberer AJ, Goll C, Smith SA, Berger SA, Izquierdo-Carrasco F: **RAxML-Light: a tool for computing TeraByte Phylogenies.** *Bioinformatics* 2012, **28:**2064–2066.
52. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends Ecol Evol* 1996, **11:**367–372.
53. Stamatakis A: **RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688–2690.
54. Pfeiffer W, Stamatakis A: **Hybrid MPI/Pthreads Parallelization of the RAxML Phylogenetics Code.** In *IPDPS Workshops, IEEE.* Atlanta; 2010:1–8.
55. Ott M, Zola J, Stamatakis A, Aluru S: **Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L.** In *Proceedings of the 2007 ACM/IEEE conference on Supercomputing.* Reno; 2007:4–11.
56. Le SQ, Gascuel O: **An improved general amino acid replacement matrix.** *Mol Biol Evol* 2008, **25:**1307–1320.
57. Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A: **How many bootstrap replicates are necessary?** *J Comput Biol* 2010, **17:**337–354.
58. Aberer AJA, Krompass D, Stamatakis A: **Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice.** *Syst Biol* 2012, **62:**162–166.
59. Stöver BC, Müller KF: **TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses.** *BMC Bioinforma* 2010, **11:**7.

doi:10.1186/1471-2148-14-52
Cite this article as: Peters *et al.*: The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. *BMC Evolutionary Biology* 2014 **14:**52.

9. T. A. White *et al.*, *PLOS Pathog.* **6**, e1001249 (2010).
10. G. Hu, J. Liu, K. A. Taylor, K. H. Roux, *J. Virol.* **85**, 2741–2750 (2011).
11. P. D. Kwong *et al.*, *Nature* **393**, 648–659 (1998).
12. M. Pancera *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1166–1171 (2010).
13. E. E. Tran *et al.*, *PLOS Pathog.* **8**, e1002797 (2012).
14. Y. Mao *et al.*, *Nat. Struct. Mol. Biol.* **19**, 893–899 (2012).
15. A. Harris *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11440–11445 (2011).
16. C. G. Moscoso *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6091–6096 (2011).
17. S. R. Wu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18844–18849 (2010).
18. R. Roy, S. Hohng, T. Ha, *Nat. Methods* **5**, 507–516 (2008).
19. Z. Zhou *et al.*, *ACS Chem. Biol.* **2**, 337–346 (2007).
20. C. W. Lin, A. Y. Ting, *J. Am. Chem. Soc.* **128**, 4542–4543 (2006).
21. Materials and Methods are available as supplementary materials on *Science* Online.
22. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
23. Q. Zheng *et al.*, *Chem. Soc. Rev.* **43**, 1044–1056 (2014).
24. N. G. Mukherjee, L. A. Lyon, J. M. Le Doux, *Nanotechnology* **20**, 065103 (2009).
25. K. Henzler-Wildman, D. Kern, *Nature* **450**, 964–972 (2007).
26. P. D. Kwong *et al.*, *Nature* **420**, 678–682 (2002).
27. U. Olshevsky *et al.*, *J. Virol.* **64**, 5701–5707 (1990).
28. F. Qin, *Biophys. J.* **86**, 1488–1501 (2004).
29. H. Haim *et al.*, *PLOS Pathog.* **5**, e1000360 (2009).
30. N. Sullivan *et al.*, *J. Virol.* **72**, 4694–4703 (1998).
31. J. Arthos *et al.*, *J. Biol. Chem.* **277**, 11456–11464 (2002).
32. A. L. DeVico, *Curr. HIV Res.* **5**, 561–571 (2007).
33. T. Zhou *et al.*, *Science* **329**, 811–817 (2010).
34. L. M. Walker *et al.*, *Science* **326**, 285–289 (2009).
35. L. M. Walker *et al.*, *Nature* **477**, 466–470 (2011).
36. M. Pancera *et al.*, *Nat. Struct. Mol. Biol.* **20**, 804–813 (2013).
37. J.-P. Julien *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 4351–4356 (2013).
38. J.-P. Julien *et al.*, *PLOS Pathog.* **9**, e1003342 (2013).
39. P. D. Kwong, J. R. Mascola, G. J. Nabel, *Nat. Rev. Immunol.* **13**, 693–701 (2013).
40. Z. Li *et al.*, *Antimicrob. Agents Chemother.* **57**, 4172–4180 (2013).
41. J. B. Munro, K. Y. Sanbonmatsu, C. M. Spahn, S. C. Blanchard, *Trends Biochem. Sci.* **34**, 390–400 (2009).
42. Subsequent to the submission of this manuscript, the structure of a trimeric prefusion HIV-1 Env (*43*) was determined in complex with broadly neutralizing antibodies PGT122 (*35*) and 35O22 (*44*). To determine the conformational state these antibodies captured in the crystal lattice, we measured smFRET on labeled JR-FL virions in the presence of either PGT122 or 35O22 or both. These data indicated that PGT122 strongly stabilized the ground state. In contrast, 35O22 had little effect on Env conformation. HIV-1 Env complexed with both antibodies exhibited slight ground state stabilization.
43. M. Pancera *et al.*, *Nature* **514**, 455–461 (2014).
44. J. Huang *et al.*, *Nature* 10.1038/nature13601 (2014).
45. J. S. McLellan *et al.*, *Nature* **480**, 336–343 (2011).

ACKNOWLEDGMENTS

We thank J. Jin, L. Agosto, T. Wang, R. B. Altman, and M. R. Wasserman for assistance; C. Walsh, J. Binely, A. Trkola, and M. Krystal for reagents; and members of the Structural Biology Section, Vaccine Research Center, for critically reading the manuscript. We thank I. Wilson and J. Sodroski for encouraging us to extend our approach to a R5-tropic Env. The data presented in this paper are tabulated in the main paper and in the supplementary materials. This work was supported by NIH grants R21 AI100696 to W.M. and S.C.B; P01 56550 to W.M., S.C.B., and A.B.S.; and R01 GM098859 to S.C.B.; by the Irvington Fellows Program of the Cancer Research Institute to J.B.M; by a fellowship from the China Scholarship Council-Yale World Scholars to X.M.; by grants from the International AIDS Vaccine Initiative's (IAVI) Neutralizing Antibody Consortium to D.R.B., W.C.K., and P.D.K.; and by funding from the NIH Intramural Research Program (Vaccine Research Center) to P.D.K. IAVI's work is made possible by generous support from many donors including: the Bill & Melinda Gates Foundation and the U.S. Agency for International Development (USAID). This study is made possible by the generous support of the American people through USAID. The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the U.S. government. Reagents from the NIH are subject to nonrestrictive material transfer agreements. Patent applications pertaining to this work are the U.S. and World Application US2009/006049 and WO/2010/053583, Synthesis of JRC-II-191 (A.B.S., J.R.C.), U.S. Patent Application 13/202,351, Methods and Compositions for Altering Photopysical Properties of Fluorophores via Proximal Quenching (S.C.B., Z.Z.); U.S. Patent Application 14/373,402 Dye Compositions, Methods of Preparation, Conjugates Thereof, and Methods of Use (S.C.B., Z.Z.); and International and US Patent Application PCT/US13/42249 Reagents and Methods for Identifying Anti-HIV Compounds (S.C.B., J.B.M., W.M.). S.C.B. is a co-founder of Lumidyne Corporation.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/346/6210/759/suppl/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S3
References (*46*–*56*)

4 April 2014; accepted 15 September 2014
Published online 8 October 2014;
10.1126/science.1254426

INSECT PHYLOGENOMICS

Phylogenomics resolves the timing and pattern of insect evolution

Bernhard Misof,[1]*† Shanlin Liu,[2,3]* Karen Meusemann,[1,4]* Ralph S. Peters,[5]* Alexander Donath,[1]* Christoph Mayer,[1]* Paul B. Frandsen,[6]* Jessica Ware,[7]* Tomáš Flouri,[8]* Rolf G. Beutel,[9]* Oliver Niehuis,[1]* Malte Petersen,[1]* Fernando Izquierdo-Carrasco,[8]* Torsten Wappler,[10]* Jes Rust,[10]* Andre J. Aberer,[8] Ulrike Aspöck,[11,12] Horst Aspöck,[13] Daniela Bartel,[12] Alexander Blanke,[1,18] Simon Berger,[8] Alexander Böhm,[12] Thomas R. Buckley,[14] Brett Calcott,[15] Junqing Chen,[3] Frank Friedrich,[16] Makiko Fukui,[17] Mari Fujita,[18] Carola Greve,[1] Peter Grobe,[1] Shengchang Gu,[3] Ying Huang,[2,3] Lars S. Jermiin,[19] Akito Y. Kawahara,[20] Lars Krogmann,[21] Martin Kubiak,[16] Robert Lanfear,[22,23,24] Harald Letsch,[25] Yiyuan Li,[2,3] Zhenyu Li,[3] Jiguang Li,[3] Haorong Lu,[3] Ryuichiro Machida,[18] Yuta Mashimo,[18] Pashalia Kapli,[8,26] Duane D. McKenna,[27] Guanliang Meng,[2,3] Yasutaka Nakagaki,[18] José Luis Navarrete-Heredia,[28] Michael Ott,[29] Yanxiang Ou,[3] Günther Pass,[12] Lars Podsiadlowski,[30] Hans Pohl,[9] Björn M. von Reumont,[31] Kai Schütte,[32] Kaoru Sekiya,[18] Shota Shimizu,[18] Adam Slipinski,[4] Alexandros Stamatakis,[8,33] Wenhui Song,[2,3] Xu Su,[2,3] Nikolaus U. Szucsich,[12] Meihua Tan,[2,3] Xuemei Tan,[3] Min Tang,[2,3] Jingbo Tang,[3] Gerald Timelthaler,[12] Shigekazu Tomizuka,[18] Michelle Trautwein,[34] Xiaoli Tong,[35] Toshiki Uchifune,[18,36] Manfred G. Walzl,[12] Brian M. Wiegmann,[37] Jeanne Wilbrandt,[1] Benjamin Wipfler,[9] Thomas K. F. Wong,[19] Qiong Wu,[2,3] Gengxiong Wu,[3] Yinlong Xie,[3] Shenzhou Yang,[2,3] Qing Yang,[2,3] David K. Yeates,[4] Kazunori Yoshizawa,[38] Qing Zhang,[2,3] Rui Zhang,[2,3] Wenwei Zhang,[3] Yunhui Zhang,[3] Jing Zhao,[2,3] Chengran Zhou,[2,3] Lili Zhou,[2,3] Tanja Ziesmann,[1] Shijie Zou,[3] Yingrui Li,[3] Xun Xu,[3] Yong Zhang,[2,3] Huanming Yang,[3] Jian Wang,[3] Jun Wang,[3,39,40,41,42]*† Karl M. Kjer,[43]*† Xin Zhou[2,3]*†

Insects are the most speciose group of animals, but the phylogenetic relationships of many major lineages remain unresolved. We inferred the phylogeny of insects from 1478 protein-coding genes. Phylogenomic analyses of nucleotide and amino acid sequences, with site-specific nucleotide or domain-specific amino acid substitution models, produced statistically robust and congruent results resolving previously controversial phylogenetic relationships. We dated the origin of insects to the Early Ordovician [~479 million years ago (Ma)], of insect flight to the Early Devonian (~406 Ma), of major extant lineages to the Mississippian (~345 Ma), and the major diversification of holometabolous insects to the Early Cretaceous. Our phylogenomic study provides a comprehensive reliable scaffold for future comparative analyses of evolutionary innovations among insects.

Insects (*1*) were among the first animals to colonize and exploit terrestrial and freshwater ecosystems. They have shaped Earth's biota, exhibiting coevolved relationships with many groups, from flowering plants to humans. They were the first to master flight and establish social societies. However, many aspects of insect evolution are still poorly understood (*2*). The oldest known fossil insects are from the Early Devonian [~412 million years ago (Ma)], which has led to the hypothesis that insects originated in the Late Silurian with the earliest terrestrial ecosystems (*3*). Molecular data, however, point to a Cambrian or at least Early Ordovician origin (*4*), which implies that early diversification of insects occurred in marine or coastal environments. Because of the absence of insect fossils from the Cambrian to the Silurian, these conclusions remain highly controversial. Furthermore, the phylogenetic relationships among major clades of polyneopteran insect orders—including grasshoppers and crickets (Orthoptera), cockroaches (Blattodea), and termites (Isoptera)—have remained elusive, as has the phylogenetic position of the enigmatic Zoraptera. Even the closest extant relatives of Holometabola (e.g.,

beetles, moths and butterflies, flies, sawflies, wasps, ants, and bees) are unknown. Thus, in order to understand the origins of physiological and morphological innovations in insects (e.g., wings and metamorphosis), it is important to reliably reconstruct the tempo and mode of insect diversification. We therefore conducted a phylogenomic study on 1478 single-copy nuclear genes obtained from genomes and transcriptomes representing key taxa from all extant insect orders and other arthropods (144 taxa) and estimated divergence dates with a validated set of 37 fossils (*5*).

Phylogenomic analyses of transcriptome and genome sequence data (*6*) can be compromised by sparsely populated data matrices, gene paralogy, sequence misalignment, and deviations from the underlying assumptions of applied evolutionary models, which may result in biased statistical confidence in phylogenetic relationships and temporal inferences. We addressed these obstacles by removing confounding factors in our analysis (*5*) (fig. S2).

We sequenced more than 2.5 gigabases (Gb) of cDNA from each of 103 insect species, which represented all extant insect orders (*5*). Additionally, we included published transcript sequence data that met our standards (table S2) and official gene sets of 14 arthropods with sequenced draft genomes (*5*), of which 12 served as references during orthology prediction of transcripts (tables S2 and S4). Comparative analysis of the reference species' official gene sets identified 1478 single-copy nuclear genes present in all these species (tables S3 and S4). Functional annotation of these genes revealed that many serve basic cellular functions (tables S14 and S15 and figs. S4 to S6). A graph-based approach using the best reciprocal genome- and transcriptome-wide hit criterion identified, on average, 98% of these genes in the 103 de novo sequenced transcriptomes, but only 79% and 62% in the previously published transcriptomes of in- and out-group taxa, respectively (tables S12 and S13).

After transcripts had been assigned and aligned to the 1478 single-copy nuclear-encoded genes, we checked for highly divergent, putatively misaligned transcripts. Of the 196,027 aligned transcripts, 2033 (1%) were classified as highly divergent. Of these, 716 were satisfactorily realigned with an automated refinement. However, alignments of 1317 transcripts could not be improved, and these transcripts were excluded from our analyses (supplementary data file S5, http://dx.doi.org/10.5061/dryad.3c0f1).

Nonrandom distribution of missing data among taxa can inflate statistical support for incorrect tree topologies (*7*). Because we detected a nonrandom distribution of missing data, we only considered data blocks if they contained information from at least one representative of each of the 39 predefined taxonomic groups of undisputed monophyly (table S6). In this representative data set, the extent of missing data was still between 5 and 97.7% in pairwise sequence comparisons, with high percentages primarily because of the data scarcity in some previously published out-group taxa (table S19 and figs. S7 to S10).

We inferred maximum-likelihood phylogenetic trees (Fig. 1) with both nucleotide (second-codon positions only and applying a site-specific rate model) and amino acid–sequence data (applying a protein domain–based partitioning scheme to improve the biological realism of the applied evolutionary models) from the representative data set (*5*) (figs. S21, S22, and S23, A and D). Trees from both data sets were fully congruent. The absence of taxa that cannot be robustly placed on the tree (rogue taxa) in the amino acid–sequence data set and the presence of a few rogue taxa that did not bias tree inference in the nucleotide sequence data set (*5*) indicated a sufficiently representative taxonomic sampling.

To detect confounding signal derived from nonrandom data coverage, we randomized amino acids within taxa, while preserving the distribution of data coverage in the representative data set (*5*). This approach revealed no evidence of biased node support that could be attributed to nonrandom data coverage (*5*) (figs. S11 and S12 and table S20). Phylogenomic data may violate the assumption of time-reversible evolutionary processes, irrespective of what partition scheme one applies, which could lead to incorrect tree estimates and biased node support. Because sections in the amino acid–sequence alignments of the representative data set violating these assumptions were present, we tested whether the observed compositional heterogeneity across taxa biased node support but found no evidence for this (*5*) (fig. S20). We next discarded data strongly violating the assumption of time-reversible evolutionary processes (tables S21 and S22, data files S6 to S8, and figs. S13 to S19). Results from phylogenetic analysis of this filtered data set (*5*) were fully congruent with those obtained from analyzing the unfiltered representative data set. The nucleotide sequence data of the representative data set containing also first and third codon positions strongly violated the assumption of time-reversible evolutionary processes, but still supported largely congruent topologies (fig. S23, B to D). In summary, our phylogenetic inferences are unlikely to be biased by any of the above-mentioned confounding factors.

Our phylogenomic study suggests an Early Ordovician origin of insects (Hexapoda) at ~479 Ma [confidence interval (CI), 509 to 452 Ma] and a radiation of ectognathous insects in the Early Silurian ~441 Ma (CI 465 to 421 Ma) (Figs. 1 and 2). These estimates imply that insects colonized land at roughly the same time as plants (*8*), in agreement with divergence date estimates on the basis of other molecular data (*4*).

The early diversification pattern of insects has remained unclear (*2*, *7*, *9*). We received support for a monophyly of insects, including Collembola and Protura as closest relatives (*10*), and Diplura as closest extant relatives of bristletails (Archaeognatha), silverfish (Zygentoma), and winged insects (Pterygota) (Fig. 1). Furthermore, our analyses corroborate Remipedia, cave-dwelling crustaceans, as the closest extant relatives of insects (*11*, *12*).

A close phylogenetic relationship of bristletails to a clade uniting silverfish and winged insects (Dicondylia) is generally accepted. However, the monophyly of silverfish has been questioned, with the relict *Tricholepidion gertschi* considered more distantly related to winged insects than

[1]Zoologisches Forschungsmuseum Alexander Koenig (ZFMK)/Zentrum für Molekulare Biodiversitätsforschung (ZMB), Bonn, Germany. [2]China National GeneBank, BGI-Shenzhen, China. [3]BGI-Shenzhen, China. [4]Australian National Insect Collection, Commonwealth Scientific and Industrial Research Organization (Australia) (CSIRO), National Research Collections Australia, Canberra, ACT, Australia. [5]Abteilung Arthropoda, Zoologisches Forschungsmuseum Alexander Koenig (ZFMK), Bonn, Germany. [6]Department of Entomology, Rutgers University, New Brunswick, NJ 08854, USA. [7]Department of Biological Sciences, Rutgers University, Newark, NJ 08854, USA. [8]Scientific Computing, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany. [9]Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum Jena, FSU Jena, Germany. [10]Steinmann-Institut, Bereich Paläontologie, Universität Bonn, Germany. [11]2. Zoologische Abteilung (Insekten), Naturhistorisches Museum Wien, Vienna, Austria. [12]Department of Integrative Zoology, Universität Wien, Vienna, Austria. [13]Institut für Spezifische Prophylaxe und Tropenmedizin, Medizinische Parasitologie, Medizinische Universität Wien (MUW), Vienna, Austria. [14]Manaaki Whenua Landcare Research, Auckland, New Zealand. [15]Center for Advanced Modeling, Emergency Medicine Department, Johns Hopkins University, Baltimore, MD 21209, USA. [16]Biozentrum Grindel und Zoologisches Museum, Universität Hamburg, Hamburg, Germany. [17]Evolutionary Morphology Laboratory, Graduate School of Science and Engineering, Ehime University, Japan. [18]Sugadaira Montane Research Center/ Hexapod Comparative Embryology Laboratory, University of Tsukuba, Japan. [19]Land and Water Flagship, CSIRO, Canberra, ACT, Australia. [20]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA. [21]Entomology, Staatliches Museum für Naturkunde Stuttgart (SMNS), Germany. [22]Ecology Evolution and Genetics, Research School of Biology, Australian National University, Canberra, ACT, Australia. [23]National Evolutionary Synthesis Center, Durham, NC 27705, USA. [24]Department of Biological Sciences, Macquarie University, Sydney, Australia. [25]Department für Botanik und Biodiversitätsforschung, Universität Wien, Vienna, Austria. [26]Natural History Museum of Crete, University of Crete, Post Office Box 2208, Gr-71409, Iraklio, and Biology Department, University of Crete, Iraklio, Crete, Greece. [27]Department of Biological Sciences and Feinstone Center for Genomic Research, University of Memphis, Memphis, TN 38152, USA. [28]Centro Universitario de Ciencias Biólogicas y Agropecuarias, Centro de Estudios en Zoología, Universidad de Guadalajara, Zapopan, Jalisco, México. [29]Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, Garching, Germany. [30]Institute of Evolutionary Biology and Ecology, Zoology and Evolutionary Biology, University of Bonn, Bonn, Germany. [31]Department of Life Sciences, The Natural History Museum London, London, UK. [32]Abteilung Entomologie, Biozentrum Grindel und Zoologisches Museum, Universität Hamburg, Hamburg, Germany. [33]Fakultät für Informatik, Karlsruher Institut für Technologie, Karlsruhe, Germany. [34]California Academy of Sciences, San Francisco, CA 94118, USA. [35]Department of Entomology, College of Natural Resources and Environment, South China Agricultural University, China. [36]Yokosuka City Museum, Yokosuka, Kanagawa, Japan. [37]Department of Entomology, North Carolina State University, Raleigh, NC 27695, USA. [38]Systematic Entomology, Hokkaido University, Sapporo, Japan. [39]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [40]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia. [41]Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, China. [42]Department of Medicine, University of Hong Kong, Hong Kong. [43]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08854, USA.

*Major contributors.

†Corresponding author. E-mail: xinzhou@genomics.cn (X.Z.), b.misof.zfmk@uni-bonn.de (B.M.), kjer@aesop.rutgers.edu (K.M.K), wangj@genomics.cn (J.W.)

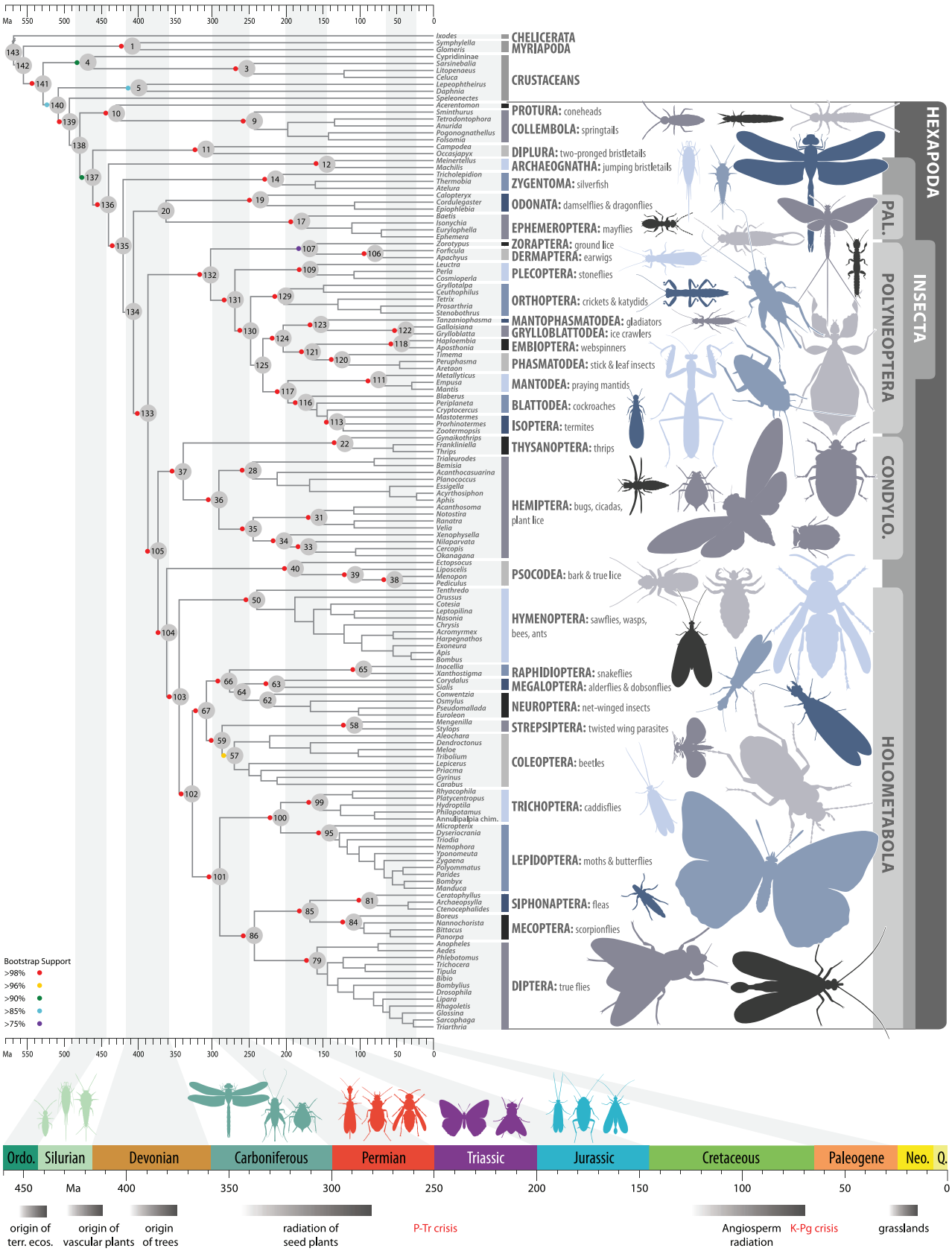

Fig. 1. Dated phylogenetic tree of insect relationships. The tree was inferred through a maximum-likelihood analysis of 413,459 amino acid sites divided into 479 metapartitions. Branch lengths were optimized and node ages estimated from 1,050,000 trees sampled from trees separately generated for 105 partitions that included all taxa (*5*). All nodes up to orders are labeled with numbers (gray circles). Colored circles indicate bootstrap support (*5*) (left key). The time line at the bottom of the tree relates the geological origin of insect clades to major geological and biological events. CONDYLO, Condylognatha; PAL, Palaeoptera.

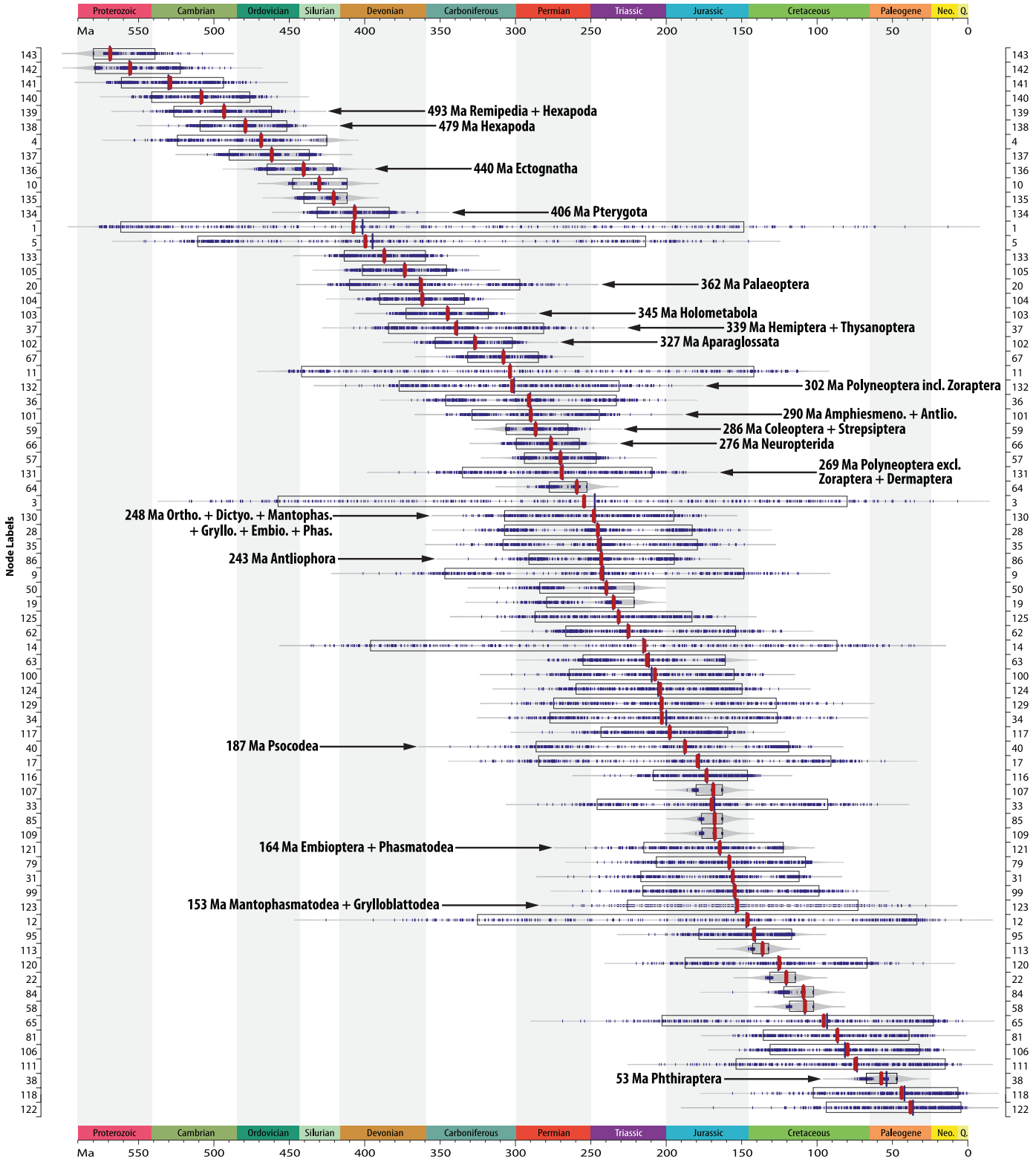

Fig. 2. Sorted ordinal and interordinal node age estimates. For each labeled node (numbers on the left and right of the figure correspond to the node labels in the tree of Fig. 1), the median (red bar), and the range of the upper and lower confidence interval (black rectangle) of age estimates are illustrated. These medians and upper and lower confidence intervals are derived from uniformly sampled trees over all 105 metapartitions (5). Additionally, we present medians of age estimates separately derived from each metapartition. Within the bean plot (gray scale), blue bars indicate the distribution of median age estimates, large blue bars indicate the inferred median of medians. All node age estimates refer to the estimated common origin of included species. Stem-lineage representatives can, of course, be older. The maximum root age of the tree was set to 580 Ma to coincide with the oldest Ediacaran fossils (5).

other silverfish (*13*). We find that silverfish are monophyletic, consistent with recently published morphological studies (*14*), and estimate that *Tricholepidion* diverged from other silverfish in the Late Triassic (~214 Ma) (Figs. 1 and 2). This result implies parallel and independent loss of the ligamentous head endoskeleton, abdominal styli, and coxal vesicles in winged insects and silverfish (*5*).

The diversification of insects is undoubtedly related to the evolution of flight. Fossil winged insects exist from the Late Mississippian (~324 Ma) (*15*), which implies a pre-Carboniferous origin of insect flight. The description of †*Rhyniognatha* (~412 Ma) from a mandible, potentially indicative of a winged insect, suggested an Early Devonian to Late Silurian origin of winged insects (*3*). Our results corroborate an origin of winged insect lineages during this time period (*16*) (Figs. 1 and 2), which implies that the ability to fly emerged after the establishment of complex terrestrial ecosystems.

Ephemeroptera and Odonata are, according to our analyses, derived from a common ancestor. However, node support is low for Palaeoptera (Ephemeroptera + Odonata) and for a sister group relationship of Palaeoptera to modern winged insects (Neoptera), which indicates that additional evidence, including extensive taxon sampling and the analysis of genomic meta-characters (*17*), will be necessary to corroborate these relationships.

We find strong support for the monophyly of Polyneoptera, a group that comprises earwigs, stoneflies, grasshoppers, crickets, katydids (Orthoptera), Embioptera, Phasmatodea, Mantophasmatodea, Grylloblattodea, cockroaches, mantids, termites, and Zoraptera (*18*–*20*). We estimated the origin of the polyneopteran lineages at ~302 Ma (CI 377 to 231 Ma) in the Pennsylvanian (Figs. 1 and 2), consistent with the idea that at least part of the rich Carboniferous neopteran insect fauna was of polyneopteran origin. Finally, our analyses suggest that the major diversity within living cockroaches, mantids, termites, and stick insects evolved after the Permian mass extinction.

Given that the oldest known fossil hemipterans date to the Middle Pennsylvanian (~310 Ma) (*21*), it had been thought that the stylet marks on liverworts from the Late Devonian (~380 Ma) (*22*) could not have been of hemipteran origin. Our study indicates that true bugs (Hemiptera) and their sister lineage, thrips (Thysanoptera), all of which possess piercing-sucking mouthparts, originated ~373 Ma (CI 401 to 346 Ma), which gives support to the possibility of a hemipteroid origin of Early Paleozoic stylet marks.

True bugs, thrips, bark lice (Psocoptera), and true lice (Phthiraptera) (together called Acercaria) were thought to be the closest extant relatives of Holometabola (Acercaria + Holometabola = Eumetabola) (*10*). However, convincing morphological features and fossil intermediates supporting a monophyly of Acercaria are lacking (*13*). We recovered bark and true lice (Psocodea) as likely closest extant relatives of Holometabola (*5*), which suggests that both groups started to diverge in the Devonian-Mississippian ~362 Ma (CI 390 to 334 Ma) (Figs. 1 and 2). However, this result did not receive support in all statistical tests and, therefore, should be further investigated in future studies that embrace additional types of characters (*17*).

We estimated that the radiation of parasitic lice occurred ~53 Ma (CI 67 to 46 Ma), which implies that they diversified well after the emergence of their avian and mammalian hosts in the Late Cretaceous–Early Eocene and contradicts the hypothesis that parasitic lice originated on feathered theropod dinosaurs ~130 Ma (*23*).

Within Holometabola, our study recovered phylogenetic relationships fully congruent with those suggested in recent studies (*2*, *24*, *25*). Although we estimated the origin of stem lineages of many holometabolous insect orders in the Late Carboniferous, we dated the spectacular diversifications within Hymenoptera, Diptera, and Lepidoptera to the Early Cretaceous, contemporary with the radiation of flowering plants (*21*, *26*). The almost linear increase in interordinal insect diversity suggests that the process of diversification of extant insects may not have been severely affected by the Permian and Cretaceous biodiversity crises (Fig. 2).

With this study, we have provided a robust phylogenetic backbone tree and reliable time estimates of insect evolution. These data and analyses establish a framework for future comparative analyses on insects, their genomes, and their morphology.

REFERENCES AND NOTES

1. The term "insects" is used here in a broad sense and synonymous to Hexapoda (including the ancestrally wingless Protura, Collembola, and Diplura).
2. M. D. Trautwein, B. M. Wiegmann, R. Beutel, K. M. Kjer, D. K. Yeates, *Annu. Rev. Entomol.* **57**, 449–468 (2012).
3. M. S. Engel, D. A. Grimaldi, *Nature* **427**, 627–630 (2004).
4. O. Rota-Stabelli *et al.*, *Curr. Biol.* **23**, 392–398 (2013).
5. Materials and methods are available as supplementary material on *Science* Online.
6. Transcriptome refers to the sequencing of all of the mRNAs of an individual or many individuals present at the time of preservation.
7. E. Dell'Ampio *et al.*, *Mol. Biol. Evol.* **31**, 239–249 (2014).
8. C. C. Labandeira, *Arthro Syst. Phylo.* **64**, 53–94 (2006).
9. K. Meusemann *et al.*, *Mol. Biol. Evol.* **27**, 2451–2464 (2010).
10. W. Hennig, *Die Stammesgeschichte der Insekten* (Kramer, Frankfurt am Main, 1969).
11. With the notable exception of Cephalocarida, for which RNA-sequencing data were unavailable to us.
12. B. M. von Reumont *et al.*, *Mol. Biol. Evol.* **29**, 1031–1045 (2012).
13. N. P. Kristensen, *Annu. Rev. Entomol.* **26**, 135–157 (1981).
14. A. Blanke, M. Koch, B. Wipfler, F. Wilde, B. Misof, *Front. Zool.* **11**, 16 (2014).
15. J. Prokop, A. Nel, I. Hoch, *Geobios* **38**, 383–387 (2005).
16. These estimates are robust whether or not †*Rhyniognatha* (Pragian stage, ~ 412 Ma) is used as a calibration point.
17. O. Niehuis *et al.*, *Curr. Biol.* **22**, 1309–1313 (2012).
18. H. Letsch, S. Simon, *Syst. Entomol.* **38**, 783–793 (2013).
19. K. Ishiwata, G. Sasaki, J. Ogawa, T. Miyata, Z.-H. Su, *Mol. Phylogenet. Evol.* **58**, 169–180 (2011).
20. K. Yoshizawa, *Syst. Entomol.* **36**, 377–394 (2011).
21. A. Nel *et al.*, *Nature* **503**, 257–261 (2013).
22. C. C. Labandeira, S. L. Tremblay, K. E. Bartowski, L. VanAller Hernick, *New Phytol.* **202**, 247–258 (2014).
23. V. S. Smith *et al.*, *Biol. Lett.* **7**, 782–785 (2011).
24. R. G. Beutel *et al.*, *Cladistics* **27**, 341–355 (2011).
25. B. M. Wiegmann *et al.*, *BMC Biol.* **7**, 34 (2009).
26. J. A. Doyle, *Annu. Rev. Earth Planet. Sci.* **40**, 301–326 (2012).

ACKNOWLEDGMENTS

The data reported in this paper are tabulated in the supplementary materials and archived at National Center for Biotechnology Information, NIH, under the Umbrella BioProject ID PRJNA183205 ("The 1KITE project: evolution of insects"). Supplementary files are archived at the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.3c0f1. Funding support: China National GeneBank and BGI-Shenzhen, China; German Research Foundation (NI 1387/1-1; MI 649/6, MI 649/10, RE 345/1-2, BE1789/8-1, BE 1789/10-1, STA 860/4, Heisenberg grant WA 1496/8-1); Austria Science Fund FWF; NSF (DEB 0816865); Ministry of Education, Culture, Sports, Science and Technology of Japan Grant-in-Aid for Young Scientists (B 22770090); Japan Society for the Promotion of Science (P14071); Deutsches Elektronen-Synchrotron (I-20120065); Paul Scherrer Institute (20110069); Schlinger Endowment to CSIRO Ecosystem Sciences; Heidelberg Institute for Theoretical Studies; University of Memphis-FedEx Institute of Technology; and Rutgers University. The authors declare no conflicts of interest.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/346/6210/763/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S27
Tables S1 to S26
Data File Captions S1 to S14
References (*27*–*187*)

17 June 2014; accepted 23 September 2014
10.1126/science.1257570

Decisive Data Sets in Phylogenomics: Lessons from Studies on the Phylogenetic Relationships of Primarily Wingless Insects

Emiliano Dell'Ampio,[†,1] Karen Meusemann,[*,†,2,3] Nikolaus U. Szucsich,[†,1] Ralph S. Peters,[†,4] Benjamin Meyer,[5] Janus Borner,[6] Malte Petersen,[2] Andre J. Aberer,[7] Alexandros Stamatakis,[7,8] Manfred G. Walzl,[1] Bui Quang Minh,[9] Arndt von Haeseler,[10] Ingo Ebersberger,[11] Günther Pass,[1] and Bernhard Misof[*,2]

[1]Department of Integrative Zoology, University of Vienna, Vienna, Austria

[2]Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Molekulare Biodiversitätsforschung (zmb), Bonn, Germany

[3]CSIRO Ecosystem Sciences, Australian National Insect Collection, Acton, ACT, Australia

[4]Zoologisches Forschungsmuseum Alexander Koenig, Abteilung Arthropoda, Bonn, Germany

[5]Institut für Systemische Neurowissenschaften, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

[6]Biozentrum Grindel & Zoologisches Museum, Universität Hamburg, Hamburg, Germany

[7]Heidelberg Institute for Theoretical Studies (HITS), Scientific Computing Group, Heidelberg, Germany

[8]Karlsruher Institut für Technologie, Fakultät für Informatik, Karlsruhe, Germany

[9]Center for Integrative Bioinformatics Vienna (CIBIV), Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria

[10]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

[11]Institute for Cell Biology and Neuroscience, Goethe-Universität Frankfurt, Frankfurt am Main, Germany

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: mail@karen-meusemann.de; b.misof.zfmk@uni-bonn.de.

Associate editor: Nicolas Vidal

Abstract

Phylogenetic relationships of the primarily wingless insects are still considered unresolved. Even the most comprehensive phylogenomic studies that addressed this question did not yield congruent results. To get a grip on these problems, we here analyzed the sources of incongruence in these phylogenomic studies by using an extended transcriptome data set. Our analyses showed that unevenly distributed missing data can be severely misleading by inflating node support despite the absence of phylogenetic signal. In consequence, only decisive data sets should be used which exclusively comprise data blocks containing all taxa whose relationships are addressed. Additionally, we used Four-cluster Likelihood Mapping (FcLM) to measure the degree of congruence among genes of a data set, as a measure of support alternative to bootstrap. FcLM showed incongruent signal among genes, which in our case is correlated neither with functional class assignment of these genes nor with model misspecification due to unpartitioned analyses. The herein analyzed data set is the currently largest data set covering primarily wingless insects, but failed to elucidate their interordinal phylogenetic relationships. Although this is unsatisfying from a phylogenetic perspective, we try to show that the analyses of structure and signal within phylogenomic data can protect us from biased phylogenetic inferences due to analytical artifacts.

Key words: **phylogenomics, ESTs, likelihood quartet mapping, conflicting hypotheses, Entognatha, Nonoculata, Ellipura, Protura, Diplura, Collembola, missing data.**

Article

Introduction

Despite enormous efforts to resolve the tree of life, several deep nodes are still considered unresolved. A good example for such problems are the unresolved phylogenetic relationships of primarily wingless insects.

Most phylogenetic studies including multigene and phylogenomic analyses have recovered the monophyly of Hexapoda, the insect clade in a broad taxonomic sense (Regier et al. 2008, 2010; von Reumont et al. 2009, 2012; Meusemann et al. 2010; Trautwein et al. 2012). Furthermore, the monophyly of Ectognatha, which comprises insects in a strict taxonomic sense, namely jumping bristletails, silverfishes and firebrats, and winged insects, is well supported (reviewed in Grimaldi 2010; Trautwein et al. 2012). By contrast, phylogenetic relationships among the entognathous primarily wingless insects, the Protura (coneheads), Collembola (springtails), and Diplura (two-pronged bristletails), are unclear. Many authors consider these entognathous insects as being monophyletic, considering entognathy in which mouth parts are concealed in gnathal pouches (first discussed in detail by Hennig 1953) to have evolved in the last common ancestor of the three groups.

Within Entognatha, either a clade uniting Protura and Collembola, referred to as Ellipura (Börner 1910), or a clade uniting Protura and Diplura, referred to as Nonoculata (Luan et al. 2005), has been proposed (Ellipura [Hennig 1953; Kristensen 1981, 1997; Shao et al. 1999; Bitsch and Bitsch 2000, 2004; Carapelli et al. 2000; Zhang et al. 2001]; Nonoculata [Giribet and Wheeler 2001; Giribet et al. 2004; Luan et al. 2005; Kjer et al. 2006; Mallatt and Giribet 2006; Misof et al. 2007; Dell'Ampio et al. 2009; von Reumont et al. 2009; Mallatt et al. 2010]). Other authors consider a paraphyly of Entognatha to be more likely, with Diplura as closest relatives to Ectognatha. Possible arguments for this hypothesis include the evolutionary origin of paired pretarsal claws and paired cerci (Kukalová-Peck 1987; Koch 1997; Beutel and Gorb 2006), the ultrastructure of the sperm (Dallai et al. 2011), and the differentiation process of the embryonic amnion (Machida 2006) in the last common ancestor of Diplura and Ectognatha.

Meusemann et al. (2010) and von Reumont et al. (2012) published the most relevant data sets and analyses covering the phylogenetic relationships among primarily wingless insects by including expressed sequence tag (EST) data of representatives of Protura, Collembola, and Diplura. Although both studies recovered the monophyly of Entognatha, Meusemann et al. found strong evidence for Protura and Diplura as closest relatives (i.e., Nonoculata) and von Reumont et al. for Protura and Collembola as closest relatives (i.e., Ellipura). These incongruent results are puzzling because taxon sampling of the primarily wingless insects is comparable in both studies, as well as the strategies used for orthology assignment, alignment masking, matrix optimization, and tree inference.

These special circumstances put us into the exceptionally favorable position to analyze possible sources of incongruence among these two large phylogenomic data sets. Most phylogenomic studies are based on concatenated supermatrices with low gene data coverage. Focusing on relationships among specific groups, many data blocks within such supermatrices therefore may not contain data for all taxa under consideration. Consequently, our starting hypothesis was that extensive missing data may mislead proper tree reconstruction. To tackle this problem, we complement the publicly available EST data of primarily wingless insects with additional EST data from representatives of Japygidae (Diplura) and Zygentoma (silverfishes and firebrats). We took particular care to concatenate a data set that contains only gene data blocks for which entognathous hexapods and outgroups had gene data coverage. We call such a data set in the following a decisive data set. Note that the term decisiveness has been used before in the context of phylogenomic data sets (Steel and Sanderson 2010; Sanderson et al. 2010), albeit based on a distinct criterion. The concatenated data set is the largest known data set covering primarily wingless insects. It was this data set that allowed us to analyze the effect of the observed uneven distribution of missing data on the extent of bootstrap support (BS). Complementary to the application of BS measures, we applied a Four-cluster Likelihood Mapping (FcLM) approach (Strimmer and von Haeseler 1997), which has been shown to be effective in disentangling signal among four groups of species. The application of bootstrapping and FcLM helped to assess the effect of the uneven distribution of missing data in indecisive data sets. Complementary to the previously mentioned analyses, we addressed the problem of incongruent signal among genes in a multigene data set by comparing tree reconstructions based on the entire decisive data set with tree reconstructions based on subsets of genes that support incongruent hypotheses. Altogether, our approach provides potential explanations for contradictory results among phylogenomic studies by pointing out underestimated sources of error and incongruence.

Results

Orthology Assignment, Alignment, and Alignment Masking

Using the reference set of 1,886 1:1 orthologous genes (OGs), we identified between 52 and 682 putative 1:1 orthologous transcripts in the transcriptome assemblies of primarily wingless hexapods (table 1) and up to 1,886 for all taxa (supplementary table S1, Supplementary Material online). We excluded 20 OGs that were present in the five reference species but absent from all other species from subsequent analyses. After alignment masking (i.e., the exclusion of multiple sequence alignment sections in which sequence similarity cannot be distinguished from random similarity of sequences), the concatenated superalignment was composed of 73 taxa with a total alignment length of 881,235 amino acid sites, partitioned into 1,866 genes (supplementary fig. S1; for gene annotations, see supplementary table S2, Supplementary Material online).

Relationships among Entognathous Hexapod Lineages

The data set *M_Ento*, which is decisive for addressing relationships among the three entognathous groups, Protura, Collembola, and Diplura (73 taxa, 117 genes, 32,883 aligned aa sites), moderately supported a clade Protura + Diplura (Nonoculata) (fig. 1). This is compatible with the results of the FcLM approach (topology T_1 favored, fig. 2). Tree reconstruction supported Collembola as closest relatives to a clade comprising Nonoculata and Ectognatha. The clade Nonoculata + Ectognatha received moderate support (fig. 1; supplementary fig. S2, Supplementary Material online).

Our tree reconstructions based on a selected optimal subset (SOS) extracted from a complete data matrix by optimizing information content and data saturation in iterative steps of gene and/or taxon exclusion (see MARE manual; Meusemann et al. 2010; Meyer and Misof 2010) (62 taxa, 253 genes, alignment length 55,429 aa positions) yielded monophyletic Entognatha with moderate support and Nonoculata with low support (fig. 3*a* and table 2; supplementary fig. S3*a*, Supplementary Material online). It should be kept in mind that this SOS is indecisive for addressing the relationships of Entognatha, with only one-third of all genes (79) of this data set being covered by all three entognathous groups (supplementary table S3, Supplementary Material online). The tree based on the data set SOS_{ω}, in which

Table 1. Primarily Wingless Hexapod Species Included in This Study, and Their Number of OGs in the Original Supermatrix and in Three Data Subsets.

Order	Family	Species	Source	No. of Contigs	Total no. of OGs	No. of OGs in *M_Ento*	No. of OGs in SOS	No. of OGs in SOS_ω
Protura	Acerentomidae	*Acerentomon* sp.[a]	NCBI[a]	1,999	191	117	91	12
Diplura	Campodeidae	*Campodea fragilis*	NCBI	6,407	370	77	116	64
Diplura	Japygidae	*Megajapyx* sp.	this study	57,602	547	105	164	89
Collembola	Neanuridae	*Anurida maritima*	NCBI	3,504	328	55	105	60
Collembola	Onychiuridae	*Onychiurus arcticus*	NCBI	9,981	795	103	183	114
Collembola	Isotomidae	*Cryptopygus antarcticus*	NCBI	1,897	199	49	78	35
Collembola	Isotomidae	*Folsomia candida*	NCBI	5,967	442	60	122	78
Collembola	Entomobryidae	*Orchesella cincta*	NCBI	754	52	10	—	—
Archaeognatha	Machilidae	*Lepismachilis y-signata*	NCBI	2,288	270	60	107	54
Zygentoma	Lepismatidae	*Tricholepisma aurea*	NCBI	344	54	22	—	—
Zygentoma	Lepismatidae	*Thermobia domestica*	this study	45,358	682	96	194	124

NOTE.—*M_Ento* is the decisive data set in which all OGs are covered by Protura, Diplura, and Collembola; SOS and SOS_ω are indecisive to address the relationships of entognathous hexapod orders.

[a]*Acerentomon* sp.: erroneously assigned as *A. franzi* in Meusemann et al. (2010) and NCBI.

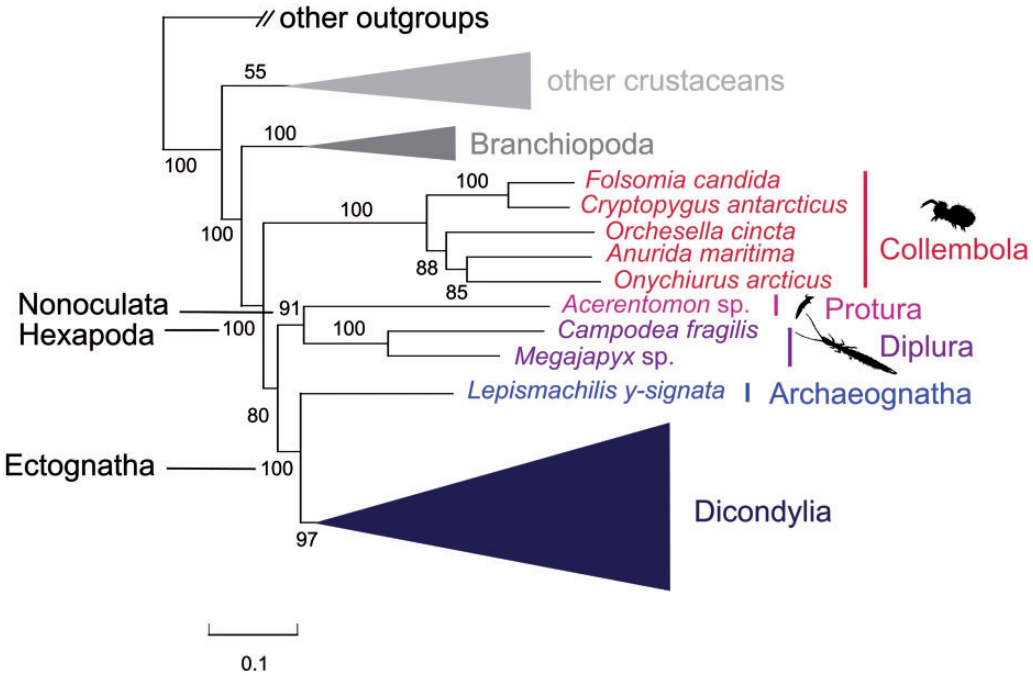

FIG. 1. Simplified phylogenetic tree of the decisive data set *M_Ento*. Best ML tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA), based on 117 OGs that are covered by Protura, Diplura, and Collembola. BS is derived from 1,000 bootstrap replicates. Rogue taxa (supplementary material [section 4], Supplementary Material online) were pruned prior to tree inference. The tree was rooted with *Capitella* sp. For the full tree, see supplementary figure S2, Supplementary Material online.

these 79 genes were removed to artificially create a maximally indecisive data set, showed Entognatha with strong support (table 2) and additionally, diplurans were paraphyletic with respect to Protura (fig. 3*b*; supplementary fig. S3*b*, Supplementary Material online). Both SOS data sets (11 taxa from the supermatrix, which included the collembolan *Orchesella cincta* were removed in the optimization process) did not contain any rogue taxa, that is, taxa that assume incongruent phylogenetic positions in a set of bootstrap trees (Aberer and Stamatakis 2011) (supplementary material [section 3], Supplementary Material online). Tree reconstructions of all data sets strongly supported monophyletic Ectognatha and monophyletic Hexapoda (table 2).

Incongruent Signal among Genes

Based on the *M_Ento* data set, the FcLM approach helped to identify a predominant signal for topology T_1 (Protura + Diplura) – (Collembola + remaining taxa) in 51 genes (12,548 aligned aa positions) (data set *M_Nono*, derived from Nonoculata), a predominant signal for topology T_2

(Protura + Collembola) – (Diplura + remaining taxa) in 35 genes (11,789 aligned aa positions) (data set *M_Elli*, derived from Ellipura), and a predominant signal for topology T_3 (Diplura + Collembola) – (Protura + remaining taxa) in 31 genes (8,546 aligned aa positions) (data set *M_DiCo*) (fig. 4*a* and *b*). Tree inferences from data sets *M_Nono*, *M_Elli*, and *M_DiCo* (rogue taxa pruned, see Materials and Methods section) yielded maximal BS support for Nonoculata, Ellipura, and Diplura + Collembola, respectively (table 2; supplementary fig. S4, Supplementary Material online). However, although tree reconstruction of our data subsets *M_Nono*, *M_Elli*, and *M_DiCo* showed maximal BS support for incongruent topologies among the entognathous insect orders, the results from the FcLM approach indicated that signal for alternative topologies was present in all data sets (fig. 4*a* and *b*; supplementary table S4, Supplementary Material online), which is not reflected by the trees. To identify possible reasons for incongruent signal among genes, we assessed the correlation between functional classes of genes and the different phylogenetic hypotheses that are supported by the data subsets. We found no correlation (supplementary material [section 4], table S5 and fig. S5, Supplementary Material online). Additionally, we tested whether model misspecification can explain the observed incongruence among genes and analyzed the data set *M_Ento* and data subsets *M_Nono*, *M_Elli*, and *M_Dico* using partitioned phylogenetic analyses (Minh et al. 2013) with the best model selected for each gene (partition) separately (supplementary material [section 5], table S6, and figs. S6–S9, Supplementary Material online). With respect to the phylogenetic relationships addressed in our study, resulting topologies did not differ from unpartitioned analyses, and BS only differed to a minor degree (table 2).

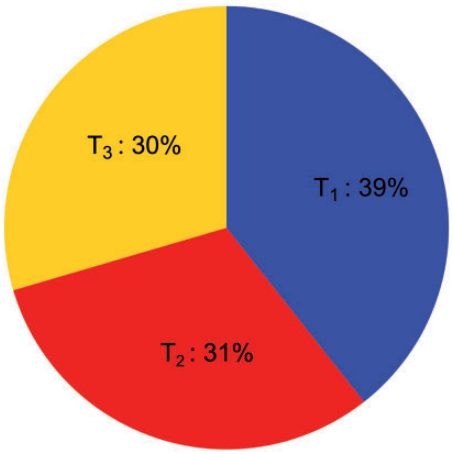

Fig. 2. Results of the FcLM for all OGs in data set *M_Ento*. The chart shows the proportion of quartets (summed up for 117 OGs) that show predominant support for T_1 ([Protura + Diplura] – [Collembola + remaining taxa], Nonoculata hypothesis, blue), T_2 ([Protura + Collembola] – [Diplura + remaining taxa], Ellipura hypothesis, red), and T_3 ([Diplura + Collembola] – [Protura + remaining taxa], yellow), see fig. 5. Quartets mapping in remaining Voronoi cells (gray) and T* (fig. 5) were not considered.

Discussion

The Importance of Data Set Decisiveness

Incongruences in proposed relationships among Protura, Collembola, and Diplura in the studies of Meusemann et al. (2010) and von Reumont et al. (2012), which both supported monophyly of Entognatha, motivated us to look for new approaches to uncover and analyze possible sources of incongruent signal in phylogenomic data sets.

Both SOS data sets in Meusemann et al. (2010) and von Reumont et al. (2012) were compiled with MARE (Meyer and Misof 2010) and were intended to address pancrustacean and arthropod relationships. Both data sets showed only low decisiveness for addressing the relationships of the three entognathous lineages: only 28 out of 128 genes in Meusemann et al. (2010) and 22 out of 316 genes in von Reumont et al. (2012) contained representatives of Protura, Diplura, and Collembola.

Despite low gene data coverage in both studies, the monophyly of Entognatha received high BS. By contrast, our decisive data set for addressing the relationships among these three insect orders lacks clear support for Entognatha

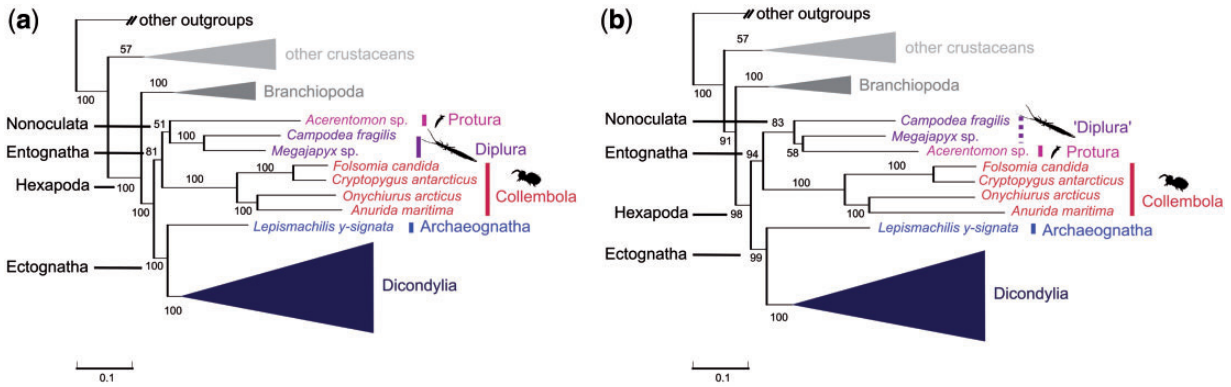

Fig. 3. Simplified phylogenetic trees of data sets SOS (*a*) and SOS_ω (*b*). Best ML tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) (*a*) based on 253 OGs, 79 of which are covered by Protura, Diplura, and Collembola (SOS) and (*b*) based on 174 OGs, none of which are covered by Protura, Diplura, and Collembola (SOS_ω). BS is derived from 1,000 bootstrap replicates. Trees were rooted with *Capitella* sp. For the full trees, see supplementary figure S3*a* and S3*b*, Supplementary Material online.

Table 2. BS (%) for Selected Clades in Tree Reconstructions with Various Data Sets.

Clade	Data Set					Data Subset of *M_Ento*		
	M_Ento	SOS	SOS_ω	Meusemann et al. (2010)	von Reumont et al. (2012)	*M_Nono*	*M_Elli*	*M_DiCo*
Hexapoda	100 (100)	100	98	100	99	72 (100)	100 (100)	100 (100)
Diplura	100 (100)	100	—[a]	N.A.	N.A.	100 (100)	100 (100)	100 (100)
Collembola	100 (100)	100	100	100	100	100 (100)	100 (100)	100 (100)
(Protura, Diplura)[b]	91 (96)	51	83[a]	100	—	100 (100)	— (—)	— (—)
(Protura, Collembola)[c]	— (—)	—	—	—	98	— (—)	100 (100)	— (—)
(Diplura, Collembola)	— (—)	—	—	—	—	— (—)	— (–)	99 (100)
Entognatha	— (—)	81	94	86	98	— (—)	— (—)	— (—)
((Protura, Diplura), Ectognatha)	80 (96)	—	—	—	—	98 (100)	— (—)	— (—)
((Collembola, Diplura), Ectognatha)	— (—)	—	—	—	—	— (—)	— (—)	60 (83)
(Diplura, Ectognatha)	— (—)	—	—	—	—	— (—)	66 (100)	— (—)
Ectognatha	100 (100)	100	99	100	100	100 (100)	100 (100)	95 (84)

NOTE.—BS was assessed with RAxML from 1,000 bootstrap replicates (see Materials and Method). BS printed in brackets was assessed from partitioned ML analyses of data sets *M_Ento*, and its subsets using the Ufboot algorithm of IQ-TREE with 5,000 bootstrap replicates (supplementary material [section 5], Supplementary Material online). *M_Ento* is the decisive data set in which all OGs are covered by Protura, Diplura, and Collembola; SOS, SOS_ω, and the data sets from Meusemann et al. (2010; data set SOS, ML tree) and von Reumont et al. (2012; data set SOS, ML tree Set 1_{red}) are indecisive to address the relationships of entognathous hexapod orders. *M_Nono*, *M_Elli*, and *M_DiCo* are subsets of *M_Ento* with predominant signal for different topologies and point out conflict of signal among genes.

[a]Diplurans are paraphyletic: *Campodea* + (*Acerentomon*,*Megajapyx*).

[b]Nonoculata hypothesis.

[c]Ellipura hypothesis.

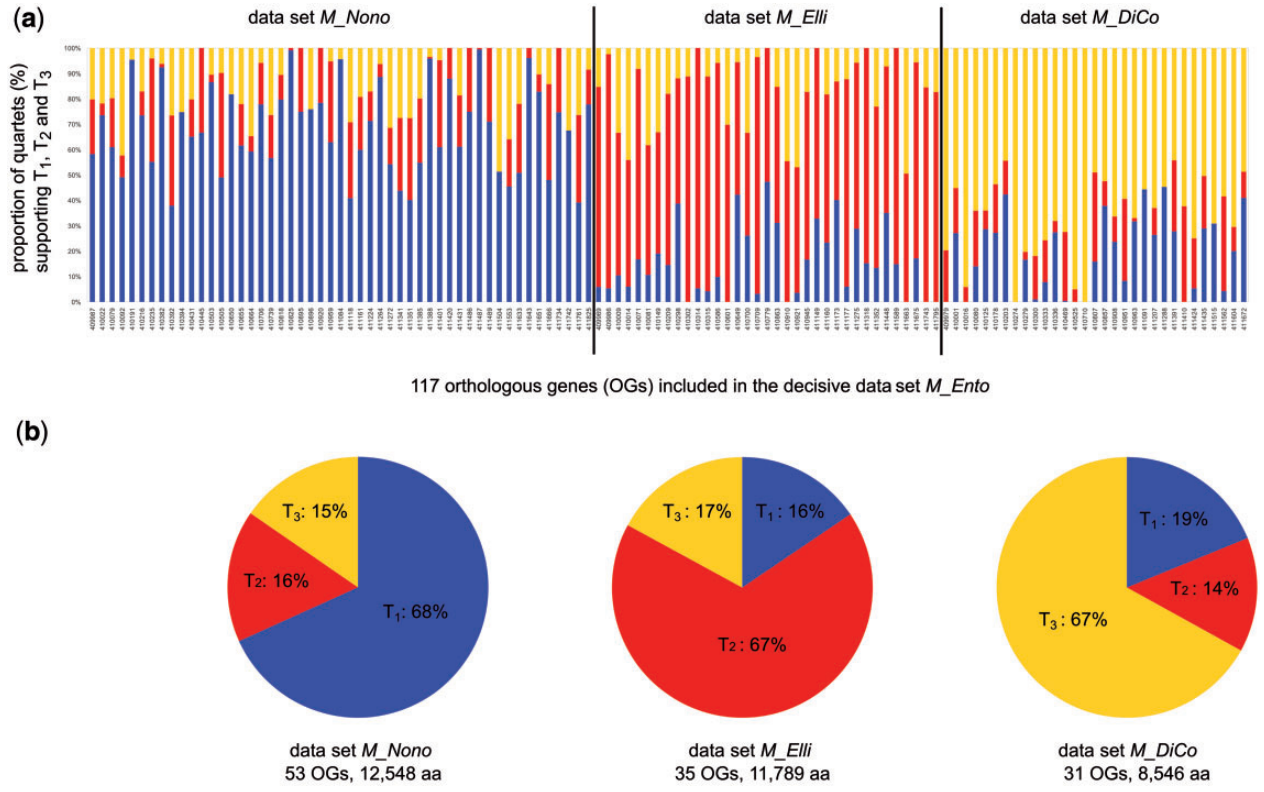

FIG. 4. Detailed results of the FcLM *Mapping* for all OGs included in data set *M_Ento* and data subsets *M_Nono*, *M_Elli*, *M_DiCo*. (*a*) Histogram of FcLM results. Each bar refers to an OG (for OG-IDs, see supplementary table S2, Supplementary Material online). *Y* axis: amount of quartets (in %), that predominantly support T_1 ([Protura + Diplura] – [Collembola + remaining taxa], blue), T_2 ([Protura + Collembola] – [Diplura + remaining taxa], red), and T_3 ([Diplura + Collembola] – [Protura + remaining taxa], yellow), quartets that show ambiguous support are not considered (fig. 5). OGs with predominant support for T_1 are classified into data set *M_Nono* (51 genes, 12,548 aligned aa positions); OGs with predominant support for T_2 are classified into data set *M_Elli*; (35 genes, 11,789 aligned aa positions); OGs with predominant support for T_3 are classified into data set *M_DiCo* (31 genes, 8,546 aligned aa positions). (*b*) FcLM results for data set *M_Nono* (left), *M_Elli* (middle), and *M_DiCo* (right). Each chart shows the proportion of quartets (summed up for the OGs included in the data sets) that show predominant support for T_1, T_2, and T_3 (see above and fig. 5). Quartets that show ambiguous support (fig. 1) are not considered.

(fig. 1). This puzzling result might be explained by the presence of an uneven distribution of missing data. We gained indirect evidence for this hypothesis with the analyses of the worst case data set SOS_{ω}. This data set is maximally indecisive for testing the monophyly of Entognatha, that is, none of the included genes were common to all three entognathous insect groups. Any inferred support for this clade in the SOS_{ω} analysis can be considered an artifact. Remarkably, bootstrapping delivered high, clearly artificial support for monophyletic Entognatha in the SOS_{ω} tree (fig. 3*b*).

We conclude from this indirect evidence that the support for Entognatha in Meusemann et al. (2010), von Reumont et al. (2012) and in our data set indecisive concerning this question (fig. 3*a*) probably results from an artificial signal due to uneven distribution of missing data (Philippe et al. 2011) among Protura, Diplura, and Collembola.

Based on the analyses of the decisive and indecisive data sets, we reject the hypothesis that missing data are unproblematic as long as many characters have been sampled overall (Wiens 2006). Missing data can be misleading as shown by the worst case SOS_{ω} data set analysis, in which relationships received high BS although the data set was maximally indecisive. Therefore, we strongly advocate the exclusive use of decisive data sets in phylogenomic studies.

Incongruent Signal between Genes in a Multigene Data Set

Even decisive data sets can contain incongruent signal (Degnan and Rosenberg 2009; Knowles 2009; Philippe et al. 2011). Using FcLM, we identified groups of genes that support different relationships of Protura, Collembola, and Diplura in the decisive data set *M_Ento* (fig. 4*a* and *b*). Additionally, we assessed conflict within the data with split analyses relying on NeighborNetworks (supplementary material [section 6] and figs. S10–S13, Supplementary Material online). This analysis corroborates the results of FcLM that all analyzed data sets did contain incongruent signal. Additional to the problem of indecisiveness discussed earlier, this incongruent signal among genes may partly be responsible for the contradictory results of Meusemann et al. (2010) and von Reumont et al. (2012). However, incongruent signal among genes is difficult to address and rectify. We analyzed two potential sources of conflict and can conclude that both can be excluded. First, we tested for homoplasy due to analogous selection regimes in functional complexes but found no correlation between predicted gene function and phylogenetic signal (supplementary material [section 4], fig. S5, and table S5, Supplementary Material online). Second, we were able to indirectly exclude model misspecifications as sources of incongruent signal because unpartitioned and partitioned maximum likelihood (ML) analyses yielded topologically congruent results and almost identical BS (table 2; supplementary material [section 5], table S6, and figs. S6–S9, Supplementary Material online). With respect to the FcLM, it may well be that this likelihood mapping approach selects sets of genes with congruent substitution processes. A possible solution, but certainly not a fully satisfying one, would be to increase the number of genes to minimize noise and confounding signal.

Relationships of Protura, Collembola, and Diplura

Monophyly of Entognatha

The monophyly of Entognatha has never been maximally supported and this has not changed in our analyses (table 2). Studies encompassing representatives of Protura, Collembola, and Diplura are limited to only a few analyses (Colgan et al. 1998; Carapelli et al. 2000; Edgecombe et al. 2000; Giribet et al. 2001, 2005). Monophyletic Entognatha were recovered in all recent studies based on nuclear rRNA genes (Gao et al. 2008; Dell'Ampio et al. 2009; von Reumont et al. 2009; Mallatt et al. 2010). However, BS was low, which was either explained by character choice (Dell'Ampio et al. 2009) or the influence of nonstationary processes across taxa (von Reumont et al. 2009). From the morphological point of view, most apomorphies suggesting the monophyly of Entognatha represent reductions (malpighian papillae vs. tubules; reduction to loss of compound eyes). The only exception is the evolution of mouthparts that are concealed in gnathal pouches (Beutel and Gorb 2006). Diplura as closest relatives to Ectognatha is the only relation that contradicts monophyletic Entognatha, and for which morphological evidence has been published (Kukalová-Peck 1991; Koch 1997; Beutel and Gorb 2006; Dallai et al. 2011). In general, morphological support for any clade encompassing more than one of the entognathous lineages Protura, Diplura, and Collembola is weak, largely because character polarization is problematic. This is due to the lack of applicability of characters and/or missing comparative studies in the crustacean groups that are discussed to be most closely related to Hexapoda (Szucsich and Pass 2008).

Ellipura versus Nonoculata

Molecular analyses mostly support Nonoculata (Protura + Diplura) (Giribet et al. 2004; Luan et al. 2005; Kjer et al. 2006; Mallatt and Giribet 2006; Misof et al. 2007; Dell'Ampio et al. 2009; von Reumont et al. 2009; Mallatt et al. 2010; see Dell'Ampio et al. 2011 for a review) while most morphologists merge Protura and Collembola into Ellipura (Börner 1910; Hennig 1953; Kristensen 1981, 1997; Kukalová-Peck 1987; Bitsch and Bitsch 2000, 2004; Beutel and Gorb 2006). Molecular evidence for Ellipura is weak and limited to three mitochondrial single-gene analyses (Shao et al. 1999; Carapelli et al. 2000; Zhang et al. 2001), and morphological support for Nonoculata is nearly missing (Szucsich and Pass 2008). These controversies call for phylogenomic approaches.

The majority of the 117 genes that compose the decisive data set *M_Ento* contain predominant signal for Nonoculata (fig. 4*a*). Also, the FcLM analysis of *M_Ento* (fig. 2) and the phylogenetic tree of *M_Ento* (fig. 1) yielded monophyletic Nonoculata, albeit not being well supported. In summary, Nonoculata is slightly favored over Ellipura in our study, but the question of the phylogenetic relationships of the three entognathous hexapod orders remains unsettled.

Conclusions

Clades may be incorrect, even if receiving high BS support (e.g., monophyly of Entognatha in Meusemann et al. [2010], von Reumont et al. [2012], and in data sets SOS and SOS_{ω} of

this study). This is a trivial conclusion and different reasons are mentioned in the literature (Lehtonen 2011, Simmons and Freudenstein 2011). We show that an uneven distribution of missing data (i.e., the use of indecisive data sets) can lead to strongly supported, yet incorrect, clades. To avoid misleading phylogenetic conclusions from seemingly robust trees based on phylogenomic data sets, we advise 1) using only data sets that are decisive for the phylogenetic question of interest, 2) including an alternative measure of support (Salichos and Rokas 2013); our method of choice was the FcLM approach, and 3) analyzing and documenting the inferred incongruence of signal between genes.

In our decisive data set, we found strong incongruence among genes that is neither correlated with functional classes of genes nor with model misspecifications in unpartitioned analyses. Based upon these notes of caution, we found no signal for the monophyly of Entognatha, and we found no strong signal for Ellipura or Nonoculata despite extending our data set with additional data from key taxa. In other words, the phylogeny and evolution of early hexapods remains enigmatic. Despite this, we show that there are valuable lessons to be learned from the analyses of phylogenomic data of primarily wingless insects, particularly in terms of incongruence among genes and data decisiveness.

Materials and Methods

Taxon Sampling and New Transcriptome Data

Our taxon sampling included 73 species: 46 hexapods, and, as outgroup species, 25 crustaceans, the chelicerate *Ixodes scapularis*, and the polychaete worm *Capitella* sp., both present in the reference set of taxa used for orthology assignment (discussed later). Transcriptome assemblies of 71 species were obtained from the Deep Metazoan Phylogeny database (http://www.deep-phylogeny.org/, last accessed November 4, 2013). We only used species for which more than 1,000 contigs were available (status: December 2011), with two exceptions: the springtail *Orchesella cincta* (Collembola, Entomobryidae, 754 contigs) and the silverfish *Tricholepisma aurea* (Zygentoma, Lepismatidae, 344 contigs), the only publicly available zygentoman transcriptome assembly (supplementary table S1, Supplementary Material online).

We generated new transcriptome data for *Megajapyx* sp. (Diplura, Japygidae) and the firebrat *Thermobia domestica* (Packard 1837) (Zygentoma, Lepismatidae) (table 1). Extraction of RNA, complementary deoxyribonucleic acid (cDNA) library construction, library normalization, and 454 pyrosequencing of ~1,000,000 ESTs per species using the GS-FLX Titanium System, ROCHE were carried out at the Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. Vector clipping, trimming, and soft masking of raw reads and assembly into contigs was conducted at the Center for Integrative Bioinformatics (CIBIV), Vienna, Austria. Steps at the MPIMG and the CIBIV were done as described in von Reumont et al. (2012) and Simon et al. (2012), for details see supplementary material (section 1; Supplementary Material online). Raw sequence reads were deposited at the National Center for Biotechnology Information (NCBI), Sequence Read Archive (accession numbers *Megajapyx* sp.: SRR400673; *T. domestica*: SRR400672). Transcriptome assemblies of *Megajapyx* sp. (accession numbers JT047774–JT094274) and *T. domestic*a (accession numbers T494145–JT533227) were deposited at the Transcriptome Shotgun Assembly (TSA) Database, NCBI Bioproject ID PRJNA81579 and PRJNA81581 (http://www.ncbi.nlm.nih.gov/bioproject, last accessed November 4, 2013). For submission, we excluded contigs shorter than 200 bp, according to the submission guidelines; the full transcriptome assemblies are available at http://zfmk.de/bioinformatics/Full_Transcriptome_Assemblies.zip (last accessed November 4, 2013).

Orthology Assignment

To identify 1:1 OGs in our transcriptome assemblies, we used the Hidden Markov Model based Search for Orthologs using Reciprocity (HaMStR) pipeline (Ebersberger et al. 2009; http://www.deep-phylogeny.org/hamstr/, last accessed November 4, 2013), version 4. As reference set for clusters of OGs, we used a set of 1,886 1:1 OGs (represented by amino acid sequences) based on five reference species (supplementary material [section 2] and table S2, Supplementary Material online). We defined orthology being present if bi-directional best hits were found between our transcript sequences and the reference species *Daphnia pulex*, *Ixodes scapularis*, *Apis mellifera*, and *Capitella* sp.

Alignment, Alignment Masking, and Concatenation

We aligned amino acid sequences using MAFFT L-INS-i (Katoh and Toh 2008) v.6.850 for each gene separately. Afterwards, randomly similar aligned sections were identified with a modified version of ALISCORE (Misof B and Misof K 2009; Kück et al. 2010; Meusemann et al. 2010; for modifications, see Meusemann et al. 2010) using the following options: default sliding window size; *-r*: maximum number of pairwise sequence comparisons; *-e*: special scoring for gappy amino acid data. Identified randomly similar aligned sections were masked with ALICUT v.2.0 (Kück 2009; www.utilities.zfmk.de, last accessed November 4, 2013). Masked alignments were concatenated into supermatrices with FASconCAT v.1.0 (Kück and Meusemann 2010).

Design of Decisive and Indecisive Data Sets

We extracted all genes from the supermatrix that contain at least one representative of each 1) Protura, 2) Diplura, 3) Collembola, and 4) remaining species to generate a decisive data set among entognathous lineages. The resulting data set is called *M_Ento*.

We generated two additional data subsets from the original supermatrix: 1) A so-called selected optimal subset (SOS), generated with MARE v.0.1.2-rc (Meyer and Misof 2010; http://mare.zfmk.de, last accessed November 4, 2013), applying taxon weighting *-t 1.5*. This approach is analogous to Meusemann et al. (2010) and von Reumont et al. (2012). 2) From this SOS data set, we compiled a data set called SOS_{ω} by removing all genes that were covered by all three entognathous lineages to receive a maximally indecisive

"worst case" data set in which each gene contained maximally two entognathous lineages.

Four-Cluster Likelihood Mapping

Additional to tree reconstruction with BS, we applied the FcLM approach using the *M_Ento* data set (Strimmer and von Haeseler 1997). We binned sequenced species into four clusters: 1) Protura (1 species), 2) Diplura (2 species), 3) Collembola (5 species), and 4) remaining species (65 species) (supplementary table S1, Supplementary Material online). Next, we 1) estimated the tree-likeness of each gene, that is the amount of quartets that showed support for one out of the three possible topologies and 2) evaluated which of the three possible topologies was supported by the majority of those quartets (predominant support): T_1 (Protura + Diplura) and (Collembola + remaining taxa), T_2 (Protura + Collembola), and (Diplura + remaining taxa), or T_3 (Diplura + Collembola) and (Protura + remaining taxa) (fig. 5). The competing hypotheses of Meusemann et al. (2010) and von Reumont et al. (2012) are represented by either T_1 (Nonoculata hypothesis) or T_2 (Ellipura hypothesis); the third topology T_3 does not represent a currently debated hypothesis. FcLM was conducted using TREE-PUZZLE v.5.2 (Schmidt et al. 2002; http://www.tree-puzzle.de, last accessed November 4, 2013), applying the BLOSUM62 substitution matrix (Henikoff S and Henikoff JG 1992) as the BLOSUM62 substitution matrix is implemented in the software MARE (Meyer and Misof, 2010; http://mare.zfmk.de, last accessed November 4, 2013).

For each gene in the data set *M_Ento*, we calculated the proportions of quartets that predominantly supported either topology T_1, T_2, or T_3. According to the topology that was supported by the majority of quartets, we classified each gene into one of three groups, supporting Nonoculata, Ellipura, or Diplura + Collembola (fig. 5 and supplementary table S4, Supplementary Material online). Quartets for which the support remained ambiguous (T_{12}, T_{23}, T_{13}, and T*; fig. 5) were not used for classification (see supplementary fig. S14 [Supplementary Material online] for the results with all quartets). All classified genes (supplementary table S4, Supplementary Material online) were subsequently concatenated into three submatrices called *M_Nono* (genes supporting Nonoculata), *M_Elli* (genes supporting Ellipura), and *M_DiCo* (genes supporting Diplura + Collembola).

Phylogenetic Tree Inference

ML tree reconstruction was done from all data sets: *M_Ento*, *M_Nono*, *M_Elli*, and *M_DiCo*, SOS, and SOS_ω (discussed earlier). We estimated evolutionary models for each data set with ModelGenerator v.0.85 (Keane et al. 2006). The

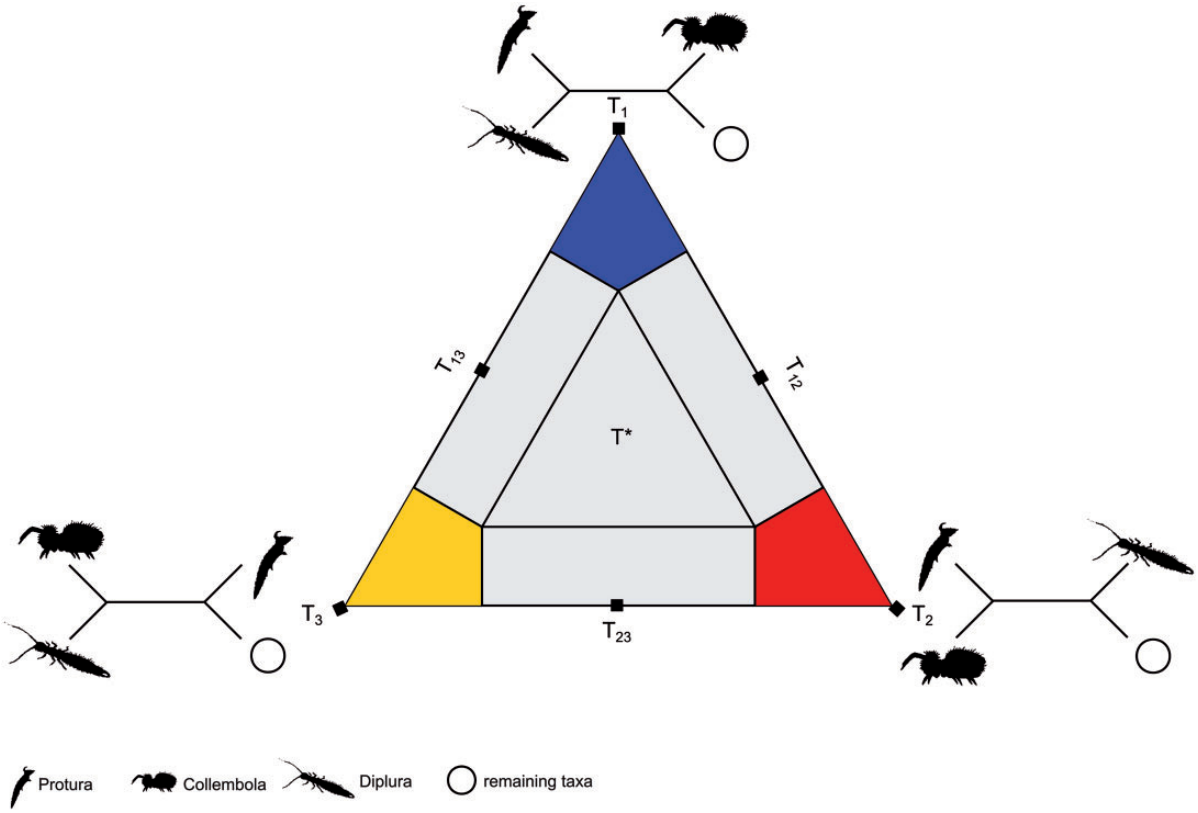

FIG. 5. 2D simplex graph. Voronoi cells are areas, in which quartets show predominant or maximal support for either of the three topologies T_1, T_2, T_3, or in which quartets show ambiguous support T_{12}, T_{13}, T_{23}, and T*. For further explanations, refer to Strimmer and von Haeseler (1997, fig. 3). Voronoi cell corresponding to T_1 (blue): quartets show support for (Protura + Diplura) – (Collembola + remaining taxa); Voronoi cell corresponding to T_2 (red): quartets show support for (Protura + Collembola) – (Diplura + remaining taxa); Voronoi cell corresponding to T_3 (yellow): quartets show support for (Diplura + Collembola) – (Protura + remaining taxa); Voronoi cells corresponding to T_{12}, T_{13}, T_{23} (gray) do not show clear support for T_1, T_2, and T_3; in T* all topologies are equally likely.

best fitting model was selected based upon the Akaike Information Criterion (AIC; Akaike 1974). ML trees were inferred with RAxML (Stamatakis 2006), v.7.2.8-ALPHA, HYBRID (Ott et al. 2007; Pfeiffer and Stamatakis 2010) using the CAT model of rate heterogeneity (Stamatakis 2006) and the LG protein substitution matrix (Le and Gascuel 2008). Final tree searches were conducted under the GAMMA model of rate heterogeneity (Yang 1996). Bootstrap analyses were performed with the rapid algorithm (Stamatakis 2006), which also included subsequent searches for the best scoring ML tree. We obtained BS for each node from 1,000 rapid bootstrap replicates, and checked a posteriori if sufficient bootstrap trees were computed using the bootstopping criteria (Pattengale et al. 2010, default settings). ML analyses were conducted on a Linux cluster at the Cologne High Efficient Operating Platform for Science (CHEOPS), Regionales Rechenzentrum Köln (RRZK), using eight nodes with 12 cores each.

After tree inference, we scrutinized our trees for rogue taxa (Aberer et al. 2013; Aberer and Stamatakis 2011, see supplementary material [section 3], figs. S2 and S4, table S7, Supplementary Material online, for details and results). We removed sequences corresponding to taxa that were identified as rogues from the concatenated alignments and repeated the tree inferences. All trees were edited with Treegraph v.2.0 (Stöver and Müller 2010), and rooted with *Capitella* sp. Data sets are deposited at Dryad: http://doi.org/10.5061/dryad.mk8p7 (last accessed November 4, 2013).

Supplementary Material

Supplementary material (sections 1–6), tables S1–S7, and figures S1–S14 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

K.M. and B.Mi. provided EST data for *Thermobia domestica*, and G.P., E.D.A. and N.U.S. for *Megajapyx* sp. J.B. and M.P. wrote Perl scripts for the analysis pipeline and B.Me. adopted the FcLM approach. Processing and sequence assembly of EST data were performed by I.E. and A.v.H. The study was conceived by B.Mi., K.M., and E.D.A. Orthology assignment and subsequent analyses were conducted by K.M. Rogue taxa analyses were performed by A.J.A. and A.S. Partitioned ML analyses were provided by B.Q.M. The manuscript was written by K.M., N.U.S., R.S.P., E.D.A., and B.Mi with useful comments and revisions from A.J.A., B.Q.M., M.G.W., A.v.H., I.E., A.S., and G.P. All authors read and approved the final manuscript. The authors thank Martin Streinzer for help in collecting *Megajapyx* sp. They acknowledge Michael Kube and Richard Reinhardt (MPIMG, Berlin, Germany) for extraction of RNA, generating cDNA libraries, and ESTs. They thank Sascha Strauss for help with processing and assembling the EST data and John Plant (University of Vienna, Austria) for examining the English. They acknowledge the Cologne High Efficient Operating Platform for Science (CHEOPS, HPC cluster at the RRZK, University of Cologne, Cologne, Germany; available from: http://rrzk.uni-koeln.de/cheops.html) for the opportunity to perform analyses. Finally, the authors thank two anonymous reviewers for helpful comments that considerably improved the manuscript. This work was supported by the Austrian Science Foundation (FWF) grant P 20497-B17 to E.D.A., N.U.S. and G.P., the German Science Foundation (DFG): priority program SPP 1174 "Deep Metazoan Phylogeny" (http://www.deep-phylogeny.org), and by institutional funding of the Heidelberg Institute for Theoretical Studies. A.v.H. and I.E. were funded by the German Science Foundation (DFG) grant HA1628/9. A.v.H. and B.Q.M. were supported by the Austrian Science Foundation (FWF) grant I760. K.M. and B.M. were funded by the German Science Foundation (DFG) grant MI 649/6.

References

Aberer AJ, Komprass D, Stamatakis A. 2013. Pruning Rogue Taxa improves phylogenetic accuracy: an efficient algorithm and web service. *Syst Biol.* 62(1):162–166.

Aberer AJ, Stamatakis A. 2011. A simple and accurate method for rogue taxon identification. *IEEE BIBM* 2011:118–122.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Aut Control.* 19:716–723.

Bitsch C, Bitsch J. 2000. The phylogenetic interrelationships of the higher taxa of apterygote hexapods. *Zool Scr.* 29:131–156.

Bitsch C, Bitsch J. 2004. Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. *Zool Scr.* 33:511–550.

Beutel RG, Gorb SN. 2006. A revised interpretation of attachment structures in Hexapoda with special emphasis on Mantophasmatodea. *Arthropod Syst Phylogeny.* 64:3–25.

Börner C. 1910. Die phylogenetische Bedeutung der Protura. *Sdr Biolog Centralbl.* 30:633–641.

Carapelli A, Frati F, Nardi F, Dallai R, Simon C. 2000. Molecular phylogeny of the apterygote insects based on nuclear and mitochondrial genes. *Pedobiologia* 44:361–373.

Colgan DJ, McLauchlan A, Wilson GDF, Livingston S, Macaranas J, Edgecombe D, Cassis G, Gray MR. 1998. Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Aust J Zool.* 46:419–437.

Dallai R, Mercati D, Carapelli A, Nardi F, Machida R, Sekiya K, Frati F. 2011. Sperm accessory microtubules suggest the placement of Diplura as the sister-group of Insecta s.s. *Arthropod Struct Dev.* 40: 77–92.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends Ecol Evol.* 24: 332–340.

Dell'Ampio E, Szucsich NU, Carapelli A, Frati F, Steiner G, Steinacher A, Pass G. 2009. Testing for misleading effects in the phylogenetic reconstruction of ancient lineages of hexapods: influence of character dependence and character choice in analyses of 28S rRNA sequences. *Zool Scr.* 38:155–170.

Dell'Ampio E, Szucsich NU, Pass G. 2011. Protura and molecular phylogenetics: status quo of a young love. *Soil Organisms* 83:347–358.

Ebersberger I, Strauss S, Von Haeseler A. 2009. HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol Biol.* 9:157.

Edgecombe GD, Wilson GDF, Colgan DJ, Gray MR, Cassis G. 2000. Arthropod cladistics: combined analysis of Histone H3 and U2 snRNA sequences and morphology. *Cladistics* 16:155–203.

Gao Y, Bu Y, Luan Y. 2008. Phylogenetic relationships of basal hexapods reconstructed from nearly complete 18S and 28S rRNA gene sequences. *Zool Sci.* 25:1139–1145.

Giribet G, Edgecombe GD, Carpenter JM, D'Haese CA, Wheeler WC. 2004. Is Ellipura monophyletic? A combined analysis of basal hexapod relationships with emphasis on the origin of insects. *Org Div Evol.* 4:319–340.

Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* 413:157–161.
Giribet G, Richter S, Edgecombe GD, Wheeler WC. 2005. The position of crustaceans within Arthropoda—evidence from nine molecular loci and morphology. In: Koenemann S, Jenner RA, editors. Crustacea and arthropod relationships. Crustacean issues 16: Festschrift for F. R. Schram. Boca Raton (FL): Taylor & Francis. p. 307–352.
Giribet G, Wheeler WC. 2001. Some unusual small-subunit ribosomal DNA sequences of metazoans. *Am Mus Novit.* 3337:1–14.
Grimaldi DA. 2010. 400 million years on six legs: on the origin and early evolution of Hexapoda. *Arthropod Struct Dev.* 39:191–203.
Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* 89:10915–10919.
Hennig W. 1953. Kritische Bemerkungen zum phylogenetischen System der Insekten. *Beitr Entomol Sonderheft.* 3:1–85.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.
Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol.* 6:29.
Kjer KM, Carle FL, Litman J, Ware J. 2006. A molecular phylogeny of Hexapoda. *Arthropod Syst Phylogeny.* 64:3–44.
Knowles LL. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Syst Biol.* 58: 463–467.
Koch M. 1997. Monophyly and phylogenetic position of the Diplura (Hexapoda). *Pedobiologia* 41:9–12.
Kristensen NP. 1981. Phylogeny of insect orders. *Annu Rev Entomol.* 26: 135–157.
Kristensen NP. 1997. The ground plan and basal diversification of the hexapods. In: Fortey RA, Thomas RH, editors. Arthropod relationships, systematic association. Special volume series 55. London: Chapman & Hall. p. 281–293.
Kück P. 2009. ALICUT: a Perlscript which cuts ALISCORE identified RSS. Version 2.0 ed. Bonn (Germany): Department of Bioinformatics, Zoologisches Forschungs Museum A. Koenig (ZFMK). [cited 2013 Oct 30]. Available from: http://www.zfmk.utilities.de.
Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. *Mol Phylogenet Evol.* 56:1115–1118.
Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Waegele JW, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool.* 7:10.
Kukalová-Peck J. 1987. New Carboniferous Diplura, Monura and Thysanura, the hexapod ground plan, and the role of thoracic side lobes in the origin of wings (Insecta). *Can J Zool.* 65:2327–2345.
Kukalová-Peck J. 1991. Fossil history and the evolution of hexapod structures. In: Naumann ID, editor. Insects of Australia: a textbook for students and research workers. Melbourne (Australia): CSIRO, Melbourne University Press. p. 141–179.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.
Lehtonen S. 2011. Can sensitivity analysis help to detect long-branch attraction? *Mol Phylogenet Evol.* 61:899–903.
Luan Y, Mallatt JM, Xie R, Yang Y, Yin W. 2005. The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on ribosomal RNA gene sequences. *Mol Biol Evol.* 22:1579–1592.
Machida R. 2006. Evidence from embryology for reconstructing the relationships of hexapod basal clades. *Arthropod Syst Phylogeny.* 64:95–104.
Mallatt J, Giribet G. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol Phylogenet Evol.* 40:772–794.
Mallatt JM, Craig CW, Yoder MJ. 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. *Mol Phylogenet Evol.* 55:1–17.
Meusemann K, von Reumont BM, Simon S, et al. (16 co-authors). 2010. A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol.* 27:2451–2464.
Meyer B, Misof B. 2010. MARE: matrix reduction—a tool to select optimized data subsets from supermatrices for phylogenetic inference. Bonn (Germany): Zentrum für Molekulare Biodiversitätsforschung (zmb) am ZFMK. [cited 2013 Oct 30]. Available from: http://mare.zfmk.de (current version).
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol.* 30:1188–1195.
Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A. 2007. Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology* 110:409–429.
Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst Biol.* 58:21–34.
Ott M, Zola J, Stamatakis A, Aluru S. 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. Proceedings of the 2007 ACM/IEEE conference on Supercomputing. IEEE/ACM Supercomputing conference 2007 (SC2007); November 2007. Reno (NV): ACM.
Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. 2010. How many bootstrap replicates are necessary? *J Comput Biol.* 17:337–354.
Pfeiffer W, Stamatakis A. 2010. Hybrid MPI/Pthreads parallelization of the RAxML phylogenetics code. Paper presented at HICOMB workshop, held in conjunction with IPDPS 2010; April 2010; Atlanta, GA.
Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9: e1000602.
Regier JC, Shultz JW, Ganley AR, et al. (11 co-authors). 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol.* 57:920–938.
Regier J, Shultz J, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463:1079–1083.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327–331.
Sanderson MJ, McMahon MM, Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evol Biol.* 10:155.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
Shao HG, Zhang YP, Xie RD, Yin WY. 1999. Mitochondria cytochrome *b* sequences variation of Protura and molecular systematics of Apterygota. *Chin Sci Bull.* 44:2031–2036.
Simmons MP, Freudenstein JV. 2011. Spurious 99% bootstrap and jack-knife support for unsupported clades. *Mol Phylogenet Evol.* 61: 177–191.
Simon S, Narechania A, DeSalle R, Hadrys H. 2012. Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. *Genome Biol Evol.* 4:1295–1309.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
Steel M, Sanderson MJ. 2010. Characterizing phylogenetically decisive taxon coverage. *Appl Math Lett.* 23:82–86.
Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A.* 94:6815–6819.
Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7.

Szucsich NU, Pass G. 2008. Incongruent phylogenetic hypotheses and character conflicts in morphology: the root and early branches of the hexapodan tree. *Mitt Dtsch Ges Allg Angew Ent.* 16:415–430.

Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. *Annu Rev Entomol.* 57:449–468.

von Reumont BM, Meusemann K, Szucsich N, et al. (14 co-authors). 2009. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evol Biol.* 9:119.

von Reumont MB, Jenner RA, Wills MA, et al. (13 co-authors). 2012. Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. *Mol Biol Evol.* 29:1031–1045.

Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. *J Biomed Inform.* 39:34–42.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11(9):367–372.

Zhang Y, Zhang Y, Luan Y, Chen Y, Yin W. 2001. Phylogeny of higher taxa of Hexapoda according to 12sRNA sequences. *Chin Sci Bull.* 46: 840–842.

Current Biology 22, 1309–1313, July 24, 2012 ©2012 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2012.05.018

Report

Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera

Oliver Niehuis,[1,3,9,*] Gerrit Hartig,[1,4,9] Sonja Grath,[4,9] Hans Pohl,[5] Jörg Lehmann,[6] Hakim Tafer,[6] Alexander Donath,[1] Veiko Krauss,[6] Carina Eisenhardt,[7] Jana Hertel,[6] Malte Petersen,[1] Christoph Mayer,[1] Karen Meusemann,[1] Ralph S. Peters,[2] Peter F. Stadler,[6] Rolf G. Beutel,[5] Erich Bornberg-Bauer,[4] Duane D. McKenna,[8] and Bernhard Misof[1,*]

[1]Center for Molecular Biodiversity Research
[2]Department Arthropoda
Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany
[3]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
[4]Institute for Evolution and Biodiversity, Evolutionary Bioinformatics Group, University of Muenster, 48149 Muenster, Germany
[5]Institute of Systematic Zoology and Evolutionary Biology with Phyletisches Museum, University of Jena, 07743 Jena, Germany
[6]Institute of Computer Science, University of Leipzig, 04107 Leipzig, Germany
[7]Institute of Biology II, University of Leipzig, 04103 Leipzig, Germany
[8]Department of Biological Sciences, University of Memphis, Memphis, TN 38152, USA

Summary

The phylogeny of insects, one of the most spectacular radiations of life on earth, has received considerable attention [1–3]. However, the evolutionary roots of one intriguing group of insects, the twisted-wing parasites (Strepsiptera), remain unclear despite centuries of study and debate [1, 2, 4–11]. Strepsiptera exhibit exceptional larval developmental features, consistent with a predicted step from direct (hemimetabolous) larval development to complete metamorphosis that could have set the stage for the spectacular radiation of metamorphic (holometabolous) insects [1, 12, 13]. Here we report the sequencing of a Strepsiptera genome and show that the analysis of sequence-based genomic data (comprising more than 18 million nucleotides from nearly 4,500 genes obtained from a total of 13 insect genomes), along with genomic metacharacters, clarifies the phylogenetic origin of Strepsiptera and sheds light on the evolution of holometabolous insect development. Our results provide overwhelming support for Strepsiptera as the closest living relatives of beetles (Coleoptera). They demonstrate that the larval developmental features of Strepsiptera, reminiscent of those of hemimetabolous insects, are the result of convergence. Our analyses solve the long-standing enigma of the evolutionary roots of Strepsiptera and reveal that the holometabolous mode of insect development is more malleable than previously thought.

Results and Discussion

We sequenced the genome of *Mengenilla moldrzyki* (Figure 1A), a newly discovered species belonging to the early-divergent strepsipteran family Mengenillidae [14]. The draft genome of *M. moldrzyki* was sequenced from genomic DNA using 454-pyrosequencing technology to an estimated coverage of $\geq 14\times$. De novo assembly of the genome from the obtained reads produced 13,919 scaffolds and 87,021 nonredundant contigs spanning a total of 165 Mb. We inferred 16,772 ab initio models of nuclear-encoded protein-coding (NEPC) genes, of which 13,296 were supported by extrinsic evidence (e.g., transcripts). We also annotated protein domains, DNA methylation-related proteins, noncoding RNAs, and the complete mitochondrial genome (see Tables S1–S7 available online; Figure S1).

The Strepsiptera genome sequence data were exploited to test the following four current competing hypotheses about the phylogenetic origin of Strepsiptera (Figure 1B): (1) Strepsiptera are the sister group of all remaining insects with complete metamorphosis (Holometabola) [15], (2) Strepsiptera are the sister group of beetles (Coleoptera) [8], (3) Strepsiptera are a derived lineage of polyphagan beetles [9], and (4) Strepsiptera are the sister group of Diptera [5, 16, 17]. For this purpose, we assessed orthology among the predicted NEPC genes in the *M. moldrzyki* genome and those of 11 other insect species with sequenced genomes (representing Coleoptera, Diptera, Hymenoptera, Lepidoptera, and Acercaria; [18–28]) using a Markov Cluster algorithm implemented in the software OrthoMCL [29] (Figure 1B). In total, we identified 15,614 groups of orthologous NEPC genes; 4,485 of these groups contained sequences of at least one representative per insect order.

After removing ambiguously aligned sites (identified at the amino acid level), we evaluated the aligned amino acid and correspondingly aligned nucleotide sequences of the 4,485 groups of orthologous NEPC genes for their degree of substitutional saturation, relative compositional variance, and for the ratio of potential synapomorphic to potential autapomorphic characters. Compositional heterogeneity among sequences was lowest and the number of potentially informative characters for inferring inter- and intraordinal phylogenetic relationships was highest for RY-recoded (A and G → R; T and C → Y) second codon positions only, as compared to nonrecoded or differently recoded data sets or data subsets (Figure S3). The complete matrix of RY-recoded second codon positions from the 4,485 groups of orthologous NEPC genes consisted of approximately 1.8 million characters—the largest data set ever compiled for inferring the phylogenetic origin of Strepsiptera or any other insect order (Table S8).

We analyzed the RY-recoded second codon positions using maximum likelihood (ML) tree inference. The inferred phylogenetic tree (Figure 1B; Figure S2) was fully resolved and received maximal statistical support for all branches. All intra- and interordinal relationships are fully consistent with the current view of insect phylogenetic relationships [2]

[9]These authors contributed equally to this work
*Correspondence: o.niehuis.zfmk@uni-bonn.de (O.N.), b.misof.zfmk@uni-bonn.de (B.M.)

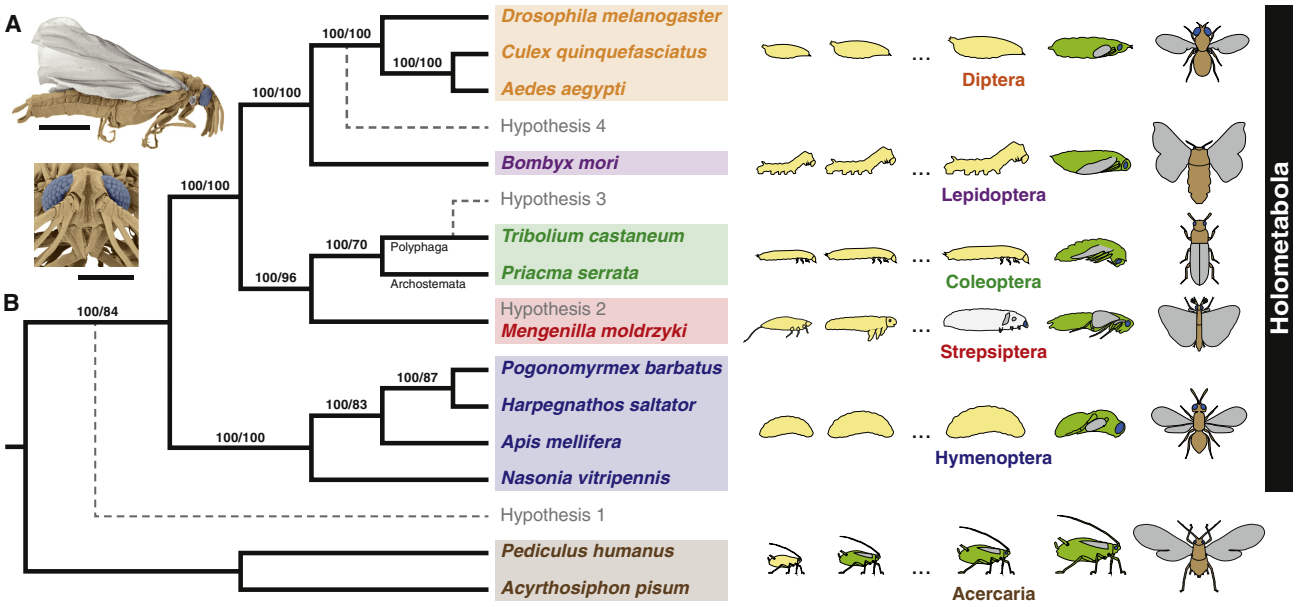

Figure 1. Evolutionary Origin of Twisted-Wing Parasites Inferred from Genomic Evidence

(A) *Mengenilla moldrzyki* male in lateral (top; scale bar represents 1 mm) and frontal (bottom; scale bar represents 500 μm) view (colored SEM micrographs; wings in gray, compound eyes in blue).

(B) Phylogenetic relationships and larval development of holometabolous insects. Numbers above branches are bootstrap support values from analyzing 4,485 protein-coding genes (RY-recoded 2nd codon positions only; ML optimality criterion) and 8,983 near intron pairs (MP optimality criterion). Recent hypotheses on the phylogenetic origin of Strepsiptera are shown in gray. Insect metamorphosis according to Truman and Riddiford [12], with pronymph (yellow) and nymphal stages (green) of insects with direct development (e.g., Acercaria) being equivalent to larval stages (yellow; nymphoid late larval stage of Strepsiptera in white) and pupa (green) of insects with complete metamorphosis (Holometabola); gray, wing buds and wings; blue, compound eyes.

(e.g., Hymenoptera are monophyletic and placed as sister to all remaining Holometabola, Diptera are monophyletic and next to Lepidoptera, and Coleoptera are more closely related to Diptera and Lepidoptera than to Hymenoptera). *M. moldrzyki* is placed as the sister taxon of the flour beetle, *Tribolium castaneum* (Figure 1B; note that the beetle *Priacma serrata* was not included at this step of our investigation). This result implies that Strepsiptera are either the sister group or a highly derived group of Coleoptera.

In addition to the primary sequence-based phylogenetic analyses, we investigated two genomic metacharacter sets as further evidence for the phylogenetic position of Strepsiptera. Specifically, we studied the phylogenetic signal of near intron pairs (NIPs) and that of gene order alignments along the lines with earlier studies that successfully used them to resolve the phylogeny of other holometabolous insects [30] and that of vertebrates [31]. The phylogenetic utility of NIPs is based on the fact that exons smaller than about 50 nucleotides are rare. Hence, introns found in close spatial proximity in orthologous genes of different species are unlikely to have ever coexisted in a single ancestral gene sequence. It is more likely that one intron is lost before the other intron is gained. We identified a total of 8,748 NIPs by studying the gene models of the 4,485 groups of orthologous NEPC genes. Phylogenetic analysis of the NIP characters, of which 1,173 were parsimony informative, under the maximum parsimony (MP) optimality criterion resulted in exactly the same topology as inferred from the primary sequence data (Figure 1B; Figure S2; note that *Priacma serrata* was not included at this step of our investigation).

The second independent approach for phylogenetic reconstruction was based on gene order information. Whereas this approach allows the genome of the species that has to be placed in the tree to be fragmented, all others must be fully assembled at the chromosome level. Accordingly, we used gene orders for *Anopheles gambiae* [32] (replacing *Aedes aegypti* and *Culex quinquefasciatus*), *Apis mellifera*, *Drosophila melanogaster*, and *Tribolium castaneum*, for which at least partial chromosome assemblies exist, and *Nasonia vitripennis*, for which we could exploit linkage map information to map a major fraction of its genome to individual linkage groups. This choice allowed for testing all four aforementioned conflicting phylogenetic scenarios under the assumption that Coleoptera and Diptera are more closely related to each other than to Hymenoptera. Phylogenetic analysis of the spatial arrangement of 791 and 1,433 genes, respectively, depending on whether or not *A. mellifera* with its partial chromosome assembly was part of the analysis, resulted in a topology consistent with those obtained with the previous two methods (Figure 2).

Given the overwhelming support for a close phylogenetic relationship of twisted-wing parasites and beetles, which is also reflected by (1) their high similarity in the protein domain content, (2) the primary sequence information of noncoding RNAs, and (3) the results of other phylogenetic analyses, including those of amino acid sequences (Figures S2 and S4), we next addressed the remaining question of whether or not Strepsiptera are highly derived beetles. For this purpose, we screened contig sequences from an early draft genome of *Priacma serrata* (Archostemata), a representative of an early-divergent lineage of beetles that is the sister group of all remaining extant Coleoptera [33, 34]. We identified the sequences of 3,018 of the 4,485 studied orthologous genes in the *P. serrata* draft genome and aligned them to the

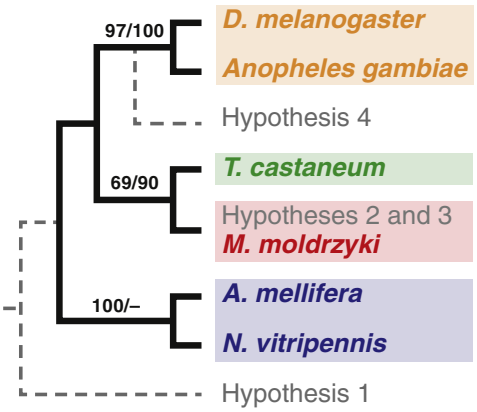

Figure 2. Phylogenetic Relationships of Strepsiptera to other Holometabolous Insects Inferred from Gene Order Distances

Numbers above branches are bootstrap support values from estimating distances with and without *Apis mellifera* (791/1,433 genes). Recent hypotheses on the phylogenetic origin of Strepsiptera are shown in gray. Abbreviations: *A*, *Apis*; *D*, *Drosophila*; *N*, *Nasonia*; *M*, *Mengenilla*; *T*, *Tribolium*.

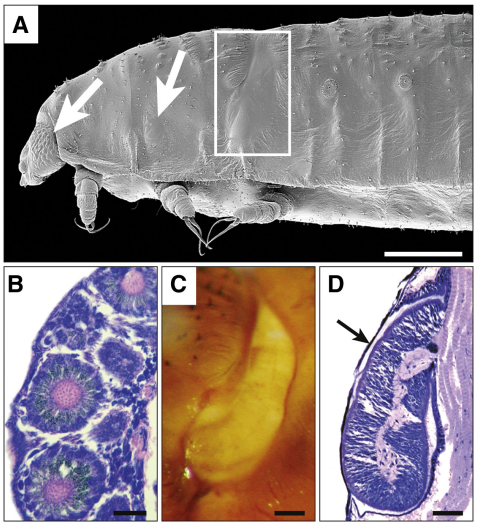

Figure 3. Peculiar Larval Developmental Features of Strepsiptera Reminiscent of Those of Hemimetabolous Insects

(A) SEM micrograph of late larval male of *Mengenilla chobauti*. Left arrow points to left compound eye; right arrow points to bud of left forewing on mesothorax; white rectangle defines sector with bud of left hindwing on metathorax; scale bar represents 500 μm.

(B) Cross-section through compound eye of late larval male of *Eoxenos laboulbenei* (Strepsiptera: Mengenillidae). Tissue was stained with basic fuchsin and methylene blue; scale bar represents 20 μm.

(C) Light microscopic image of left wing bud on metathorax of late larval male of *M. chobauti*; scale bar represents 100 μm.

(D) Cross-section through left wing bud on metathorax of late larval male of *E. laboulbenei*. Tissue stained with basic fuchsin and methylene blue; arrow points to the cuticula from a preceding larval stage; scale bar represents 100 μm.

sequences of the corresponding orthologs from the aforementioned insect species. The new matrix of RY-recoded second codon positions consisted of approximately 1.7 million characters (Table S9). We then repeated the sequence-based phylogenetic analysis, this time including the data from *P. serrata*. The inferred relationships of holometabolous insects were identical with the previously obtained ones, and all branches of the phylogenetic tree again received maximal statistical support (Figure 1B; Figure S2). The sequence data overwhelmingly support a sister group relationship between the archostematan beetle (*P. serrata*) and the polyphagan beetle (*T. castaneum*), indicating that Coleoptera represent a monophyletic group that does not include Strepsiptera. The analysis of NIPs provided additional and independent support for Strepsiptera being the sister group of beetles (Figure 1B; Figure S2).

The first sequenced genome of a twisted-wing parasite allowed the critical evaluation of current hypotheses on the phylogenetic origin of the enigmatic insect order Strepsiptera and provided strong support for Strepsiptera as the closest living relatives of beetles. Although our taxon sampling did not include Neuropterida (alderflies, dobsonflies, snakeflies, ant lions, and relatives), a close phylogenetic relationship between Neuropterida and Strepsiptera appears unlikely from a morphological point of view and would, among other unlikely events, require the independent evolution of posteromotorism, flight with the hindwings only, and a pupa with immobile mandibles (pupa adectica) in Coleoptera and Strepsiptera [2]. A sister group relationship of Strepsiptera and Coleoptera, which is in accordance with morphological evidence [2] and results of some molecular analyses [8, 10, 35], implies that the appearance of compound eyes and the presence of wing buds in late larval Strepsiptera are due to convergence instead of representing ancestral hemimetabolous developmental traits (Figure 1B). This shows that the sequence of holometabolous development, with late instar larvae exhibiting wing imaginal discs and only the pupal stage featuring visible wing buds, is not immutable. The striking similarity of the wing buds and complex eyes of the Strepsiptera late instar larvae (Figure 3) to those of hemimetabolous insect nymphs suggests the reuse of a pre-existing developmental program (homoiology), possibly triggered by a simple change of developmental timing (heterochrony). Our analyses demonstrate that the development of wing imaginal discs and the absence of compound eyes in larval stages are ground plan features of the extremely successful Holometabola and that Strepsiptera are consequently not the "missing link" between hemi- and holometabolous insects.

Experimental Procedures

Genome Sequencing and Assembly

The genome of *Mengenilla moldrzyki* was sequenced using a GS XLR 70 (Titanium) sequencer (Roche, Indianapolis, IN, USA) and tissue samples collected at the type locality (Tunisia, Parc Nationale du Jebil, N 32°58′40″/E 009°02′33″). Five PicoTiterPlates were dedicated to an unpaired shotgun (fragmented) library with genomic DNA from a single male. Two PicoTiterPlates were dedicated to a 3 kb mate-pair library with genomic DNA from 14 males. A normalized complementary DNA library from seven adult males was sequenced on an additional PicoTiterPlate. The genome and the complementary transcriptome data were assembled with Newbler 2.3 (Roche). The coverage of the sequenced *M. moldrzyki* genome was estimated with the *l*-mer approach implemented in the software GSP 1.06 (http://gsizepred.sourceforge.net). Genome sequences of *Priacma serrata* were obtained using an Illumina Hi-Seq 2000 sequencer (San Diego, CA, USA) to sequence two paired-end fragment libraries with 500 bp inserts using DNA from two adult males collected in Montana (Gallatin National Forest, N 45°35′27″/W 111°01′30″). The obtained sequence reads were assembled with CLCbio's Genomics Workbench 4.7.1 (Cambridge, MA, USA). Sequence data of the genome shotgun projects have been deposited in the Dryad data repository (http://datadryad.org/;doi:10.5061/dryad.ts058) and at DDBJ/EMBL/GenBank under the accession numbers AGDA00000000 and AGRH00000000.

Gene Annotation and Orthology

We used MAKER 2.02 with the ab initio gene prediction programs Augustus 2.4, GeneMark-ES 2.3a, and SNAP 2010-07-28 to infer models of NEPC genes [36–40]. We provided MAKER transcript sequences of *M. moldrzyki* and those of other Strepsiptera species downloaded from GenBank (release 179.0; October 5, 2010) and amino acid sequences downloaded from the UniprotKB and TrEMBL protein databases (October 5, 2010) as extrinsic evidence. Mitochondrial genes were annotated with MITOS (http://mitos.bioinf.uni-leipzig.de). Protein domains of NEPC genes were annotated with Pfam_scan.pl 1.3 and HMMER 3.0 and domains from the Pfam database version 24 [41, 42]. DNA methylation-related proteins were searched for and annotated with BLAST 2.2.24+ using amino acid sequences of corresponding proteins in *Apis mellifera* from RefSeq version 48 as query [43]. Noncoding RNAs were annotated with transfer RNA (tRNA)scan-SE 1.21 (tRNA genes and tRNA pseudogenes), RNAmmer 1.2 (18/28S and 5S ribosomal RNA), BLAST 2.2.8 (ncRNAs in general), rfam_scan.pl 1.0 and Infernal 1.02 (snoRNAs), and GotohScan 2.0 (microRNAs), using sequence data from the Rfam database 10.0, GenBank, and miRBase 16.0 [44–49]. Orthology of NEPC genes among species with annotated genome was assessed with OrthoMCL 2.0 [29]. Orthologous NEPC genes in the early draft genome of *Priacma serrata* were identified by reciprocal search using BLAST 2.2.20 and amino acid sequences of NEPC genes from *Tribolium castaneum* and *M. moldrzyki* as queries [46]. The annotated mitochondrial genome of *M. moldrzyki* has been deposited at DDBJ/EMBL/GenBank under the accession number JQ398619. All other annotations are available from the Dryad data repository (http://datadryad.org/;doi:10.5061/dryad.ts058).

Phylogenetic Analyses

Orthologous amino acid sequences were aligned with MAFFT 6.833b (L-INS-i algorithm), and the resulting alignments were refined with MUSCLE 3.7 [50, 51]. The amino acid alignments were used as blueprints to align the corresponding coding sequences using PAL2NAL 13 [52]. To improve the signal-to-noise ratio in the amino acid alignments of orthologous genes, we used ALISCORE 2.0 to identify and subsequently remove regions in the alignment, whose amino acid pattern-matches did not differ from a random pattern-match [53]. Substitutional saturation was assessed by calculating the observed distances between sequences and comparing them with corrected distances calculated with MEGA5 (Tamura-Nei substitution model) and using a guide tree inferred under the ML optimality criterion when analyzing the amino acid supermatrix of all 4,485 orthologous NEPC genes with RAxML 7.2.8-ALPHA (LG substitution matrix, empirically estimated amino acid frequencies [+F]; rate heterogeneity among sites modeled with gamma distribution [+Γ]) [54, 55]. The relative compositional variance (RCV) among sequences was calculated with the formula given by Phillips and Penny [56] and excluding constant sites. The signal-to-noise ratio in the data was assessed by calculating the proportion of internal branch lengths to all branch lengths using the minimum evolution (ME) optimality criterion and measuring the branch lengths in the above guide tree with PHYLIP 3.69 [57]. Partition schemes and substitution model parameters were evaluated with ModelGenerator 0.85 [58]. Matrices of RY-recoded second codon positions were analyzed with RAxML using the GTRGAMMA model and specifying 14 partitions, each uniting genes with a similar purine (R) frequency. The concatenated nucleotide sequence alignment of 13 noncoding RNAs (*bantam*, *mir-124*, *mir-133*, *mir-184*, *mir-190*, *mir-263*, *mir-275*, *mir-277*, *mir-305*, *mir-7*, *mir-9*, *U2*, and *U6atac*) was also analyzed with RAxML 7.2.8-ALPHA under the maximum likelihood (ML) optimality criterion, using a mixed RNA-DNA substitution model (S7D model and GTRGAMMA model for paired and unpaired nucleotides, respectively). Near intron pair (NIP) characters were analyzed under the MP optimality criterion using PAUP* 4.0b10 (heuristic tree search: random stepwise addition of taxa [1,000 replicates] and TBR branch-swapping) [59]. Gene order alignments were studied with the program TIBA using the double cut-and-join (DCJ) model for distance correction [31]. Statistical bootstrap support values were estimated from 1,000 (sequence-based and gene order analyses) and 10,000 (NIP character analysis) replicates. The primary sequence-based data matrices, the NIP character matrices, and the gene order alignments have been deposited in the Dryad data repository (http://datadryad.org/;doi:10.5061/dryad.ts058).

Accession Numbers

Sequence data of the genome shotgun projects have been deposited in the Dryad data repository (http://datadryad.org/;doi:10.5061/dryad.ts058) and at DDBJ/EMBL/GenBank under the accession numbers AGDA00000000 and AGRH00000000. The annotated mitochondrial genome of *M. moldrzyki* has been deposited at DDBJ/EMBL/GenBank under the accession number JQ398619. All other annotations are available from the Dryad data repository (http://datadryad.org/;doi:10.5061/dryad.ts058).

Supplemental Information

Supplemental Information includes four figures, nine tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.cub.2012.05.018.

Acknowledgments

We are grateful to Robert C. Edgar, Carson Holt, and Mario Stanke for help installing or using their software. We further acknowledge Berthold Fartmann, Jürgen Gadau, Gerald Nyakatura, Christopher D. Smith, and Xin-Xing Tan for technical advice and Thomas P. Niesel for IT support. We thank J. Wolfgang Wägele for helpful comments on an early draft of the manuscript. Michael Ivie kindly assisted with locating field sites for collecting *Priacma serrata*. Lars Podsiadlowski helped with submission of the mitochondrial data to GenBank. We are grateful for allocated computer time from the Cologne High Efficiency Operating Platform for Sciences (CHEOPS) at the University of Cologne. Sequencing and assembly of *P. serrata* was supported by a grant to D.D.M. from the University of Memphis FedEx Institute of Technology. O.N. acknowledges Judith Korb for providing her research facilities in an early phase of this project.

Received: April 3, 2012
Revised: May 4, 2012
Accepted: May 4, 2012
Published online: June 14, 2012

References

1. Kristensen, N.P. (1999). Phylogeny of endopterygote insects, the most successful lineage of living organisms. Eur. J. Entomol. *96*, 237–253.
2. Beutel, R.G., Friedrich, F., Hörnschemeyer, T., Pohl, H., Hünefeld, F., Beckman, F., Meier, R., Misof, B., Whiting, M.F., and Vilhemsen, L. (2011). Morphological and molecular evidence converge upon a robust phylogeny of the megadiverse Holometabola. Cladistics *27*, 341–355.
3. Trautwein, M.D., Wiegmann, B.M., Beutel, R., Kjer, K.M., and Yeates, D.K. (2012). Advances in insect phylogeny at the dawn of the postgenomic era. Annu. Rev. Entomol. *57*, 449–468.
4. Pierce, W.D. (1909). A monographic revision of the twisted winged insects comprising the order Strepsiptera Kirby. Bull. U. S. Nat. Mus. *66*, 1–232.
5. Whiting, M.F., and Wheeler, W.C. (1994). Insect homeotic transformation. Nature *368*, 696.
6. Carmean, D., and Crespi, B.J. (1995). Do long branches attract flies? Nature *373*, 666.
7. Proffitt, F. (2005). Parasitology. Twisted parasites from "outer space" perplex biologists. Science *307*, 343.
8. Wiegmann, B.M., Trautwein, M.D., Kim, J.-W., Cassel, B.K., Bertone, M.A., Winterton, S.L., and Yeates, D.K. (2009). Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. *7*, 34.
9. McKenna, D.D., and Farrell, B.D. (2010). 9-genes reinforce the phylogeny of holometabola and yield alternate views on the phylogenetic placement of Strepsiptera. PLoS ONE *5*, e11887.
10. Longhorn, S.J., Pohl, H.W., and Vogler, A.P. (2010). Ribosomal protein genes of holometabolan insects reject the Halteria, instead revealing a close affinity of Strepsiptera with Coleoptera. Mol. Phylogenet. Evol. *55*, 846–859.
11. McMahon, D.P., Hayward, A., and Kathirithamby, J. (2011). Strepsiptera. Curr. Biol. *21*, R272.
12. Truman, J.W., and Riddiford, L.M. (1999). The origins of insect metamorphosis. Nature *401*, 447–452.
13. Pohl, H., and Beutel, R.G. (2008). The evolution of Strepsiptera (Hexapoda). Zoology (Jena) *111*, 318–338.
14. Pohl, H., Niehuis, O., Gloyna, K., Misof, B., and Beutel, R.G. (2012). A new species of *Mengenilla* (Insecta, Strepsiptera) from Tunisia. ZooKeys. *198*, 79–102.
15. Kristensen, N.P. (1981). Phylogeny of insect orders. Annu. Rev. Entomol. *26*, 135–157.

16. Whiting, M.F., Carpenter, J.C., Wheeler, Q.D., and Wheeler, W.C. (1997). The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. Syst. Biol. *46*, 1–68.
17. Wheeler, W.C., Whiting, M., Wheeler, Q.D., and Carpenter, J.M. (2001). The phylogeny of the extant hexapod orders. Cladistics *17*, 113–169.
18. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of *Drosophila melanogaster*. Science *287*, 2185–2195.
19. Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., Zhao, P., Zha, X., Cheng, T., Chai, C., et al.; Biology Analysis Group. (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science *306*, 1937–1940.
20. Honeybee Genome Sequencing Consortium. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature *443*, 931–949.
21. Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., et al. (2007). Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science *316*, 1718–1723.
22. Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W., Gibbs, R., Beeman, R.W., Brown, S.J., Bucher, G., et al.; Tribolium Genome Sequencing Consortium. (2008). The genome of the model beetle and pest *Tribolium castaneum*. Nature *452*, 949–955.
23. Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F., et al. (2010). Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science *330*, 86–88.
24. Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., et al. (2010). Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science *329*, 1068–1071.
25. International Aphid Genomics Consortium. (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. *8*, e1000313.
26. Kirkness, E.F., Haas, B.J., Sun, W., Braig, H.R., Perotti, M.A., Clark, J.M., Lee, S.H., Robertson, H.M., Kennedy, R.C., Elhaik, E., et al. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc. Natl. Acad. Sci. USA *107*, 12168–12173.
27. Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Beukeboom, L.W., Desplan, C., Elsik, C.G., Grimmelikhuijzen, C.J., et al.; Nasonia Genome Working Group. (2010). Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science *327*, 343–348.
28. Smith, C.R., Smith, C.D., Robertson, H.M., Helmkampf, M., Zimin, A., Yandell, M., Holt, C., Hu, H., Abouheif, E., Benton, R., et al. (2011). Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc. Natl. Acad. Sci. USA *108*, 5667–5672.
29. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.
30. Krauss, V., Thümmler, C., Georgi, F., Lehmann, J., Stadler, P.F., and Eisenhardt, C. (2008). Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. Mol. Biol. Evol. *25*, 821–830.
31. Lin, Y., Rajan, V., and Moret, B.M.E. (2011). Bootstrapping phylogenies inferred from rearrangement data. In Proceedings of the 11th Workshop on Algorithms in Bioinformatics WABI'11, Lecture Notes in Computer Science, Vol. 6833, Przytycka, T.M., and Sagot, M.-F., eds. (Berlin, Heidelberg: Springer), pp. 175–187.
32. Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. Science *298*, 129–149.
33. Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O.S., Wild, R., Hammond, P.M., Ahrens, D., Balke, M., Caterino, M.S., et al. (2007). A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science *318*, 1913–1916.
34. Lawrence, J.F., Ślipiński, A., Seago, A.E., Thayer, M.K., Newton, A.F., and Marvaldi, A.E. (2011). Phylogeny of the Coleoptera based on morphological characters of adults and larvae. Ann. Zool. *61*, 1–217.
35. Ishiwata, K., Sasaki, G., Ogawa, J., Miyata, T., and Su, Z.-H. (2011). Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. Mol. Phylogenet. Evol. *58*, 169–180.
36. Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188–196.
37. Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19*, ii215–ii225.
38. Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. *26*, 1107–1115.
39. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. *33*, 6494–6506.
40. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.
41. Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.
42. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res. *40* (Database issue), D290–D301.
43. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.
44. Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.
45. Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100–3108.
46. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
47. Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R.D., Nawrocki, E.P., Kolbe, D.L., Eddy, S.R., and Bateman, A. (2011). Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res. *39*, D141–D145.
48. Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics *25*, 1335–1337.
49. Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B., and Stadler, P.F. (2009). Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. Nucleic Acids Res. *37*, 1602–1615.
50. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.
51. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.
52. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609–W612.
53. Misof, B., and Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol. *58*, 21–34.
54. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739.
55. Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.
56. Phillips, M.J., and Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. *28*, 171–185.
57. Felsenstein, J. (1989). PHYLIP – phylogeny inference package (version 3.2). Cladistics *5*, 164–166.
58. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., and Mclnerney, J.O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that *ad hoc* assumptions for choice of matrix are not justified. BMC Evol. Biol. *6*, 29.
59. Swofford, D.L. (2003). PAUP*. Phylogenetic Analysis using Parsimony (* and other methods) (Sunderland, MA: Sinauer).

C

Supplemental material to chapter 2

Supplemental Figures:

Supplemental Tables:

C.1 Supplemental Figures

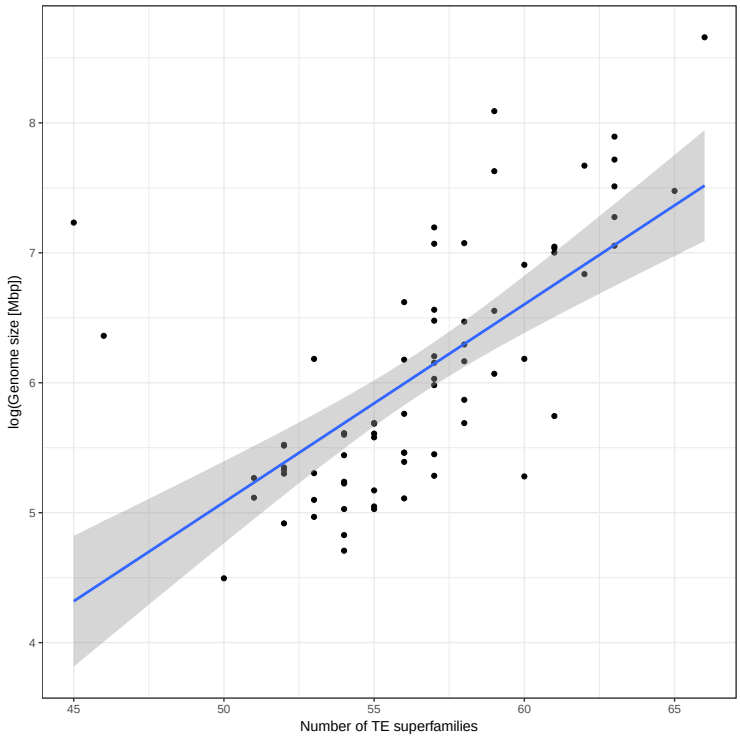

Figure C.1: The number of TE superfamilies is significantly correlated to genome size as well.

C.2 Supplemental Tables

Table C.1: Patterns employed to exclude non-TE search hits. Note that these are regular expressions for use with a compatible parser such as GNU grep or Perl.

Pattern
transcripta
transpos
gag[-/]pol
env(elope)? protein
env\b
pol p(olyp)?rotein
gag(-like)? protein
reverse transcrpitase
retro
integras
replicas
t-element
transporase
piggybac
copia

Table C.2: TE coverage by classes in 73 arthropod genomes.

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Acromyrmex echinatior	295944863	14028877	5564391	3332023	321166	56838168	80084625	27.0606572414132
Acyrthosiphon pisum	541675471	46850908	7139839	2352412	38433072	44085709	138861940	25.6356337760031
Aedes aegypti	1383971543	284584199	170617062	74484211	19268167	224707931	773661570	55.9015518717208
Agrilus planipennis	353849136	11965854	20658335	7721017	805978	45704697	86855881	24.5460203695396
Anopheles gambiae	265011681	14556836	8430174	6904028	2383954	13962246	46237238	17.4472452782185
Anoplophora glabripennis	707712193	80890038	14708125	3927741	65766	193744915	293336585	41.4485701816925
Apis mellifera	250270657	1477978	93806	661247	0	8415504	10648535	4.25480762612934
Athalia rosae	163837890	2189910	199632	415788	18659	4286428	7110417	4.33991001715171
Atta cephalotes	317690795	14790388	3104389	1917120	80553	53276653	73169103	23.0315464443973
Belgica antarctica	89583723	89347	225787	64091	0	1927767	2306992	2.57523568204461
Blattella germanica	2055425512	118099518	96796661	3368694	41260001	566565387	826090261	40.1907175023913
Bombus terrestris	248654244	3914003	2676528	1987158	9100	17941728	26528517	10.6688374078184

Table C.2 –continued

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Bombyx mori	481819406	14859808	52996846	2822023	45329577	66994932	183003186	37.9816968185794
Camponotus floridanus	232685334	3588249	1189283	1820352	245858	19473445	26317187	11.310204449757
Catajapyx silvestris	312272917	11564208	2999527	4563615	1450868	67274723	87852941	28.1333846828606
Centruroides exilicauda	931068862	60232232	19922674	2827127	40425	129049264	212071722	22.7772327757214
Ceratitis capitata	484773492	53984782	51057991	8093448	4481264	31292957	148910442	30.7175298273116
Cimex lectularius	650492763	26133805	77196299	8906041	14201408	70965264	197402817	30.346658445453
Copidosoma floridanum	645712421	24140234	15472247	28300530	1045178	77891713	146849902	22.7423071361361
Culex quinquefasciatus	579042118	148830919	19232476	12525560	10422015	82116713	273127683	47.1688802091595
Danaus plexippus	272853388	2556945	10476056	777496	1140231	13930574	28881302	10.5849160282371
Daphnia pulex	197206209	3913500	1756846	11804780	1769569	20903178	40147873	20.3583209694985
Drosophila ananassae	230993012	5851179	18214713	36010716	5878	32843785	92926271	40.2290399157183
Drosophila erecta	152712140	1620381	7925820	12413050	5791	6927427	28892469	18.9195626490468

Table C.2 –continued

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Drosophila grimshawi	200467819	4081630	5726513	22750394	1989	7938263	40498789	20.202139775861
Drosophila melanogaster	143726002	1864630	6201553	14975264	0	4411181	27452628	19.1006690633474
Drosophila miranda	136728780	1193299	2169497	932607	12128	5964753	10272284	7.5128908485836
Drosophila mojavensis	193826310	4423019	6200643	12547097	0	15174518	38345277	19.7833188899897
Drosophila persimilis	188374079	3017923	10737250	21690609	44193	17715388	53205363	28.244524555844
Drosophila pseudoobscura	152696384	1814141	4620512	9081564	6593	7604926	23127736	15.1462237638843
Drosophila sechellia	166592095	4545125	10981352	16246960	0	5975207	37748644	22.6593248617229
Drosophila simulans	124966452	454842	3082710	4363044	10511	1017167	8928274	7.14453667933215
Drosophila virilis	206026697	2388960	7217526	14406968	2878	21536844	45553176	22.1103267990556
Drosophila willistoni	235516348	6979192	13395864	30867252	12943	23060371	74315622	31.5543369413999
Drosophila yakuba	165693946	2762860	7240428	18655858	4743	8220312	36884201	22.2604397386975
Ephemera danica	475911277	1587870	4342127	657361	5788	103286144	109879290	23.0881879270955

Table C.2 –continued

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Euperipatoides rowelli	2681872052	81520224	68827726	43632247	1743681	591339224	787063102	29.3475261585671
Eurytemora affinis	494890867	4913699	2873823	2975606	0	118249976	129013104	26.0690007843689
Frankliniella occidentalis	415803855	2366622	800197	769380	500775	36574762	41011736	9.86324092642191
Gerris buenoi	1000194699	23294121	11729071	3527948	5436752	200880060	244867952	24.4820285735188
Halyomorpha halys	1150099797	15071472	137983767	10173492	12033787	277888335	453150853	39.4010027809787
Harpegnathos saltator	294465601	22402907	4860671	2536115	401040	38188633	68389366	23.2249083654427
Heliconius melpomene	273786188	3309141	10787521	1806703	5675476	62279040	83857881	30.6289669367835
Helicoverpa punctigera	432318525	1626618	8168605	684817	14403617	49299046	74182703	17.1592700081497
Homalodisca vitripennis	2247672265	33258334	50290153	1400275	5082654	223402079	313433495	13.9448041371814
Hyalella azteca	1181648033	17877377	22314001	1745793	6144	92142233	134085548	11.3473339146158
Ixodes scapularis	1765382190	50926831	55077295	19711825	11668591	568841548	706226090	40.0041472039547
Ladona fulva	1158111285	35625466	48467767	1277019	4772032	127344552	217486836	18.7794419082964

Table C.2 –continued

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Latrodectus hesperus	1137104758	64136256	2775971l	8159512	8420167	79950721	188426367	16.570713091678
Leptinotarsa decemlineata	1176182208	82875924	133162799	8299074	1354549	138863056	364555402	30.99480671620 57
Limnephilus lunatus	1333324643	27934987	62350860	529611	28296255	293532450	412644163	30.9485139396767
Limulus polyphemus	1828256766	57602361	65031258	72495679	42695658	370231034	608055990	33.2587851612523
Linepithema humile	219500750	3146879	1635451	1634508	124443	16888693	23429974	10.67421136374 25
Locusta migratoria	5759798599	548656473	922471727	110839523	122578589	1955778475	3660324787	63.549527367771l
Loxosceles reclusa	3262503565	277060302	237214458	41355923	45464181	479023789	1080118653	33.1070489726806
Lucilia cuprina	470583961	6587453	19759307	7925151	2180	85424374	119698465	25.4361548459149
Machilis hrabei	2144866089	88401084	56089285	7823768	31164907	429393676	612872720	28.5739386315599
Mayetiola destructor	185827756	2334815	634675	1332276	23062	14823623	19148451	10.3044084544615
Mengenilla moldrzyki	155727465	11658097	2671635	5169197	18013	55690413	75207355	48.2942138690821
Musca domestica	750403944	153293508	14279601	10470028	131797	218138285	396313219	52.8133177029251

Table C.2 –continued

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Nasonia vitripennis	295780872	9357282	9442807	12159327	93355	25283154	56335925	19.0465071723773
Oncopeltus fasciatus	1098693218	24473805	59736772	6583561	17420216	123890779	232105133	21.1255634600632
Onthophagus taurus	270546467	20788548	19283682	2900456	35890	51627557	94636133	34.9796225577767
Orussus abietinus	201220334	2757902	424153	1413563	38403	34929859	39563880	19.6619691526802
Pachypsylla venusta	701795784	18682836	13497373	754417	9711493	129459210	172105329	24.5235626835855
Parasteatoda tepidariorum	1443909906	71202706	13909498	2196300	34445396	332296216	454050116	31.4458758204544
Pediculus humanus	110781312	2419628	1040406	939937	226962	2807314	7434247	6.71074106795197
Pogonomyrmex barbatus	235645958	6927372	1525631	3764006	113867	18340243	30671119	13.0157628250089
Solenopsis invicta	396009169	15467507	7574738	8638636	0	82023961	113704842	28.7126791248614
Strigamia maritima	176210797	2555637	1485499	20620465	342139	48848531	73852271	41.9113199970374
Tribolium castaneum	210248733	8676449	1786295	624856	32802	32662849	43783251	20.8245017105525
Trichogramma pretiosum	196221301	1988709	1912973	1512183	72212	19378013	24864090	12.671453034551

Table C.2 –continued

Species	Genome size	DNA	LINE	LTR	SINE	Unknown	Total	Coverage [%]
Zootermopsis nevadensis	485009472	14695064	26646385	236056	9305656	70248697	121131858	24.9751530625777

Table C.3: Download URLs for the genome assemblies of 73 arthropod species.

Species	Order	URL
Aedes aegypti	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000004015.1_Aedes_aegypti/GCA_000004015.1_Aedes_aegypti_genomic.fna.gz
Atta cephalotes	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000143395.2_Attacep1.0/GCA_000143395.2_Attacep1.0_genomic.fna.gz
Acromyrmex echinatior	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000204515.1_Aech_3.9/GCA_000204515.1_Aech_3.9_genomic.fna.gz
Anopheles gambiae	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000005575.1_AgamP3/GCA_000005575.1_AgamP3_genomic.fna.gz
Apis mellifera	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000002195.1_Amel_4.5/GCA_000002195.1_Amel_4.5_genomic.fna.gz
Acyrthosiphon pisum	Hemiptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000142985.2_Acyr_2.0/GCA_000142985.2_Acyr_2.0_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Belgica antarctica	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000775305.1_ASM77530v1/GCA_000775305.1_ASM77530v1_genomic.fna.gz
Bombyx mori	Lepidoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000151625.1_ASM15162v1/GCF_000151625.1_ASM15162v1_genomic.fna.gz
Bombus terrestris	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000214255.1_Bter_1.0/GCA_000214255.1_Bter_1.0_genomic.fna.gz
Camponotus floridanus	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000147175.1_CamFlo_1.0/GCA_000147175.1_CamFlo_1.0_genomic.fna.gz
Culex quinquefasciatus	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000209185.1_CulPip1.0/GCA_000209185.1_CulPip1.0_genomic.fna.gz
Drosophila ananassae	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005115.1_dana_caf1/GCF_000005115.1_dana_caf1_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Drosophila erecta	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005135.1_dere_caf1/GCF_000005135.1_dere_caf1_genomic.fna.gz
Drosophila grimshawi	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005155.2_dgri_caf1/GCF_000005155.2_dgri_caf1_genomic.fna.gz
Drosophila melanogaster	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000001215.4_Release_6_plus_ISO1_MT/GCA_000001215.4_Release_6_plus_ISO1_MT_genomic.fna.gz
Drosophila miranda	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000269505.2_DroMir_2.2/GCA_000269505.2_DroMir_2.2_genomic.fna.gz
Drosophila mojavensis	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005175.2_dmoj_caf1/GCF_000005175.2_dmoj_caf1_genomic.fna.gz
Drosophila persimilis	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005195.2_dper_caf1/GCF_000005195.2_dper_caf1_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Danaus plexippus	Lepidoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000235995.1_DanPle_1.0/GCA_000235995.1_DanPle_1.0_genomic.fna.gz
Drosophila pseudoobscura	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000001765.3_Dpse_3.0/GCF_000001765.3_Dpse_3.0_genomic.fna.gz
Daphnia pulex	Cladocera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000187875.1_V1.0/GCA_000187875.1_V1.0_genomic.fna.gz
Drosophila sechellia	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005215.3_dsec_caf1/GCF_000005215.3_dsec_caf1_genomic.fna.gz
Drosophila simulans	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000754195.2_ASM75419v2/GCA_000754195.2_ASM75419v2_genomic.fna.gz
Drosophila virilis	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005245.1_dvir_caf1/GCF_000005245.1_dvir_caf1_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Drosophila willistoni	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000005925.1_dwil_caf1/GCF_000005925.1_dwil_caf1_genomic.fna.gz
Drosophila yakuba	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000005975.1_dyak_caf1/GCA_000005975.1_dyak_caf1_genomic.fna.gz
Heliconius melpomene	Lepidoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000313835.2_ASM31383v2/GCA_000313835.2_ASM31383v2_genomic.fna.gz
Harpegnathos saltator	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000147195.1_HarSal_1.0/GCA_000147195.1_HarSal_1.0_genomic.fna.gz
Ixodes scapularis	Ixodida	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000208615.1_JCVI_ISG_i3_1.0/GCA_000208615.1_JCVI_ISG_i3_1.0_genomic.fna.gz
Linepithema humile	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000217595.1_Lhum_UMD_V04/GCA_000217595.1_Lhum_UMD_V04_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Locusta migratoria	Orthoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000516895.1_LocustGenomeV1/GCA_000516895.1_LocustGenomeV1_genomic.fna.gz
Limulus polyphemus	Xiphosura	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000517525.1_Limulus_polyphemus-2.1.2/GCA_000517525.1_Limulus_polyphemus-2.1.2_genomic.fna.gz
Mayetiola destructor	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000149185.1_Mdes_1.0/GCA_000149185.1_Mdes_1.0_genomic.fna.gz
Musca domestica	Diptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000371365.1_Musca_domestica-2.0.2/GCF_000371365.1_Musca_domestica-2.0.2_genomic.fna.gz
Mengenilla moldrzyki	Strepsiptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000281935.1_Memo_1.0/GCA_000281935.1_Memo_1.0_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Nasonia vitripennis	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000002325.2_Nvit_2.1/GCA_000002325.2_Nvit_2.1_genomic.fna.gz
Pogonomyrmex barbatus	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000187915.1_Pbar_UMD_V03/GCA_000187915.1_Pbar_UMD_V03_genomic.fna.gz
Pediculus humanus	Phthiraptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000006295.1_JCVI_LOUSE_1.0/GCA_000006295.1_JCVI_LOUSE_1.0_genomic.fna.gz
Solenopsis invicta	Hymenoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000188075.1_Si_gnG/GCA_000188075.1_Si_gnG_genomic.fna.gz
Strigamia maritima	Myriapoda	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000239455.1_Smar_1.0/GCA_000239455.1_Smar_1.0_genomic.fna.gz
Tribolium castaneum	Coleoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000002335.2_Tcas_3.0/GCA_000002335.2_Tcas_3.0_genomic.fna.gz

Table C.3 –continued

Species	Order	URL
Zootermopsis nevadensis	Isoptera	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000696155.1_ZooNev1.0/GCA_000696155.1_ZooNev1.0_genomic.fna.gz
Agrilus planipennis Fairmaire	Coleoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Emerald_ash_borer/NCBI-submitted/Aplan.agp.contamination-free.scaffolds.50.fa
Anoplophora glabripennis	Coleoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Asian_long-horned_beetle/Agla_Btl03082013.genome.fa
Athalia rosae	Hymenoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Turnip_sawfly/Aros01112013-genome.fa
Blattella germanica	Blattodea	ftp://ftp.hgsc.bcm.edu/I5K-pilot/German_cockroach/Bgermanica.scaffolds
Catajapyx silvestris	Diplura	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Silvestris_Northern_Forcepstail/forcepstail.consistent.scaffolds

Table C.3 –continued

Species	Order	URL
Centruroides exilicauda	Scorpiones	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Bark_scorpion/NCBI-submitted/Cscul.scaffolds.50.fa
Ceratitis capitata	Diptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Mediterranean_fruit_fly/Ccap01172013-genome.fa
Cimex lectularius	Hemiptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Bed_bug/Clec_Bbug02212013.genome.fa
Copidosoma floridanum	Hymenoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Copidosoma_floridanum/NCBI-submitted/Cflo.scaffolds.50.fa
Ephemera danica	Ephemeroptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Mayfly/Edan07162013.scaffolds.fa
Euperipatoides rowelli	Euonychophora	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Velvet_worm/pre_assembly/Erow.scaffolds.fasta
Eurytemora affinis	Calanoida	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Eurytemora_affinis/NCBI-submitted/Eaff_11172013.genome.fa

Table C.3 –continued

Species	Order	URL
Frankliniella occidentalis	Thysanoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Western_flower_thrips/NCBI-submitted/Focc.scaffolds
Gerris buenoi	Hemiptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Water_strider/NCBI-submitted/Gbue_1.0-unplaced_scaffolds.fsa
Halyomorpha halys	Hemiptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Brown_marmorated_stink_bug/Hhal.scaffolds.fa
Helicoverpa punctigera	Lepidoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Helicoverpa_punctigera/Hpun12202012.genome.fa
Homalodisca vitripennis	Hemiptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Glassy-winged_sharpshooter/NCBI-submitted/Hvit.scaffolds
Hyalella azteca	Amphipoda	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Hyalella_azteca/pre_assembly/Hazt.scaffolds.fasta

Table C.3 –continued

Species	Order	URL
Ladona fulva	Odonata	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Scarce_Chaser/Lful_Scha04012013-genome.fa
Latrodectus hesperus	Araneae	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Western_black_widow_spider/NCBI-submitted/Lhes.scaffolds
Leptinotarsa decemlineata	Coleoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Colorado_Potato_Beetle/Ldec.genome.10062013.fa
Limnephilus lunatus	Trichoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Caddisfly/NCBI-submitted/Llun.contaminationfree.scaffolds.fa
Loxosceles reclusa	Araneae	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Brown_recluse_spider/NCBI-submitted/Lrec.scaffolds
Lucilia cuprina	Diptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Sheep_blowfly/NCBIsubmitted/Lcup.scaffolds

Table C.3 –continued

Species	Order	URL
Machilis hrabei	Archaeognatha	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Hrabes_jumping_bristletail/pre-assembly/Mhar.scaffolds.fasta
Oncopeltus fasciatus	Hemiptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Milkweed_bug/NCBI-submitted/Ofas.contaminationfree.scaffolds
Onthophagus taurus	Coleoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Bull-headed_Dung_beetle/Otaur.scaffolds.fa
Orussus abietinus	Hymenoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Parasitic_wood_wasp/Oabi112420l3.genome.fa
Pachypsylla venusta	Hemiptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Hackberry_petiole_gall_psyllid/NCBI-submitted/Pven.scaffolds.50.fa
Parasteatoda tepidariorum	Araneae	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Common_house_spider/NCBI-submitted/Ptepo1282013.genome.fa

Table C.3 –continued

Species	Order	URL
Trichogramma pretiosum	Hymenoptera	ftp://ftp.hgsc.bcm.edu/I5K-pilot/Trichogramma_pretiosum/Tpre_scaffolds.50.fa

D

Supplemental material to chapter 3

Supplemental Figures:

Supplemental Tables:

D.1 Supplemental figures

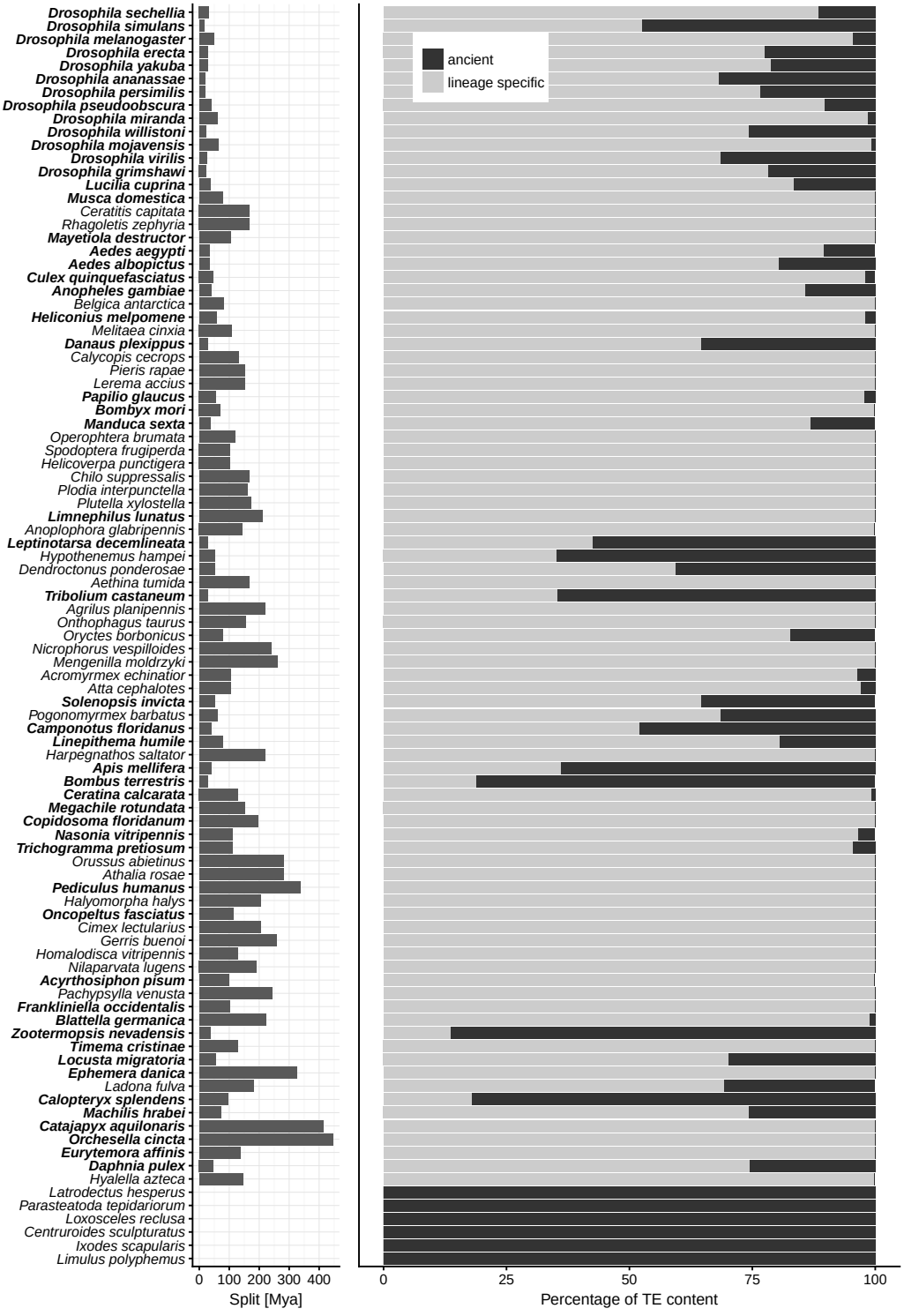

Figure D.1: Most insect transposable elements are clade-specific when analyzed at order level. TE age was determined from the RepeatMasker (Smit et al., 2015) annotation using the intra-TE-family Kimura distances and order-specific nucleotide substitution rates based on data from Misof et al. (2014). TE copies were classified as "ancient" if they were older (more divergent) than the clade the host belongs to. Bold font face denotes species for which ancestral genome size inferences and branch length estimates are available.

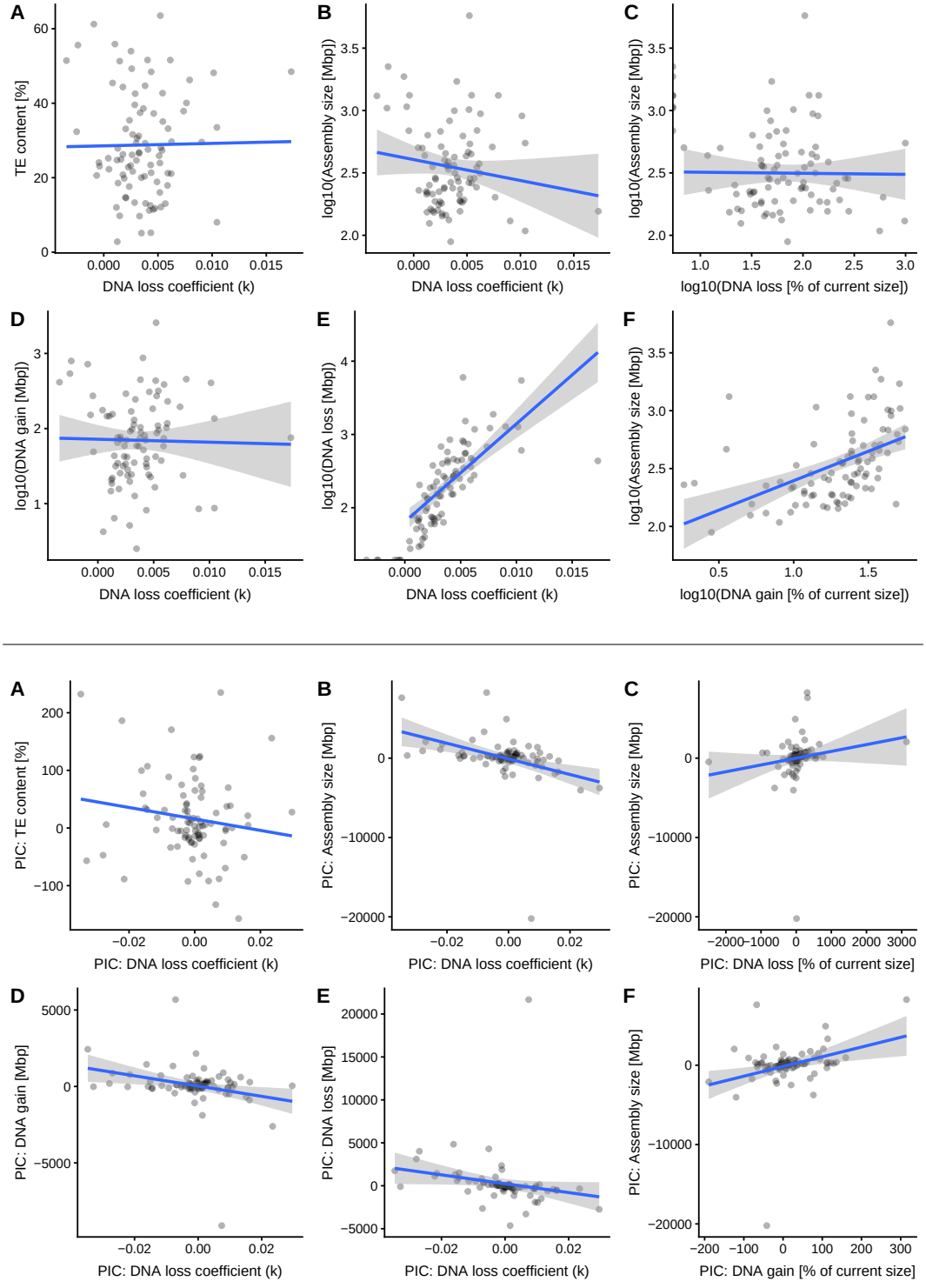

Figure D.2: Insect genome size dynamics and TE content are governed by the DNA loss coefficient. Top: without phylogenetic independent contrasts (PIC), bottom: with PIC. A: While TE content determines the genome size (Figure D.3), the TE content is not dependent on the DNA loss coefficient k. There is no correlation despite a visible trend in the regression (Pearson, $p = 0.15$). Obviously, genome (assembly) size decreases with higher k (B, $p = 0.0002$), as does the amount of DNA gained (D, $p = 0.01$). Surprisingly, the assembly size appears to remain more or less stable despite increasing amounts of DNA loss (C). A strong negative correlation is, however, found by testing for it (section D.3.2; $p \ll 0.005$).

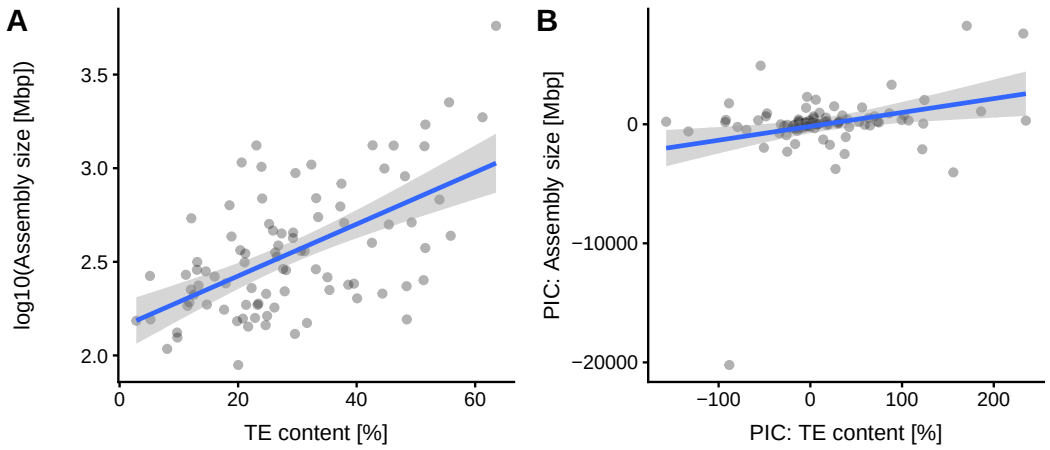

Figure D.3: TE content is a predictor for genome size. Dots: individual measurements; blue line: linear regression; shaded area: confidence interval. PIC: phylogenetic independent contrast (Felsenstein, 1985)

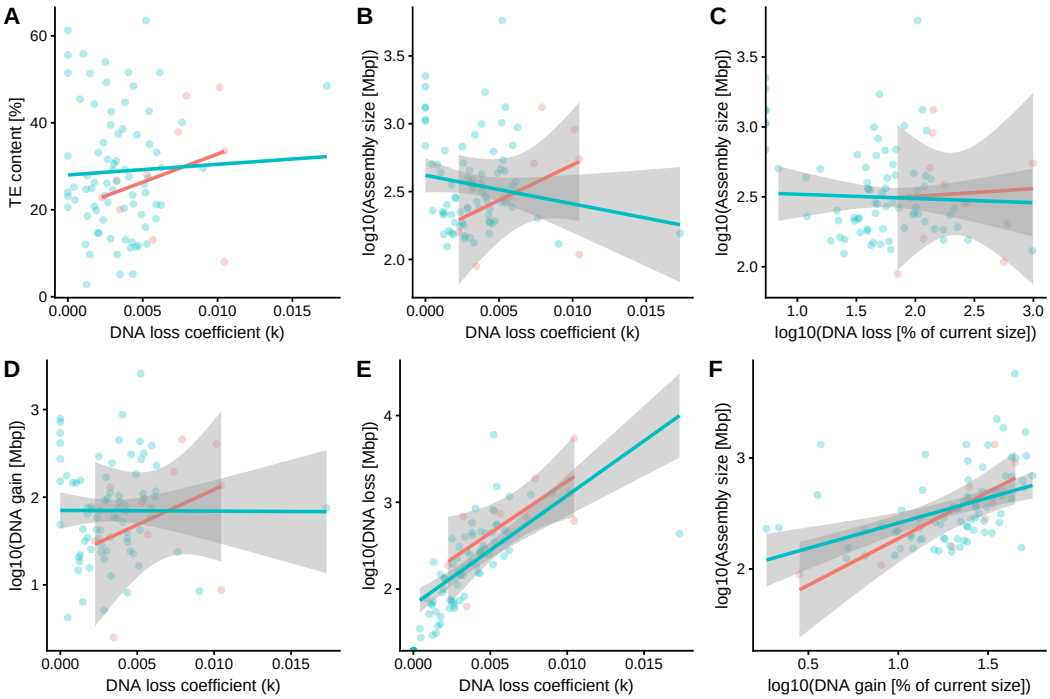

Figure D.4: The same as Fig. D.3. Red: flightless; blue: flying

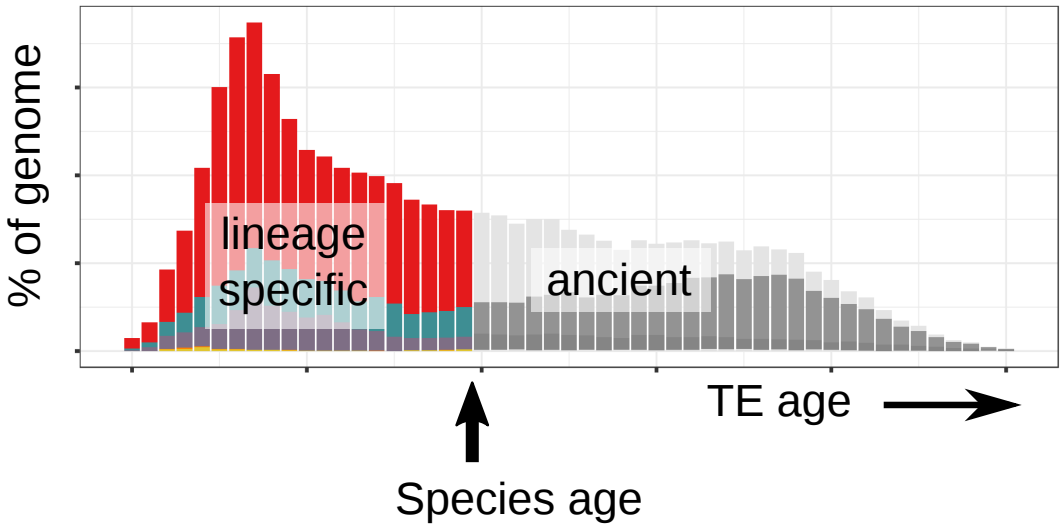

Figure D.5: The age classification analysis splits the repeat landscape into the ancestral and the lineage-specific parts. The further to the right the species' age is, the greater the lineage-specific fraction of the TE content. If the species is older than the oldest TE copy on the landscape, it will have 0 % ancestral TEs.

D.2 Data sources

D.2.1 Genome assemblies

Table D.1: NCBI accession numbers and references for the genome assemblies.

Species	Order	NCBI Accession	Reference
Drosophila yakuba	Diptera	GCA_000005975.1	Drosophila 12 Genomes Consortium (2007)
Drosophila simulans	Diptera	GCA_000754195.2	Drosophila 12 Genomes Consortium (2007)
Drosophila sechellia	Diptera	GCF_000005215.3	Drosophila 12 Genomes Consortium (2007)
Drosophila melanogaster	Diptera	GCA_000001215.4	Adams (2000)
Drosophila erecta	Diptera	GCF_000005135.1	Drosophila 12 Genomes Consortium (2007)
Drosophila ananassae	Diptera	GCF_000005115.1	Drosophila 12 Genomes Consortium (2007)
Drosophila pseudoobscura	Diptera	GCF_000001765.3	Drosophila 12 Genomes Consortium (2007)
Drosophila persimilis	Diptera	GCF_000005195.2	Drosophila 12 Genomes Consortium (2007)

Table D.1 –continued

Species	Order	NCBI Accession	Reference
Drosophila miranda	Diptera	GCA_000269505.2	McGaugh & Noor (2012)
Drosophila willistoni	Diptera	GCF_000005925.1	Drosophila 12 Genomes Consortium (2007)
Drosophila virilis	Diptera	GCF_000005245.1	Drosophila 12 Genomes Consortium (2007)
Drosophila mojavensis	Diptera	GCF_000005175.2	Drosophila 12 Genomes Consortium (2007)
Drosophila grimshawi	Diptera	GCF_000005155.2	Drosophila 12 Genomes Consortium (2007)
Rhagoletis zephyria	Diptera	GCA_001687245.1	Drosophila 12 Genomes Consortium (2007)
Ceratitis capitata	Diptera	GCA_000347755.2	Papanicolaou et al. (2016)
Lucilia cuprina	Diptera	GCA_000699065.1	i5k Initiative
Musca domestica	Diptera	GCF_000371365.1	Scott et al. (2014)
Culex quinquefasciatus	Diptera	GCA_000209185.1	Arensburger et al. (2010)
Aedes albopictus	Diptera	GCA_001444175.2	Chen et al. (2015)
Aedes aegypti	Diptera	GCA_000004015.1	Nene et al. (2007)
Anopheles gambiae	Diptera	GCA_000005575.1	Holt et al. (2002)
Belgica antarctica	Diptera	GCA_000775305.1	Kelley et al. (2014)

Table D.1 –continued

Species	Order	NCBI Accession	Reference
Mayetiola destructor	Diptera	GCA_000149185.1	Zhao et al. (2015)
Papilio glaucus	Lepidoptera	GCA_000931545.1	Cong et al. (2015a)
Melitaea cinxia	Lepidoptera	GCA_000716385.1	Ahola et al. (2014)
Heliconius melpomene	Lepidoptera	GCA_000313835.2	The Heliconius Genome Consortium et al. (2012)
Danaus plexippus	Lepidoptera	GCA_000235995.1	Zhan et al. (2011)
Calycopis cecrops	Lepidoptera	GCA_001625245.1	Cong et al. (2016)
Pieris rapae	Lepidoptera	GCA_001856805.1	Shen et al. (2016)
Lerema accius	Lepidoptera	GCA_001278395.1	Cong et al. (2015b)
Manduca sexta	Lepidoptera	GCA_000262585.1	Kanost et al. (2016)
Plutella xylostella	Lepidoptera	GCA_000325945.1	You et al. (2013)
Spodoptera frugiperda	Lepidoptera	GCA_000753635.2	Gouin et al. (2017)
Helicoverpa punctigera	Lepidoptera		i5k Initiative
Chilo suppressalis	Lepidoptera	GCA_000636095.1	Yin et al. (2014)
Operophtera brumata	Lepidoptera	GCA_001266575.1	Derks et al. (2015)
Plodia interpunctella	Lepidoptera	GCA_900182495.1	Paterson (2017)
Bombyx mori	Lepidoptera	GCF_000151625.1	International Silkworm Genome Consortium (2008)
Limnephilus lunatus	Trichoptera	GCA_000648945.1	i5k Initiative

Table D.1 –continued

Species	Order	NCBI Accession	Reference
Aethina tumida	Coleoptera	GCF_001937115.1	Evans et al. (2018)
Anoplophora glabripennis	Coleoptera	GCA_000390285.1	i5k Initiative
Leptinotarsa decemlineata	Coleoptera	GCA_000500325.1	i5k Initiative
Tribolium castaneum	Coleoptera	GCA_000002335.2	Tribolium Genome Sequencing Consortium (2008)
Agrilus planipennis	Coleoptera	GCA_000699045.1	i5k Initiative
Oryctes borbonicus	Coleoptera	GCA_001443705.1	Meyer et al. (2016)
Onthophagus taurus	Coleoptera	GCA_000648695.1	i5k Initiative
Dendroctonus ponderosae	Coleoptera	GCF_000355655.1	Keeling et al. (2013)
Hypothenemus hampei	Coleoptera	GCA_001012855.1	Vega et al. (2015)
Nicrophorus vespilloides	Coleoptera	GCF_001412225.1	Cunningham et al. (2015)
Mengenilla moldrzyki	Strepsiptera	GCA_000281935.1	Niehuis et al. (2012)
Pogonomyrmex barbatus	Hymenoptera	GCA_000187915.1	Smith et al. (2011b)
Solenopsis invicta	Hymenoptera	GCA_000188075.1	Wurm et al. (2011)
Acromyrmex echinatior	Hymenoptera	GCA_000204515.1	Nygaard et al. (2011)
Atta cephalotes	Hymenoptera	GCA_000143395.2	Suen et al. (2011)
Harpegnathos saltator	Hymenoptera	GCA_000147195.1	Bonasio et al. (2010)
Camponotus floridanus	Hymenoptera	GCA_000147175.1	Bonasio et al. (2010)
Linepithema humile	Hymenoptera	GCA_000217595.1	Smith et al. (2011a)

Table D.1 –continued

Species	Order	NCBI Accession	Reference
Megachile rotundata	Hymenoptera	GCF_000220905.1	Robinson et al. (2014)
Ceratina calcarata	Hymenoptera	GCF_001652005.1	Rehan et al. (2016)
Bombus terrestris	Hymenoptera	GCA_000214255.1	Sadd et al. (2015)
Apis mellifera	Hymenoptera	GCA_000002195.1	Honeybee Genome Sequencing Consortium (2006)
Nasonia vitripennis	Hymenoptera	GCA_000002325.2	Werren et al. (2010)
Copidosoma floridanum	Hymenoptera	GCA_000648655.1	i5k Initiative
Trichogramma pretiosum	Hymenoptera	GCA_000599845.2	i5k Initiative
Orussus abietinus	Hymenoptera	GCA_000612105.1	i5k Initiative
Athalia rosae	Hymenoptera	GCA_000344095.1	i5k Initiative
Pediculus humanus	Psocodea	GCA_000006295.1	Kirkness et al. (2010)
Halyomorpha halys	Heteroptera	GCA_000696795.1	i5k Initiative
Oncopeltus fasciatus	Heteroptera	GCA_000696205.1	i5k Initiative
Cimex lectularius	Heteroptera	GCA_000648675.1	Rosenfeld et al. (2016)
Gerris buenoi	Heteroptera	GCA_001010745.1	i5k Initiative
Nilaparvata lugens	Auchenorrhyncha	GCA_000757685.1	Xue et al. (2014)
Homalodisca vitripennis	Auchenorrhyncha	GCA_000696855.1	i5k Initiative
Pachypsylla venusta	Sternorrhyncha	GCA_000695645.1	i5k Initiative

Table D.1 –continued

Species	Order	NCBI Accession	Reference
Acyrthosiphon pisum	Sternorrhyncha	GCA_000142985.2	The International Aphid Genomics Consortium (2010)
Frankliniella occidentalis	Thysanoptera	GCA_000697945.1	i5k Initiative
Blattella germanica	Blattodea	GCA_000762945.1	i5k Initiative
Zootermopsis nevadensis	Isoptera	GCA_000696155.1	Terrapon et al. (2014)
Timema cristinae	Phasmatodea	GCA_002009905.3	i5k Initiative
Locusta migratoria	Orthoptera	GCA_000516895.1	Wang et al. (2014)
Ephemera danica	Ephemeroptera	GCA_000507165.1	i5k Initiative
Calopteryx splendens	Odonata	GCA_002093875.1	i5k Initiative
Ladona fulva	Odonata	GCA_000376725.1	i5k Initiative
Machilis hrabei	Archaeognatha		i5k Initiative
Catajapyx aquilonaris	Diplura	GCA_000934665.1	i5k Initiative
Orchesella cincta	Collembola	GCA_001718145.1	Faddeeva-Vakhrusheva et al. (2016)
Hyalella azteca	Copepoda		i5k Initiative
Eurytemora affinis	Branchiopoda	GCA_000591075.1	Eyun et al. (2017)
Daphnia pulex	Malacostraca	GCA_000187875.1	Colbourne et al. (2011)
Strigamia maritima	Myriapoda	GCA_000239455.1	Chipman et al. (2014)
Latrodectus hesperus	Araneae	GCA_000697925.1	i5k Initiative

Table D.1 –continued

Species	Order	NCBI Accession	Reference
Parasteatoda tepidariorum	Araneae	GCA_000365465.2	Schwager et al. (2017)
Loxosceles reclusa	Araneae	GCA_001188405.1	i5k Initiative
Centruroides sculpturatus	Scorpionidae	GCA_000671375.1	Schwager et al. (2017)
Ixodes scapularis	Ixodida	GCA_000208615.1	Gulia-Nuss et al. (2016)
Limulus polyphemus	Xiphosura	GCA_000517525.1	Simpson et al. (2017)

D.2.2 Insect genome size estimates

We estimated genome sizes of eight additional species using either flow cytometry (FCM) or a *k*-mer peak method adopted from (Hozza et al., 2015). For *k*-mer estimates, we downloaded genomic reads from i5k FTP server for *Limnephilus lunatus* and *Catajapyx aquilonaris*:

Catajapyx aquilonaris: `ftp://ftp.hgsc.bcm.edu/I5K-pilot/Silvestris_Northern_Forcepstail/genomic_sequence/Caqu_1Kb_1_sequence.txt.bz2`

Limnephilus lunatus: `ftp://ftp.hgsc.bcm.edu/I5K-pilot/Caddisfly/genomic_sequence/Llun_8kb_1_sequence.txt.bz2`

For *Stylops ovinae*, we used our own (unpublished) genomic short reads.

Using flow cytometry, we estimated the genome size for an additional 9 species. The results are listed in table D.2 on page 368.

Table D.2: Genome size estimates. The *c*-value (in picogram DNA per haploid cell) is converted to a value in Mbp by calculating $c \times 978$ (Doležel et al., 2003). Note that *Stylops ater* is a synonym for *S. ovinae*. FCM: flow cytometry.

Order	Family	Species	c-value	Mbp	Method
Diplura	Japygidae	*Catajapyx aquilonaris*	0.316	308.855	25-mer
Thysanura	Lepismatidae	*Thermobia domestica*	3.982	3894.055	FCM
Thysanura	Lepismatidae	*Thermobia domestica*	3.837	3752.976	FCM
Thysanura	Lepismatidae	*Thermobia domestica*	2.000	1956.000	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.413	403.741	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.427	417.919	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.454	444.068	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.433	423.355	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.480	469.614	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.462	451.856	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.411	401.903	FCM
Ephemeroptera	Ephemeridae	*Ephemera danica*	0.400	391.626	FCM
Mecoptera	Panorpidae	*Panorpa germanica*	0.591	578.002	FCM
Mecoptera	Panorpidae	*Panorpa germanica*	0.598	584.661	FCM
Megaloptera	Sialidae	*Sialis lutaria*	0.366	358.383	FCM
Megaloptera	Sialidae	*Sialis lutaria*	0.407	397.908	FCM
Megaloptera	Sialidae	*Sialis lutaria*	0.393	384.185	FCM
Megaloptera	Sialidae	*Sialis lutaria*	0.403	393.989	FCM
Neuroptera	Chrysopidae	*Chrysopa perla*	0.699	683.460	FCM
Neuroptera	Chrysopidae	*Chrysopa perla*	0.694	678.502	FCM
Neuroptera	Chrysopidae	*Chrysopa perla*	0.708	692.763	FCM
Neuroptera	Chrysopidae	*Chrysopa perla*	0.677	661.918	FCM
Neuroptera	Chrysopidae	*Chrysopa perla*	0.709	693.076	FCM
Neuroptera	Chrysopidae	*Chrysopa perla*	0.707	691.567	FCM
Strepsiptera	Stylopidae	*Stylops ater*	0.109	106.583	FCM
Strepsiptera	Stylopidae	*Stylops ovinae*	0.055	54.243	17-mer
Trichoptera	Limnephilidae	*Limnephilus lunatus*	2.567	2510.874	17-mer

D.2.3 COI barcode sequences

For the species that were part of our TE analysis, but were not represented in the BOLD database (Table D.3 on page 369), we dowloaded COI sequences from NCBI Genbank by searching for

"species_name COI". In the cases where multiple sequences were returned, we selected the longest one. If there were multiple sequences with the longest length, we selected one at random. For *Pachypsylla venusta*, the complete mitochondrial genome was available, but not just the COI sequence. We used the COI sequence of the closely related species *Bemisia tabaci* in an alignment using MAFFT and cropped the *P. venusta* sequence to the length of the *B. tabaci* COI sequence.

Table D.3: Species not represented in the BOLD database

Order	Family	Genus	Species
Chelicerata	Buthidae	Centruroides	*C. sculpturatus*
Chelicerata	Ixodidae	Ixodes	*I. scapularis*
Chelicerata	Limulidae	Limulus	*L. polyphemus*
Chelicerata	Sicariidae	Loxosceles	*L. reclusa*
Chelicerata	Theridiidae	Latrodectus	*L. hesperus*
Chelicerata	Theridiidae	Parasteatoda	*P. tepidariorum*
Coleoptera	Buprestidae	Agrilus	*A. planipennis*
Coleoptera	Cerambycidae	Anoplophora	*A. glabripennis*
Coleoptera	Curculionidae	Dendroctonus	*D. ponderosae*
Coleoptera	Curculionidae	Hypothenemus	*H. hampei*
Coleoptera	Nitidulidae	Aethina	*A. tumida*
Coleoptera	Scarabaeidae	Onthophagus	*O. taurus*
Coleoptera	Scarabaeidae	Oryctes	*O. borbonicus*

Table D.3 –continued

Order	Family	Genus	Species
Coleoptera	Silphidae	Nicrophorus	*N. vespilloides*
Crustacea	Dogielinotidae	Hyalella	*H. azteca*
Diptera	Chironomidae	Belgica	*B. antarctica*
Diptera	Tephritidae	Ceratitis	*C. capitata*
Diptera	Tephritidae	Rhagoletis	*R. zephyria*
Hemiptera	Aphalaridae	Pachypsylla	*P. venusta*
Hemiptera	Cicadellidae	Homalodisca	*H. vitripennis*
Hemiptera	Cimicidae	Cimex	*C. lectularius*
Hemiptera	Delphacidae	Nilaparvata	*N. lugens*
Hemiptera	Gerridae	Gerris	*G. buenoi*
Hemiptera	Miridae	Halyomorpha	*H. halys*
Hymenoptera	Formicidae	Acromyrmex	*A. echinatior*
Hymenoptera	Formicidae	Atta	*A. cephalotes*
Hymenoptera	Formicidae	Harpegnathos	*H. saltator*
Hymenoptera	Formicidae	Pogonomyrmex	*P. barbatus*
Hymenoptera	Orussidae	Orussus	*O. abietinus*
Hymenoptera	Tenthredinidae	Athalia	*A. rosae*
Lepidoptera	Crambidae	Chilo	*C. suppressalis*
Lepidoptera	Geometridae	Operophtera	*O. brumata*

Table D.3 –continued

Order	Family	Genus	Species
Lepidoptera	Hesperidae	Lerema	*L. accius*
Lepidoptera	Lycaenidae	Calycopis	*C. cecrops*
Lepidoptera	Noctuidae	Helicoverpa	*H. punctigera*
Lepidoptera	Noctuidae	Spodoptera	*S. frugiperda*
Lepidoptera	Nymphalidae	Melitaea	*M. cinxia*
Lepidoptera	Pieridae	Pieris	*P. rapae*
Lepidoptera	Plutellidae	Plutella	*P. xylostella*
Lepidoptera	Pyralidae	Plodia	*P. interpunctella*
Myriapoda	Linotaeniidae	Strigamia	*S. maritima*
Odonata	Libellulidae	Ladona	*L. fulva*
Strepsiptera	Mengenillidae	Mengenilla	*M. moldrzyki*

Table D.4: Divergence times in Mya and MRCA split node numbers in the ancestral reconstruction phylogeny.

Species	MRCA node	Age
Drosophila yakuba	765	21.44269125067723
Drosophila simulans	763	12.865614687731181
Drosophila sechellia	762	25.731229532150792

Table D.4 –continued

Species	MRCA node	Age [Mya]
Drosophila melanogaster	761	38.596844376570 46
Drosophila erecta	766	21.442691250677 23
Drosophila ananassae	786	12.603051120340638
Drosophila pseudoobscura	774	31.507628033726974
Drosophila persimilis	775	15.753813938518 42
Drosophila miranda	773	47.261442128934 45
Drosophila willistoni	793	16.804068211792014
Drosophila virilis	799	20.164881885285638
Drosophila mojavensis	803	49.992103238770 88
Drosophila grimshawi	812	18.904576757561074
Rhagoletis zephyria	844	118.03069917141 25
Ceratitis capitata	844	118.03069917141 25
Lucilia cuprina	839	25.797218874700604
Musca domestica	841	55.279754909604 49
Culex quinquefasciatus	887	25.929197559787 27
Aedes albopictus	882	18.520855355803 52
Aedes aegypti	882	18.520855355803 52
Anopheles gambiae	890	22.225026457795423
Belgica antarctica	896	44.450053069747 64

Table D.4 –continued

Species	MRCA node	Age [Mya]
Mayetiola destructor	868	66.61851738853375
Papilio glaucus	963	36.555952942833926
Melitaea cinxia	957	73.11190604406304
Heliconius melpomene	959	36.555952942833926
Danaus plexippus	960	18.277976393935717
Calycopis cecrops	956	91.3898825895289
Pieris rapae	955	109.66785913842705
Lerema accius	955	109.66785913842705
Manduca sexta	948	22.84747053116029 7
Plutella xylostella	907	138.6079886790834
Spodoptera frugiperda	914	63.972917766181 24
Helicoverpa punctigera	914	63.972917766181 24
Chilo suppressalis	908	127.9458356873252
Operophtera brumata	913	74.63507075303869
Plodia interpunctella	909	117.283682700467 86
Bombyx mori	954	45.694941217283 09
Limnephilus lunatus	906	175.295995538263 13
Aethina tumida	973	150.78055424458506
Anoplophora glabripennis	975	130.67648032433073

Table D.4 –continued

Species	MRCA node	Age [Mya]
Leptinotarsa decemlineata	989	25.13009224299492
Tribolium castaneum	1001	24.4120896029859 86
Agrilus planipennis	968	201.040739045220 96
Oryctes borbonicus	1028	70.364258563567 72
Onthophagus taurus	1026	140.728517284458
Dendroctonus ponderosae	995	46.909505656603 98
Hypothenemus hampei	995	46.909505656603 98
Nicrophorus vespilloides	966	221.144812965475 35
Mengenilla moldrzyki	964	241.248886885729 63
Pogonomyrmex barbatus	1054	57.318999184451 74
Solenopsis invicta	1056	50.950221479777 27
Acromyrmex echinatior	1055	101.900443117171 85
Atta cephalotes	1055	101.900443117171 85
Harpegnathos saltator	1039	210.16966409663 45
Camponotus floridanus	1048	38.212666070428 67
Linepithema humile	1060	76.425332298474 76
Megachile rotundata	1071	145.902907258557 66
Ceratina calcarata	1072	125.059634770532 74
Bombus terrestris	1082	26.054090452413732

Table D.4 –continued

Species	MRCA node	Age [Mya]
Apis mellifera	1084	39.08113575743005
Nasonia vitripennis	1106	106.99546528091116
Copidosoma floridanum	1103	187.24206435980705
Trichogramma pretiosum	1113	106.99546528091116
Orussus abietinus	1036	267.488663441655666
Athalia rosae	1036	267.488663441655666
Pediculus humanus	713	369.25168817524080
Halyomorpha halys	1141	201.576127713029680
Oncopeltus fasciatus	1151	113.386571769549730
Cimex lectularius	1141	201.576127713029680
Gerris buenoi	1139	251.970159680732420
Nilaparvata lugens	1152	188.977619721104020
Homalodisca vitripennis	1153	125.985079761475620
Pachypsylla venusta	1122	237.571864832817250
Acyrthosiphon pisum	1129	98.988276921634680
Frankliniella occidentalis	1156	98.988276921634740
Blattella germanica	1178	272.803414609984030
Zootermopsis nevadensis	1186	45.467235636153870
Timema cristinae	1188	159.135325123069040

Table D.4 –continued

Species	MRCA node	Age [Mya]
Locusta migratoria	1173	68.20085353353676
Ephemera danica	1189	396.16785128318885
Calopteryx splendens	1220	121.05128778189425
Ladona fulva	1194	231.09791318241201
Machilis hrabei	1225	93.82705441726876
Catajapyx aquilonaris	707	489.1119896305598
Orchesella cincta	706	509.08870653964 01
Hyalella azteca	690	198.39953369400536
Eurytemora affinis	632	183.13803108993875
Daphnia pulex	640	63.95296313437109
Strigamia maritima		NA
Latrodectus hesperus		NA
Parasteatoda tepidariorum		NA
Loxosceles reclusa		NA
Centruroides sculpturatus		NA
Ixodes scapularis		NA
Limulus polyphemus		NA

D.3 TE age determination

D.3.1 Divergence times and substitution rates

To obtain order-specific substitution rates, we calculated the weighted arithmetic mean as

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{D.1}$$

with n = number of branches in the tree, w_i = branch substitution rate, t_i = branch time, $x_i = \frac{w_i}{t_i}$.

Thus, longer branches have a higher influence on the mean substitution rate than shorter branches. The results are listed in Table D.6 (page 390). Note that for the TE age classification ("agesplit"), we used species-specific divergence times derived from the time-calibrated phylogeny, listed in Table D.4.

D.3.2 DNA gain and loss

Insect order divergence times were taken from Misof et al. (2014) and are listed in Table D.6 (page 390). We used the upper and lower confidence interval as maximum and minimum age, respectively, for the time calibration of the ancestral genome size reconstruction tree. We used the splits listed in Table D.7 (page 390) as calibration points to convert the ultrametric phylogeny into a chronogram.

The inferred amounts of DNA gain and loss are listed in Table D.5 (page 384).

Correlation tests under phylogenetic independent contrasts (PIC)

Some correlations are only apparent when correcting for phylogeny. This also shows the importance of considering the phylogeny when drawing conclusions in comparative studies.

TE content and *k*:

```
Pearson's product-moment correlation


data:  pic.tes and pic.k
t = -1.9038, df = 85, p-value = 0.06032
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.396008539  0.008792809
sample estimates:
     cor
-0.20223
```

Genome size and *k*:

```
Pearson's product-moment correlation


data:  pic.size and pic.k
t = -4.0119, df = 85, p-value = 0.0001291
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5623912 -0.2056495
sample estimates:
       cor
-0.3990127
```

DNA gain and k:

```
Pearson's product-moment correlation


data:  pic.gain and pic.k
t = -2.7991, df = 85, p-value = 0.006341
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.47225571 -0.08506429
sample estimates:
       cor
-0.2905071
```

DNA loss and k:

```
Pearson's product-moment correlation


data:  pic.loss and pic.k
t = -2.1293, df = 85, p-value = 0.03612
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.41596621 -0.01510393
sample estimates:
       cor
-0.2250362
```

DNA gain and genome (assembly) size:

```
Pearson's product-moment correlation


data:  pic.gain and pic.size
t = 24.438, df = 85, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9029451 0.9575565
sample estimates:
      cor
0.9356325
```

DNA loss and genome (assembly) size:

```
Pearson's product-moment correlation


data:  pic.loss and pic.size

t = -9.1316, df = 85, p-value = 2.923e-14

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 -0.7963161 -0.5788707

sample estimates:

       cor

-0.7037099
```

Table D.5: The calculated DNA losses and gains in the 96 studied species show large variation. DNA gain is defined as the amount of clade-specific TEs, while DNA loss is calculated as the difference between ancestral clade genome size and ancestral DNA of the species (assembly size – clade-specific TEs). Clade relationships after Misof et al. (2014), intra-ordinal relationships based on published phylogenies listed in Table D.8.

Species	Ancestral size [Mbp]	Assembly [Mbp]	Gain [Mbp]	loss [Mbp]
Drosophila yakuba	253.38	162.60	40.08	130.86
Drosophila simulans	253.38	124.61	12.16	140.94
Drosophila sechellia	253.38	157.25	39.16	135.29
Drosophila melanogaster	253.38	142.57	29.66	140.47
Drosophila erecta	253.38	145.08	31.48	139.77
Drosophila ananassae	253.38	213.92	94.87	134.33

Table D.5 –continued

Species	Ancestral size [Mbp]	Assembly [Mbp]	Gain [Mbp]	loss [Mbp]
Drosophila pseudoobscura	253.38	149.03	26.28	130.63
Drosophila persimilis	253.38	175.58	55.51	133.31
Drosophila miranda	253.38	132.59	12.85	133.65
Drosophila willistoni	253.38	223.61	79.19	108.96
Drosophila virilis	253.38	189.21	49.44	113.62
Drosophila mojavensis	253.38	180.21	42.21	115.39
Drosophila grimshawi	253.38	186.09	43.33	110.62
Rhagoletis zephyria	346.78	1045.32	122.69	-575.85
Ceratitis capitata	346.78	440.70	386.49	292.57
Lucilia cuprina	585.89	379.07	539.20	746.02
Musca domestica	585.89	691.74	146.18	40.32
Culex quinquefasciatus	606.87	539.96	276.99	343.90
Aedes albopictus	1011.02	1776.29	987.68	222.41
Aedes aegypti	1011.02	1310.09	802.33	503.27
Anopheles gambiae	1011.02	252.44	50.54	809.12
Belgica antarctica	149.71	88.99	2.51	63.23
Mayetiola destructor	156.92	153.14	18.54	22.32
Papilio glaucus	401.06	361.20	99.63	139.49
Melitaea cinxia	378.11	361.02	112.72	129.81

Table D.5 –continued

Species	Ancestral size [Mbp]	Assembly [Mbp]	Gain [Mbp]	loss [Mbp]
Heliconius melpomene	352.65	269.65	82.52	165.51
Danaus plexippus	330.92	272.28	30.46	89.10
Calycopis cecrops	363.37	689.13	272.59	-53.17
Pieris rapae	368.94	242.73	58.41	184.63
Lerema accius	368.94	289.62	51.91	131.23
Manduca sexta	456.58	399.66	106.07	162.99
Plutella xylostella	472.47	186.03	35.13	321.57
Spodoptera frugiperda	687.01	330.62	70.14	426.53
Helicoverpa punctigera	687.01	350.24	73.87	410.64
Chilo suppressalis	471.68	314.17	117.03	274.54
Operophtera brumata	685.17	624.73	307.85	368.29
Plodia interpunctella	513.30	364.62	74.27	222.95
Bombyx mori	499.30	431.73	183.81	251.39
Limnephilus lunatus	544.65	804.08	413.69	154.26
Aethina tumida	657.89	234.34	30.69	454.24
Anoplophora glabripennis	710.89	602.43	291.74	400.21
Leptinotarsa decemlineata	688.55	678.27	366.13	376.41
Tribolium castaneum	234.88	151.32	44.27	127.84
Agrilus planipennis	627.58	252.63	88.62	463.57

Table D.5 –continued

Species	Ancestral size [Mbp]	Assembly [Mbp]	Gain [Mbp]	loss [Mbp]
Oryctes borbonicus	909.82	423.77	125.37	611.42
Onthophagus taurus	723.50	238.61	95.65	580.53
Dendroctonus ponderosae	1349.29	201.82	40.01	1187.48
Hypothenemus hampei	1349.29	130.55	24.22	1242.96
Nicrophorus vespilloides	581.20	192.10	22.62	411.71
Mengenilla moldrzyki	438.32	155.73	75.48	358.07
Pogonomyrmex barbatus	268.29	220.21	32.03	80.11
Solenopsis invicta	426.50	354.73	117.46	189.23
Acromyrmex echinatior	343.31	288.58	80.40	135.13
Atta cephalotes	343.31	281.25	73.86	135.92
Harpegnathos saltator	368.28	283.10	70.01	155.19
Camponotus floridanus	321.96	224.63	28.21	125.53
Linepithema humile	257.70	213.27	25.64	70.07
Megachile rotundata	476.16	265.92	59.22	269.46
Ceratina calcarata	476.16	183.85	24.48	316.78
Bombus terrestris	398.11	236.41	27.08	188.77
Apis mellifera	251.45	229.11	11.74	34.09
Nasonia vitripennis	426.41	238.62	60.20	247.99
Copidosoma floridanum	334.55	454.98	175.59	55.16

Table D.5 –continued

Species	Ancestral size [Mbp]	Assembly [Mbp]	Gain [Mbp]	loss [Mbp]
Trichogramma pretiosum	254.06	181.15	26.68	99.59
Orussus abietinus	358.53	186.48	39.80	211.85
Athalia rosae	358.53	156.83	8.17	209.87
Pediculus humanus	387.75	108.40	8.70	288.06
Halyomorpha halys	953.82	1000.80	447.22	400.24
Oncopeltus fasciatus	1833.80	773.64	229.56	1289.72
Cimex lectularius	953.82	513.62	194.56	634.76
Gerris buenoi	913.26	653.32	244.59	504.52
Nilaparvata lugens	1321.64	1017.42	434.43	738.65
Homalodisca vitripennis	2444.86	1325.90	317.52	1436.48
Pachypsylla venusta	605.50	371.84	168.94	402.59
Acyrthosiphon pisum	437.89	499.89	146.41	84.40
Frankliniella occidentalis	411.61	263.81	42.29	190.10
Blattella germanica	1833.68	1710.49	842.24	965.43
Zootermopsis nevadensis	901.50	464.44	43.52	480.58
Timema cristinae	1850.01	844.26	405.77	1411.51
Locusta migratoria	9201.19	5759.80	3658.54	7099.94
Ephemera danica	1070.02	399.55	109.20	779.68
Calopteryx splendens	1174.90	1324.05	226.06	76.92

Table D.5 –continued

Species	Ancestral size [Mbp]	Assembly [Mbp]	Gain [Mbp]	loss [Mbp]
Ladona fulva	809.51	948.04	191.87	53.34
Machilis hrabei	2780.23	1322.99	605.88	2063.13
Catajapyx aquilonaris	1256.19	311.13	87.38	1032.44
Orchesella cincta	1273.33	286.75	37.42	1024.00
Hyalella azteca	5897.57	596.63	136.37	5437.31
Eurytemora affinis	906.42	387.57	129.86	648.70
Daphnia pulex	313.54	158.61	42.48	197.41
Strigamia maritima	2171.16	173.60	73.69	2071.25
Latrodectus hesperus	2171.16	726.41	192.24	1636.99
Parasteatoda tepidariorum	2171.16	1141.93	442.18	1471.40
Loxosceles reclusa	2171.16	1793.28	1077.48	1455.35
Centruroides sculpturatus	2171.16	627.51	221.45	1765.10
Ixodes scapularis	2171.16	1388.47	707.90	1490.59
Limulus polyphemus	2171.16	1706.69	608.32	1072.79

Table D.6: Divergence times and clade-specific substitution rates for the arthropod orders in this study. Substitution rates are in substitutions $\times$ position^{-1} $\times$ My^{-1}.

Clade	Species	Divergence time [Mya]	Substitution rate
Diptera	23	157.83	0.0067943
Lepidoptera	15	141.47	0.0066874
Trichoptera	1	154.32	0.004798
Coleoptera	10	269.98	0.0034029
Strepsiptera	1	107.56	0.0069976
Hymenoptera	16	239.53	0.0036226
Phthiraptera	1	187.33	0.0059665
Hemiptera: Heteroptera	4	155.56	0.0059329
Hemiptera: Auchenorrhyncha	2	169.58	0.0040139
Hemiptera: Sternorrhyncha	2	245.03	0.0049467
Thysanoptera	1	119.93	0.005508
Blattodea + Isoptera	1	172.98	0.0019926
Isoptera	1	135.83	0.0011611
Phasmatodea	1	124.71	0.0039888
Orthoptera	1	202.7	0.0037212
Ephemeroptera	1	178.82	0.00412
Odonata	2	234.73	0.0013438
Archaeognatha	1	145.65	0.0029923
Diplura	1	303.4	0.0027449
Collembola	1	242.69	0.0043442
Malacostraca	1	254.21	0.0032646
Copepoda + Branchiopoda	2	399.32	0.0038017
Myriapoda	1	407.25	0.002289
Chelicerata	6	568.82	0.0011044

Table D.7: Calibration points. Minimum and maximum age are in Mya and correspond to the boundaries of the 95 % confidence interval of the node dating by Misof et al. (2014).

Clade	Min. age [Mya]	Max. age [Mya]
Copepoda + Branchiopoda	222.87	500.75
Thysanoptera + Hemiptera	287.77	379.13
Psocodea	124.09	279.64
Hymenoptera	221.00	280.62
Lepidoptera	119.49	172.27
Diptera	114.90	202.04

D.4 Order-level phylogenies

The backbone phylogeny (order topology) was based on Misof et al. (2014). For intra-ordinal species relationships, we used the sources listed in Table D.8 (page 392) to build the constraint topology for inferring branch lengths based on COI barcode sequences. We used the constraint topology to estimate branch lengths using RAxML v8.2.11:

```
raxml -s COI_nt_seq.afa -n COI -m GTRCAT -p 1 -g CONSTRAINT.tree
```

The resulting tree was rendered ultrametric with the following short Python script using the ETE3 toolkit (Huerta-Cepas et al., 2016):

```
#!/usr/bin/python3

import sys

from ete3 import Tree

t = Tree(sys.argv[1])

t.convert_to_ultrametric()

print(t.write())
```

Table D.8: References for the constraint phylogeny.

Clade	Sources
Archaeognatha	COI
Isoptera	Cameron et al. (2012)
Odonata	Letsch et al. (2016)
Blattodea	Wang et al. (2017)
Orthoptera	Zhang et al. (2013)
Diptera	Wiegmann et al. (2011); Cranston et al. (2011)
Hemiptera	Song et al. (2012); Ortiz-Rivas & Martínez-Torres (2010); Nováková et al. (2013)
Hymenoptera	Peters et al. (2017); Branstetter et al. (2017); Ward et al. (2015)
Strepsiptera	Pohl & Beutel (2005)
Coleoptera	McKenna et al. (2015); Ahrens et al. (2014); Magro et al. (2010); Kergoat et al. (2014); Hundsdoerfer et al. (2009)
Lepidoptera	Breinholt et al. (2018); Regier et al. (2013); Kawahara et al. (2009); Mitchell et al. (2005); Abraham et al. (2001)
Malacostraca	Tsang et al. (2008); Ahyong & O'Meally (2004)
Copepoda	Eyun (2017); Blanco-Bercial et al. (2011); Figueroa (2011); Thum (2004)
Branchiopoda	Richter et al. (2007)

References

Abraham, D., Ryrholm, N., Wittzell, H., Holloway, J. D., Scoble, M. J., & Löfstedt, C. (2001). Molecular Phylogeny of the Subfamilies in Geometridae (Geometroidea: Lepidoptera). *Molecular Phylogenetics and Evolution*, 20(1), 65–77.

Adams, M. D. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461), 2185–2195.

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., Tanskanen, J., Hornett, E. A., Ferguson, L. C., Luo, S., Cao, Z., de Jong, M. A., Duplouy, A., Smolander, O.-P., Vogel, H., McCoy, R. C., Qian, K., Chong, W. S., Zhang, Q., Ahmad, F., Haukka, J. K., Joshi, A., Salojärvi, J., Wheat, C. W., Grosse-Wilde, E., Hughes, D., Katainen, R., Pitkänen, E., Ylinen, J., Waterhouse, R. M., Turunen, M., Vähärautio, A., Ojanen, S. P., Schulman, A. H., Taipale, M., Lawson, D., Ukkonen, E., Mäkinen, V., Goldsmith, M. R., Holm, L., Auvinen, P., Frilander, M. J., & Hanski, I. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications*, 5, 4737.

Ahrens, D., Schwarzer, J., & Vogler, A. P. (2014). The evolution of scarab beetles tracks the

sequential rise of angiosperms and mammals. *Proceedings of the Royal Society B: Biological Sciences*, 281(1791), 20141470–20141470.

Ahyong, S. & O'Meally, D. (2004). Phylogeny of the Decapoda Reptantia: Resolution using three molecular loci and morphology. *Raffles Bulletin of Zoology*, 52, 673–693.

Arensburger, P., Megy, K., Waterhouse, R. M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F., Campbell, C. L., Campbell, K. S., Casola, C., Castro, M. T., Chandramouliswaran, I., Chapman, S. B., Christley, S., Costas, J., Eisenstadt, E., Feschotte, C., Fraser-Liggett, C., Guigo, R., Haas, B., Hammond, M., Hansson, B. S., Hemingway, J., Hill, S. R., Howarth, C., Ignell, R., Kennedy, R. C., Kodira, C. D., Lobo, N. F., Mao, C., Mayhew, G., Michel, K., Mori, A., Liu, N., Naveira, H., Nene, V., Nguyen, N., Pearson, M. D., Pritham, E. J., Puiu, D., Qi, Y., Ranson, H., Ribeiro, J. M. C., Roberston, H. M., Severson, D. W., Shumway, M., Stanke, M., Strausberg, R. L., Sun, C., Sutton, G., Tu, Z., Tubio, J. M. C., Unger, M. F., Vanlandingham, D. L., Vilella, A. J., White, O., White, J. R., Wondji, C. S., Wortman, J., Zdobnov, E. M., Birren, B., Christensen, B. M., Collins, F. H., Cornel, A., Dimopoulos, G., Hannick, L. I., Higgs, S., Lanzaro, G. C., Lawson, D., Lee, N. H., Muskavitch, M. A. T., Raikhel, A. S., & Atkinson, P. W. (2010). Sequencing of Culex quinquefasciatus Establishes a Platform for Mosquito Comparative Genomics. *Science*, 330(6000), 86–88.

Blanco-Bercial, L., Bradford-Grieve, J., & Bucklin, A. (2011). Molecular phylogeny of the Calanoida (Crustacea: Copepoda). *Molecular Phylogenetics and Evolution*, 59(1), 103–113.

Bonasio, R., Zhang, G., Ye, C., Mutti, N. S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., Zhang, P., Huang, Z., Berger, S. L., Reinberg, D., Wang, J., & Liebig, J. (2010). Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. *Science (New York, N.Y.)*, 329(5995), 1068–1071.

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., Gates, M. W., Kula, R. R., & Brady, S. G. (2017). Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. *Current Biology*, 27(7), 1019–1025.

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., & Kawahara, A. Y. (2018). Resolving Relationships among the Megadiverse Butterflies and Moths with a Novel Pipeline for Anchored Phylogenomics. *Systematic Biology*, 67(1), 78–93.

Cameron, S. L., Lo, N., Bourguignon, T., Svenson, G. J., & Evans, T. A. (2012). A mitochondrial genome phylogeny of termites (Blattodea: Termitoidae): Robust support for interfamilial relationships and molecular synapomorphies define major clades. *Molecular Phylogenetics and Evolution*, 65(1), 163–173.

Chen, X.-G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., Zhang, C., Bonizzoni, M., Dermauw, W., Vontas, J., Armbruster, P., Huang, X., Yang, Y., Zhang, H., He, W., Peng, H., Liu, Y., Wu, K., Chen, J., Lirakis, M., Topalis, P., Van Leeuwen, T., Hall, A. B., Jiang, X., Thorpe, C., Mueller, R. L., Sun, C., Waterhouse, R. M., Yan, G., Tu, Z. J., Fang, X., & James, A. A. (2015). Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44), E5907–5915.

Chipman, A. D., Ferrier, D. E. K., Brena, C., Qu, J., Hughes, D. S. T., Schröder, R., Torres-Oliva, M., Znassi, N., Jiang, H., Almeida, F. C., Alonso, C. R., Apostolou, Z., Aqrawi, P., Arthur, W., Barna, J. C. J., Blankenburg, K. P., Brites, D., Capella-Gutiérrez, S., Coyle, M., Dearden, P. K., Pasquier, L. D., Duncan, E. J., Ebert, D., Eibner, C., Erikson, G., Evans, P. D., Extavour, C. G., Francisco, L., Gabaldón, T., Gillis, W. J., Goodwin-Horn, E. A., Green, J. E., Griffiths-Jones, S., Grimmelikhuijzen, C. J. P., Gubbala, S., Guigó, R., Han, Y., Hauser, F., Havlak, P., Hayden, L., Helbing, S., Holder, M., Hui, J. H. L., Hunn, J. P., Hunnekuhl, V. S., Jackson, L., Javaid, M., Jhangiani, S. N., Jiggins, F. M., Jones, T. E., Kaiser, T. S., Kalra, D., Kenny, N. J., Korchina, V., Kovar, C. L., Kraus, F. B., Lapraz, F., Lee, S. L., Lv, J., Mandapat, C., Manning, G., Mariotti, M., Mata, R., Mathew, T., Neumann, T., Newsham, I., Ngo, D. N., Ninova, M., Okwuonu, G., Ongeri, F., Palmer, W. J., Patil, S., Patraquim, P., Pham, C., Pu, L.-L., Putman, N. H., Rabouille, C., Ramos, O. M., Rhodes, A. C., Robertson, H. E., Robertson, H. M., Ronshaugen, M., Rozas, J., Saada, N., Sánchez-Gracia, A., Scherer, S. E., Schurko, A. M., Siggens, K. W., Simmons, D., Stief, A., Stolle, E., Telford, M. J., Tessmar-Raible, K., Thornton, R., van der Zee, M., von Haeseler, A., Williams, J. M., Willis, J. H., Wu, Y., Zou, X., Lawson, D., Muzny, D. M., Worley, K. C., Gibbs, R. A., Akam, M., & Richards, S. (25-Nov-2014). The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. *PLOS Biology*, 12(11), e1002005.

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Caceres, C. E., Carmel, L., Casola, C., Choi, J.-H., Detter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., Frohlich, T., Geiler-

Samerotte, K. A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E. V., Kultz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J. R., Muller, J., Pangilinan, J., Patwardhan, R. P., Pitluck, S., Pritham, E. J., Rechtsteiner, A., Rho, M., Rogozin, I. B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y. I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J. R., Andrews, J., Crease, T. J., Tang, H., Lucas, S. M., Robertson, H. M., Bork, P., Koonin, E. V., Zdobnov, E. M., Grigoriev, I. V., Lynch, M., & Boore, J. L. (2011). The Ecoresponsive Genome of Daphnia pulex. *Science*, 331(6017), 555–561.

Cong, Q., Borek, D., Otwinowski, Z., & Grishin, N. V. (2015a). Skipper genome sheds light on unique phenotypic traits and phylogeny. *BMC Genomics*, 16(1).

Cong, Q., Borek, D., Otwinowski, Z., & Grishin, N. V. (2015b). Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. *Cell Reports*, 10(6), 910–919.

Cong, Q., Shen, J., Borek, D., Robbins, R. K., Otwinowski, Z., & Grishin, N. V. (2016). Complete genomes of Hairstreak butterflies, their speciation, and nucleo-mitochondrial incongruence. *Scientific Reports*, 6, 24863.

Cranston, P. S., Hardy, N. B., & Morse, G. E. (2011). A dated molecular phylogeny for the Chironomidae (Diptera). *Systematic Entomology*, 37(1), 172–188.

Cunningham, C. B., Ji, L., Wiberg, R. A. W., Shelton, J., McKinney, E. C., Parker, D. J., Meagher, R. B., Benowitz, K. M., Roy-Zokan, E. M., Ritchie, M. G., Brown, S. J., Schmitz,

R. J., & Moore, A. J. (2015). The Genome and Methylome of a Beetle with Complex Social Behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biology and Evolution*, 7(12), 3383–3396.

Derks, M. F. L., Smit, S., Salis, L., Schijlen, E., Bossers, A., Mateman, C., Pijl, A. S., de Ridder, D., Groenen, M. A. M., Visser, M. E., & Megens, H.-J. (2015). The Genome of Winter Moth (Operophtera brumata) Provides a Genomic Perspective on Sexual Dimorphism and Phenology. *Genome Biology and Evolution*, 7(8), 2321–2332.

Doležel, J., Bartoš, J., Voglmayr, H., & Greilhuber, J. (2003). Letter to the editor. *Cytometry Part A*, 51A(2), 127–128.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167), 203–218.

Evans, J. D., McKenna, D., Scully, E., Cook, S. C., Dainat, B., Egekwu, N., Grubbs, N., Lopez, D., Lorenzen, M. D., Reyna, S. M., Rinkevich, F. D., Neumann, P., & Huang, Q. (2018). Genome of the small hive beetle (Aethina tumida, Coleoptera: Nitidulidae), a worldwide parasite of social bee colonies, provides insights into detoxification and herbivory. *GigaScience*, 7(12).

Eyun, S.-i. (2017). Phylogenomic Analysis of Copepoda (Arthropoda, Crustacea) Reveals Unexpected Similarities with Earlier Proposed Morphological Phylogenies. *BMC Evolutionary Biology*, 17, 23.

Eyun, S.-i., Soh, H. Y., Posavi, M., Munro, J. B., Hughes, D. S., Murali, S. C., Qu, J., Dugan, S., Lee, S. L., Chao, H., Dinh, H., Han, Y., Doddapaneni, H., Worley, K. C., Muzny, D. M., Park, E.-O., Silva, J. C., Gibbs, R. A., Richards, S., & Lee, C. E. (2017). Evolutionary History of Chemosensory-Related Gene Families across the Arthropoda. *Molecular Biology and Evolution*, 34(8), 1838–1862.

Faddeeva-Vakhrusheva, A., Derks, M. F. L., Anvar, S. Y., Agamennone, V., Suring, W., Smit, S., Straalen, V., M, N., & Roelofs, D. (2016). Gene Family Evolution Reflects Adaptation to Soil Environmental Stressors in the Genome of the Collembolan Orchesella cincta. *Genome Biology and Evolution*, 8(7), 2106–2117.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1), 1–15.

Figueroa, D. F. (2011). Phylogenetic Analysis of Ridgewayia (Copepoda: Calanoida) from the Galapagos and of a New Species from the Florida Keys With a Reevaluation of the Phylogeny of Calanoida. *Journal of Crustacean Biology*, 31(1), 153–165.

Gouin, A., Bretaudeau, A., Nam, K., Gimenez, S., Aury, J.-M., Duvic, B., Hilliou, F., Durand, N., Montagné, N., Darboux, I., Kuwar, S., Chertemps, T., Siaussat, D., Bretschneider, A., Moné, Y., Ahn, S.-J., Hänniger, S., Grenet, A.-S. G., Neunemann, D., Maumus, F., Luyten, I., Labadie, K., Xu, W., Koutroumpa, F., Escoubas, J.-M., Llopis, A., Maïbèche-Coisne, M., Salasc, F., Tomar, A., Anderson, A. R., Khan, S. A., Dumas, P., Orsucci, M., Guy, J., Belser, C., Alberti, A., Noel, B., Couloux, A., Mercier, J., Nidelet, S., Dubois, E., Liu, N.-Y., Boulogne, I., Mirabeau, O., Goff, G., Gordon, K., Oakeshott, J., Consoli, F. L., Volkoff, A.-N., Fescemyer,

H. W., Marden, J. H., Luthe, D. S., Herrero, S., Heckel, D. G., Wincker, P., Kergoat, G. J., Amselem, J., Quesneville, H., Groot, A. T., Jacquin-Joly, E., Nègre, N., Lemaitre, C., Legeai, F., d'Alençon, E., & Fournier, P. (2017). Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda , Noctuidae) with different host-plant ranges. *Scientific Reports*, 7(1), 11816.

Gulia-Nuss, M., Nuss, A. B., Meyer, J. M., Sonenshine, D. E., Roe, R. M., Waterhouse, R. M., Sattelle, D. B., de la Fuente, J., Ribeiro, J. M., Megy, K., Thimmapuram, J., Miller, J. R., Walenz, B. P., Koren, S., Hostetler, J. B., Thiagarajan, M., Joardar, V. S., Hannick, L. I., Bidwell, S., Hammond, M. P., Young, S., Zeng, Q., Abrudan, J. L., Almeida, F. C., Ayllón, N., Bhide, K., Bissinger, B. W., Bonzon-Kulichenko, E., Buckingham, S. D., Caffrey, D. R., Caimano, M. J., Croset, V., Driscoll, T., Gilbert, D., Gillespie, J. J., Giraldo-Calderón, G. I., Grabowski, J. M., Jiang, D., Khalil, S. M. S., Kim, D., Kocan, K. M., Koči, J., Kuhn, R. J., Kurtti, T. J., Lees, K., Lang, E. G., Kennedy, R. C., Kwon, H., Perera, R., Qi, Y., Radolf, J. D., Sakamoto, J. M., Sánchez-Gracia, A., Severo, M. S., Silverman, N., Šimo, L., Tojo, M., Tornador, C., Van Zee, J. P., Vázquez, J., Vieira, F. G., Villar, M., Wespiser, A. R., Yang, Y., Zhu, J., Arensburger, P., Pietrantonio, P. V., Barker, S. C., Shao, R., Zdobnov, E. M., Hauser, F., Grimmelikhuijzen, C. J. P., Park, Y., Rozas, J., Benton, R., Pedra, J. H. F., Nelson, D. R., Unger, M. F., Tubio, J. M. C., Tu, Z., Robertson, H. M., Shumway, M., Sutton, G., Wortman, J. R., Lawson, D., Wikel, S. K., Nene, V. M., Fraser, C. M., Collins, F. H., Birren, B., Nelson, K. E., Caler, E., & Hill, C. A. (2016). Genomic insights into the Ixodes scapularis tick vector of Lyme disease. *Nature Communications*, 7, 10507.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M. C., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., Atkinson, P. W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G. K., Chrystal, M. A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C. A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M. E., Hladun, S. L., Hogan, J. R., Hong, Y. S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.-J., Lobo, N. F., Lopez, J. R., Malek, J. A., McIntosh, T. C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S. D., O'Brochta, D. A., Pfannkoch, C., Qi, R., Regier, M. A., Remington, K., Shao, H., Sharakhova, M. V., Sitter, C. D., Shetty, J., Smith, T. J., Strong, R., Sun, J., Thomasova, D., Ton, L. Q., Topalis, P., Tu, Z., Unger, M. F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K. J., Wortman, J. R., Wu, M., Yao, A., Zdobnov, E. M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S. C., Zhimulev, I., Coluzzi, M., della Torre, A., Roth, C. W., Louis, C., Kalush, F., Mural, R. J., Myers, E. W., Adams, M. D., Smith, H. O., Broder, S., Gardner, M. J., Fraser, C. M., Birney, E., Bork, P., Brey, P. T., Venter, J. C., Weissenbach, J., Kafatos, F. C., Collins, F. H., & Hoffman, S. L. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. *Science (New York, N.Y.)*, 298(5591), 129–149.

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114), 931–949.

Hozza, M., Vinař, T., & Brejová, B. (2015). How Big is that Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra. In *String Processing and Information Retrieval*, Lecture Notes in Computer Science (pp. 199–209).: Springer, Cham.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6), 1635–1638.

Hundsdoerfer, A. K., Rheinheimer, J., & Wink, M. (2009). Towards the phylogeny of the Curculionoidea (Coleoptera): Reconstructions from mitochondrial and nuclear ribosomal DNA sequences. *Zoologischer Anzeiger - A Journal of Comparative Zoology*, 248(1), 9–31.

International Silkworm Genome Consortium (2008). The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect Biochemistry and Molecular Biology*, 38(12), 1036–1045.

Kanost, M. R., Arrese, E. L., Cao, X., Chen, Y.-R., Chellapilla, S., Goldsmith, M. R., Grosse-Wilde, E., Heckel, D. G., Herndon, N., Jiang, H., Papanicolaou, A., Qu, J., Soulages, J. L., Vogel, H., Walters, J., Waterhouse, R. M., Ahn, S.-J., Almeida, F. C., An, C., Aqrawi, P., Bretschneider, A., Bryant, W. B., Bucks, S., Chao, H., Chevignon, G., Christen, J. M., Clarke, D. F., Dittmer, N. T., Ferguson, L. C. F., Garavelou, S., Gordon, K. H. J., Gunaratna, R. T., Han, Y., Hauser, F., He, Y., Heidel-Fischer, H., Hirsh, A., Hu, Y., Jiang, H., Kalra, D., Klinner, C., König, C., Kovar, C., Kroll, A. R., Kuwar, S. S., Lee, S. L., Lehman, R., Li, K., Li, Z., Liang, H., Lovelace, S., Lu, Z., Mansfield, J. H., McCulloch, K. J., Mathew, T., Morton, B., Muzny, D. M., Neunemann, D., Ongeri, F., Pauchet, Y., Pu, L.-L., Pyrousis, I., Rao, X.-J., Redding, A., Roesel, C., Sanchez-Gracia, A., Schaack, S., Shukla, A., Tetreau, G., Wang, Y.,

Xiong, G.-H., Traut, W., Walsh, T. K., Worley, K. C., Wu, D., Wu, W., Wu, Y.-Q., Zhang, X., Zou, Z., Zucker, H., Briscoe, A. D., Burmester, T., Clem, R. J., Feyereisen, R., Grimmelikhuijzen, C. J. P., Hamodrakas, S. J., Hansson, B. S., Huguet, E., Jermiin, L. S., Lan, Q., Lehman, H. K., Lorenzen, M., Merzendorfer, H., Michalopoulos, I., Morton, D. B., Muthukrishnan, S., Oakeshott, J. G., Palmer, W., Park, Y., Passarelli, A. L., Rozas, J., Schwartz, L. M., Smith, W., Southgate, A., Vilcinskas, A., Vogt, R., Wang, P., Werren, J., Yu, X.-Q., Zhou, J.-J., Brown, S. J., Scherer, S. E., Richards, S., & Blissard, G. W. (2016). Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. *Insect Biochemistry and Molecular Biology*, 76, 118–147.

Kawahara, A. Y., Mignault, A. A., Regier, J. C., Kitching, I. J., & Mitter, C. (2009). Phylogeny and Biogeography of Hawkmoths (Lepidoptera: Sphingidae): Evidence from Five Nuclear Genes. *PLOS ONE*, 4(5), e5719.

Keeling, C. I., Yuen, M. M., Liao, N. Y., Roderick Docking, T., Chan, S. K., Taylor, G. A., Palmquist, D. L., Jackman, S. D., Nguyen, A., Li, M., Henderson, H., Janes, J. K., Zhao, Y., Pandoh, P., Moore, R., Sperling, F. A., W Huber, D. P., Birol, I., Jones, S. J., & Bohlmann, J. (2013). Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. *Genome Biology*, 14(3), R27.

Kelley, J. L., Peyton, J. T., Fiston-Lavier, A.-S., Teets, N. M., Yee, M.-C., Johnston, J. S., Bustamante, C. D., Lee, R. E., & Denlinger, D. L. (2014). Compact Genome of the Antarctic Midge Is Likely an Adaptation to an Extreme Environment. *Nature Communications*, 5.

Kergoat, G. J., Soldati, L., Clamens, A.-L., Jourdan, H., Jabbour-Zahab, R., Genson, G., Bouchard, P., & Condamine, F. L. (2014). Higher level molecular phylogeny of darkling beetles (Coleoptera: Tenebrionidae): Darkling beetle phylogeny. *Systematic Entomology*, 39(3), 486–499.

Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., Lee, S. H., Robertson, H. M., Kennedy, R. C., Elhaik, E., Gerlach, D., Kriventseva, E. V., Elsik, C. G., Graur, D., Hill, C. A., Veenstra, J. A., Walenz, B., Tubío, J. M. C., Ribeiro, J. M. C., Rozas, J., Johnston, J. S., Reese, J. T., Popadic, A., Tojo, M., Raoult, D., Reed, D. L., Tomoyasu, Y., Kraus, E., Mittapalli, O., Margam, V. M., Li, H.-M., Meyer, J. M., Johnson, R. M., Romero-Severson, J., VanZee, J. P., Alvarez-Ponce, D., Vieira, F. G., Aguadé, M., Guirao-Rico, S., Anzola, J. M., Yoon, K. S., Strycharz, J. P., Unger, M. F., Christley, S., Lobo, N. F., Seufferheld, M. J., Wang, N., Dasch, G. A., Struchiner, C. J., Madey, G., Hannick, L. I., Bidwell, S., Joardar, V., Caler, E., Shao, R., Barker, S. C., Cameron, S., Bruggner, R. V., Regier, A., Johnson, J., Viswanathan, L., Utterback, T. R., Sutton, G. G., Lawson, D., Waterhouse, R. M., Venter, J. C., Strausberg, R. L., Berenbaum, M. R., Collins, F. H., Zdobnov, E. M., & Pittendrigh, B. R. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences*, 107(27), 12168–12173.

Letsch, H., Gottsberger, B., & Ware, J. L. (2016). Not going with the flow: A comprehensive time-calibrated phylogeny of dragonflies (Anisoptera: Odonata: Insecta) provides evidence for the role of lentic habitats on diversification. *Molecular Ecology*, 25(6), 1340–1353.

Magro, A., Lecompte, E., Magné, F., Hemptinne, J.-L., & Crouau-Roy, B. (2010). Phylogeny of ladybirds (Coleoptera: Coccinellidae): Are the subfamilies monophyletic? *Molecular Phylogenetics and Evolution*, 54(3), 833–848.

McGaugh, S. E. & Noor, M. A. F. (2012). Genomic impacts of chromosomal inversions in parapatric Drosophila species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587), 422–429.

McKenna, D. D., Wild, A. L., Kanda, K., Bellamy, C. L., G., B. R., Caterino Michael S., Farnum Charles W., Hawks David C., Ivie Michael A., Jameson Mary Liz, Leschen Richard a. B., Marvaldi Adriana E., Mchugh Joseph V., Newton Alfred F., Robertson James A., Thayer Margaret K., Whiting Michael F., Lawrence John F., Ślipiński Adam, Maddison David R., & Farrell Brian D. (2015). The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Systematic Entomology*, 40(4), 835–880.

Meyer, J. M., Markov, G. V., Baskaran, P., Herrmann, M., Sommer, R. J., & Rödelsperger, C. (2016). Draft Genome of the Scarab Beetle Oryctes borbonicus on La Réunion Island. *Genome Biology and Evolution*, 8(7), 2093–2105.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li,

Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von, R. B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K., & Zhou, X. (2014). Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science*, 346, 763–7.

Mitchell, A., Mitter, C., & Regier, J. C. (2005). Systematics and evolution of the cutworm moths (Lepidoptera: Noctuidae): Evidence from two protein-coding nuclear genes: Molecular systematics of Noctuidae. *Systematic Entomology*, 31(1), 21–46.

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Kodira, C., Tu, Z. J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., Ren, Q., Zdobnov, E. M., Lobo, N. F., Campbell, K. S., Brown, S. E., Bonaldo, M. F., Zhu, J., Sinkins, S. P., Hogenkamp, D. G., Amedeo, P., Arensburger, P., Atkinson, P. W., Bidwell, S., Biedler, J., Birney, E., Bruggner, R. V., Costas, J., Coy, M. R., Crabtree, J., Crawford, M., deBruyn, B., DeCaprio, D., Eiglmeier, K., Eisenstadt, E., El-Dorry, H., Gelbart, W. M., Gomes, S. L., Hammond, M., Hannick, L. I., Hogan, J. R., Holmes, M. H., Jaffe, D., Johnston, J. S., Kennedy, R. C., Koo, H., Kravitz, S., Kriventseva, E. V., Kulp, D., LaButti, K., Lee, E., Li, S., Lovin, D. D., Mao, C., Mauceli, E., Menck, C. F. M., Miller, J. R., Montgomery, P., Mori, A., Nascimento, A. L., Naveira, H. F., Nusbaum, C., O'Leary, S.,

Orvis, J., Pertea, M., Quesneville, H., Reidenbach, K. R., Rogers, Y.-H., Roth, C. W., Schneider, J. R., Schatz, M., Shumway, M., Stanke, M., Stinson, E. O., Tubio, J. M. C., VanZee, J. P., Verjovski-Almeida, S., Werner, D., White, O., Wyder, S., Zeng, Q., Zhao, Q., Zhao, Y., Hill, C. A., Raikhel, A. S., Soares, M. B., Knudson, D. L., Lee, N. H., Galagan, J., Salzberg, S. L., Paulsen, I. T., Dimopoulos, G., Collins, F. H., Birren, B., Fraser-Liggett, C. M., & Severson, D. W. (2007). Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. *Science*, 316(5832), 1718–1723.

Niehuis, O., Hartig, G., Grath, S., Pohl, H., Lehmann, J., Tafer, H., Donath, A., Krauss, V., Eisenhardt, C., Hertel, J., Petersen, M., Mayer, C., Meusemann, K., Peters, R. S., Stadler, P. F., Beutel, R. G., Bornberg-Bauer, E., McKenna, D. D., & Misof, B. (2012). Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera. *Current Biology*, 22(14), 1309–1313.

Nováková, E., Hypša, V., Klein, J., Foottit, R. G., von Dohlen, C. D., & Moran, N. A. (2013). Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont Buchnera aphidicola. *Molecular Phylogenetics and Evolution*, 68(1), 42–54.

Nygaard, S., Zhang, G., Schiøtt, M., Li, C., Wurm, Y., Hu, H., Zhou, J., Ji, L., Qiu, F., Rasmussen, M., Pan, H., Hauser, F., Krogh, A., Grimmelikhuijzen, C. J. P., Wang, J., & Boomsma, J. J. (2011). The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. *Genome Research*, 21(8), 1339–1348.

Ortiz-Rivas, B. & Martínez-Torres, D. (2010). Combination of molecular data support the existence of three main lineages in the phylogeny of aphids (Hemiptera: Aphididae) and the

basal position of the subfamily Lachninae. *Molecular Phylogenetics and Evolution*, 55(1), 305–317.

Papanicolaou, A., Schetelig, M. F., Arensburger, P., Atkinson, P. W., Benoit, J. B., Bourtzis, K., Castañera, P., Cavanaugh, J. P., Chao, H., Childers, C., Curril, I., Dinh, H., Doddapaneni, H., Dolan, A., Dugan, S., Friedrich, M., Gasperi, G., Geib, S., Georgakilas, G., Gibbs, R. A., Giers, S. D., Gomulski, L. M., González-Guzmán, M., Guillem-Amat, A., Han, Y., Hatzigeorgiou, A. G., Hernández-Crespo, P., Hughes, D. S. T., Jones, J. W., Karagkouni, D., Koskinioti, P., Lee, S. L., Malacrida, A. R., Manni, M., Mathiopoulos, K., Meccariello, A., Murali, S. C., Murphy, T. D., Muzny, D. M., Oberhofer, G., Ortego, F., Paraskevopoulou, M. D., Poelchau, M., Qu, J., Reczko, M., Robertson, H. M., Rosendale, A. J., Rosselot, A. E., Saccone, G., Salvemini, M., Savini, G., Schreiner, P., Scolari, F., Siciliano, P., Sim, S. B., Tsiamis, G., Ureña, E., Vlachos, I. S., Werren, J. H., Wimmer, E. A., Worley, K. C., Zacharopoulou, A., Richards, S., & Handler, A. M. (2016). The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biology*, 17, 192.

Paterson, S. (2017). Plodia interpunctella strain Dundee, whole genome shotgun sequencing project. *NCBI nucleotide database*. {:itemType: dataset}.

Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P. A., Heraty, J., Kjer, K. M., Klopfstein, S., Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B., & Niehuis, O. (2017). Evolutionary History of the Hymenoptera. *Current Biology*.

Pohl, H. & Beutel, R. G. (2005). The phylogeny of Strepsiptera (Hexapoda). *Cladistics*, 21(4), 328–374.

Regier, J. C., Mitter, C., Zwick, A., Bazinet, A. L., Cummings, M. P., Kawahara, A. Y., Sohn, J.-C., Zwickl, D. J., Cho, S., Davis, D. R., Baixeras, J., Brown, J., Parr, C., Weller, S., Lees, D. C., & Mitter, K. T. (12-Mar-2013). A Large-Scale, Higher-Level, Molecular Phylogenetic Study of the Insect Order Lepidoptera (Moths and Butterflies). *PLOS ONE*, 8(3), e58568.

Rehan, S. M., Glastad, K. M., Lawson, S. P., & Hunt, B. G. (2016). The genome and methylome of a subsocial small carpenter bee, *Ceratina calcarata*. *Genome Biology and Evolution*, (pp. evw079).

Richter, S., Olesen, J., & Wheeler, W. C. (2007). Phylogeny of Branchiopoda (Crustacea) based on a combined analysis of morphological data and six molecular loci. *Cladistics*, 23(4), 301–336.

Robinson, G., Robertson, H., Hudson, M., Walden, K., Fischman, B., Pitts-Singer, T., James, R., Salzberg, S., Puiu, D., Magoc, T., Kelley, D., & Zimin, A. (2014). Megachile rotundata, whole genome shotgun sequencing project. *NCBI nucleotide database*. {:itemType: dataset}.

Rosenfeld, J. A., Reeves, D., Brugler, M. R., Narechania, A., Simon, S., Durrett, R., Foox, J., Shianna, K., Schatz, M. C., Gandara, J., Afshinnekoo, E., Lam, E. T., Hastie, A. R., Chan, S., Cao, H., Saghbini, M., Kentsis, A., Planet, P. J., Kholodovych, V., Tessler, M., Baker, R., DeSalle, R., Sorkin, L. N., Kolokotronis, S.-O., Siddall, M. E., Amato, G., & Mason, C. E. (2016). Genome assembly and geospatial phylogenomics of the bed bug *Cimex lectularius*. *Nature Communications*, 7, 10164.

Sadd, B. M., Barribeau, S. M., Bloch, G., de Graaf, D. C., Dearden, P., Elsik, C. G., Gadau, J., Grimmelikhuijzen, C. J., Hasselmann, M., Lozier, J. D., Robertson, H. M., Smagghe, G., Stolle, E., Van Vaerenbergh, M., Waterhouse, R. M., Bornberg-Bauer, E., Klasberg, S., Bennett, A. K., Câmara, F., Guigó, R., Hoff, K., Mariotti, M., Munoz-Torres, M., Murphy, T., Santesmasses, D., Amdam, G. V., Beckers, M., Beye, M., Biewer, M., Bitondi, M. M., Blaxter, M. L., Bourke, A. F., Brown, M. J., Buechel, S. D., Cameron, R., Cappelle, K., Carolan, J. C., Christiaens, O., Ciborowski, K. L., Clarke, D. F., Colgan, T. J., Collins, D. H., Cridge, A. G., Dalmay, T., Dreier, S., du Plessis, L., Duncan, E., Erler, S., Evans, J., Falcon, T., Flores, K., Freitas, F. C., Fuchikawa, T., Gempe, T., Hartfelder, K., Hauser, F., Helbing, S., Humann, F. C., Irvine, F., Jermiin, L. S., Johnson, C. E., Johnson, R. M., Jones, A. K., Kadowaki, T., Kidner, J. H., Koch, V., Köhler, A., Kraus, F. B., Lattorff, H. M. G., Leask, M., Lockett, G. A., Mallon, E. B., Antonio, D. S. M., Marxer, M., Meeus, I., Moritz, R. F., Nair, A., Näpflin, K., Nissen, I., Niu, J., Nunes, F. M., Oakeshott, J. G., Osborne, A., Otte, M., Pinheiro, D. G., Rossié, N., Rueppell, O., Santos, C. G., Schmid-Hempel, R., Schmitt, B. D., Schulte, C., Simões, Z. L., Soares, M. P., Swevers, L., Winnebeck, E. C., Wolschin, F., Yu, N., Zdobnov, E. M., Aqrawi, P. K., Blankenburg, K. P., Coyle, M., Francisco, L., Hernandez, A. G., Holder, M., Hudson, M. E., Jackson, L., Jayaseelan, J., Joshi, V., Kovar, C., Lee, S. L., Mata, R., Mathew, T., Newsham, I. F., Ngo, R., Okwuonu, G., Pham, C., Pu, L.-L., Saada, N., Santibanez, J., Simmons, D., Thornton, R., Venkat, A., Walden, K. K., Wu, Y.-Q., Debyser, G., Devreese, B., Asher, C., Blommaert, J., Chipman, A. D., Chittka, L., Fouks, B., Liu, J., O'Neill, M. P., Sumner, S., Puiu, D., Qu, J., Salzberg, S. L., Scherer, S. E., Muzny, D. M., Richards, S., Robinson, G. E., Gibbs, R. A., Schmid-Hempel, P., & Worley, K. C. (2015). The genomes of two key bumblebee species with

primitive eusocial organization. *Genome Biology*, 16, 76.

Schwager, E. E., Sharma, P. P., Clarke, T., Leite, D. J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T., et al. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *bioRxiv*, (pp. 106385).

Scott, J. G., Warren, W. C., Beukeboom, L. W., Bopp, D., Clark, A. G., Giers, S. D., Hediger, M., Jones, A. K., Kasai, S., Leichter, C. A., Li, M., Meisel, R. P., Minx, P., Murphy, T. D., Nelson, D. R., Reid, W. R., Rinkevich, F. D., Robertson, H. M., Sackton, T. B., Sattelle, D. B., Thibaud-Nissen, F., Tomlinson, C., van de Zande, L., Walden, K., Wilson, R. K., & Liu, N. (2014). Genome of the house fly, Musca domestica L., a global vector of diseases with adaptations to a septic environment. *Genome Biology*, 15(10), 466.

Shen, J., Cong, Q., Kinch, L. N., Borek, D., Otwinowski, Z., & Grishin, N. V. (2016). Complete genome of Pieris rapae, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. *F1000Research*, 5, 2631.

Simpson, S. D., Ramsdell, J. S., Iii, W., H, W., & Chabot, C. C. (2017). The Draft Genome and Transcriptome of the Atlantic Horseshoe Crab, Limulus polyphemus. *International Journal of Genomics*.

Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0.

Smith, C. D., Zimin, A., Holt, C., Abouheif, E., Benton, R., Cash, E., Croset, V., Currie, C. R., Elhaik, E., Elsik, C. G., Fave, M.-J., Fernandes, V., Gadau, J., Gibson, J. D., Graur, D., Grubbs, K. J., Hagen, D. E., Helmkampf, M., Holley, J.-A., Hu, H., Viniegra, A. S. I., Johnson, B. R.,

Johnson, R. M., Khila, A., Kim, J. W., Laird, J., Mathis, K. A., Moeller, J. A., Muñoz-Torres, M. C., Murphy, M. C., Nakamura, R., Nigam, S., Overson, R. P., Placek, J. E., Rajakumar, R., Reese, J. T., Robertson, H. M., Smith, C. R., Suarez, A. V., Suen, G., Suhr, E. L., Tao, S., Torres, C. W., van Wilgenburg, E., Viljakainen, L., Walden, K. K. O., Wild, A. L., Yandell, M., Yorke, J. A., & Tsutsui, N. D. (2011a). Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). *Proceedings of the National Academy of Sciences*, 108(14), 5673–5678.

Smith, C. R., Smith, C. D., Robertson, H. M., Helmkampf, M., Zimin, A., Yandell, M., Holt, C., Hu, H., Abouheif, E., Benton, R., Cash, E., Croset, V., Currie, C. R., Elhaik, E., Elsik, C. G., Favé, M.-J., Fernandes, V., Gibson, J. D., Graur, D., Gronenberg, W., Grubbs, K. J., Hagen, D. E., Viniegra, A. S. I., Johnson, B. R., Johnson, R. M., Khila, A., Kim, J. W., Mathis, K. A., Munoz-Torres, M. C., Murphy, M. C., Mustard, J. A., Nakamura, R., Niehuis, O., Nigam, S., Overson, R. P., Placek, J. E., Rajakumar, R., Reese, J. T., Suen, G., Tao, S., Torres, C. W., Tsutsui, N. D., Viljakainen, L., Wolschin, F., & Gadau, J. (2011b). Draft genome of the red harvester ant Pogonomyrmex barbatus. *Proceedings of the National Academy of Sciences*, 108(14), 5667–5672.

Song, N., Liang, A.-P., & Bu, C.-P. (2012). A Molecular Phylogeny of Hemiptera Inferred from Mitochondrial Genome Sequences. *PLoS ONE*, 7(11), e48778.

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., Bouffard, P., Caldera, E. J., Cash, E., Cavanaugh, A., Denas, O., Elhaik, E., Favé, M.-J., Gadau, J., Gibson, J. D., Graur, D., Grubbs, K. J., Hagen, D. E., Harkins, T. T., Helmkampf, M., Hu, H., Johnson,

B. R., Kim, J., Marsh, S. E., Moeller, J. A., Muñoz-Torres, M. C., Murphy, M. C., Naughton, M. C., Nigam, S., Overson, R., Rajakumar, R., Reese, J. T., Scott, J. J., Smith, C. R., Tao, S., Tsutsui, N. D., Viljakainen, L., Wissler, L., Yandell, M. D., Zimmer, F., Taylor, J., Slater, S. C., Clifton, S. W., Warren, W. C., Elsik, C. G., Smith, C. D., Weinstock, G. M., Gerardo, N. M., & Currie, C. R. (2011). The genome gequence of the leaf-cutter ant Atta cephalotes reveals insights into Its obligate symbiotic lifestyle. *PLoS Genet*, *7*(2), e1002007.

Terrapon, N., Li, C., Robertson, H. M., Ji, L., Meng, X., Booth, W., Chen, Z., Childers, C. P., Glastad, K. M., Gokhale, K., Gowin, J., Gronenberg, W., Hermansen, R. A., Hu, H., Hunt, B. G., Huylmans, A. K., Khalil, S. M. S., Mitchell, R. D., Munoz-Torres, M. C., Mustard, J. A., Pan, H., Reese, J. T., Scharf, M. E., Sun, F., Vogel, H., Xiao, J., Yang, W., Yang, Z., Yang, Z., Zhou, J., Zhu, J., Brent, C. S., Elsik, C. G., Goodisman, M. A. D., Liberles, D. A., Roe, R. M., Vargo, E. L., Vilcinskas, A., Wang, J., Bornberg-Bauer, E., Korb, J., Zhang, G., & Liebig, J. (2014). Molecular traces of alternative social organization in a termite genome. *Nature Communications*, *5*.

The Heliconius Genome Consortium, Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., Zimin, A. V., Hughes, D. S. T., Ferguson, L. C., Martin, S. H., Salazar, C., Lewis, J. J., Adler, S., Ahn, S.-J., Baker, D. A., Baxter, S. W., Chamberlain, N. L., Chauhan, R., Counterman, B. A., Dalmay, T., Gilbert, L. E., Gordon, K., Heckel, D. G., Hines, H. M., Hoff, K. J., Holland, P. W. H., Jacquin-Joly, E., Jiggins, F. M., Jones, R. T., Kapan, D. D., Kersey, P., Lamas, G., Lawson, D., Mapleson, D., Maroja, L. S., Martin, A., Moxon, S., Palmer, W. J., Papa, R., Papanicolaou, A., Pauchet, Y., Ray, D. A., Rosser,

N., Salzberg, S. L., Supple, M. A., Surridge, A., Tenger-Trolander, A., Vogel, H., Wilkinson, P. A., Wilson, D., Yorke, J. A., Yuan, F., Balmuth, A. L., Eland, C., Gharbi, K., Thomson, M., Gibbs, R. A., Han, Y., Jayaseelan, J. C., Kovar, C., Mathew, T., Muzny, D. M., Ongeri, F., Pu, L.-L., Qu, J., Thornton, R. L., Worley, K. C., Wu, Y.-Q., Linares, M., Blaxter, M. L., ffrench-Constant, R. H., Joron, M., Kronforst, M. R., Mullen, S. P., Reed, R. D., Scherer, S. E., Richards, S., Mallet, J., McMillan, W. O., & Jiggins, C. D. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405), 94–98.

The International Aphid Genomics Consortium (2010). Genome Sequence of the Pea Aphid Acyrthosiphon pisum. *PLoS Biology*, 8(2), e1000313.

Thum, R. A. (2004). Using 18S rDNA to resolve diaptomid copepod (Copepoda: Calanoida: Diaptomidae) phylogeny: An example with the North American genera. *Hydrobiologia*, 519(1-3), 135–141.

Tribolium Genome Sequencing Consortium (2008). The genome of the model beetle and pest Tribolium castaneum. *Nature*, 452(7190), 949–955.

Tsang, L., Ma, K., Ahyong, S., Chan, T.-Y., & Chu, K. (2008). Phylogeny of Decapoda using two nuclear protein-coding genes: Origin and evolution of the Reptantia. *Molecular Phylogenetics and Evolution*, 48(1), 359–368.

Vega, F. E., Brown, S. M., Chen, H., Shen, E., Nair, M. B., Ceja-Navarro, J. A., Brodie, E. L., Infante, F., Dowd, P. F., & Pain, A. (2015). Draft genome of the most devastating insect pest of coffee worldwide: The coffee berry borer, Hypothenemus hampei. *Scientific Reports*, 5, 12525.

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., Hao, S., Chen, B., Ma, Z., Yu, D., Xiong, Z., Zhu, Y., Fan, D., Han, L., Wang, B., Chen, Y., Wang, J., Yang, L., Zhao, W., Feng, Y., Chen, G., Lian, J., Li, Q., Huang, Z., Yao, X., Lv, N., Zhang, G., Li, Y., Wang, J., Wang, J., Zhu, B., & Kang, L. (2014). The Locust Genome Provides Insight into Swarm Formation and Long-Distance Flight. *Nature Communications*, 5.

Wang, Y.-H., Wu, H.-Y., Rédei, D., Xie, Q., Chen, Y., Chen, P.-P., Dong, Z.-E., Dang, K., Damgaard, J., Štys, P., Wu, Y.-Z., Luo, J.-Y., Sun, X.-Y., Hartung, V., Kuechler, S. M., Liu, Y., Liu, H.-X., & Bu, W.-J. (2017). When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera-Heteroptera. *Cladistics*.

Ward, P. S., Brady, S. G., Fisher, B. L., & Schultz, T. R. (2015). The evolution of myrmicine ants: Phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae): Phylogeny and evolution of myrmicine ants. *Systematic Entomology*, 40(1), 61–81.

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., The Nasonia Genome Working Group, Beukeboom, L. W., Desplan, C., Elsik, C. G., Grimmelikhuijzen, C. J. P., Kitts, P., Lynch, J. A., Murphy, T., Oliveira, D. C. S. G., Smith, C. D., v. d. Zande, L., Worley, K. C., Zdobnov, E. M., Aerts, M., Albert, S., Anaya, V. H., Anzola, J. M., Barchuk, A. R., Behura, S. K., Bera, A. N., Berenbaum, M. R., Bertossa, R. C., Bitondi, M. M. G., Bordenstein, S. R., Bork, P., Bornberg-Bauer, E., Brunain, M., Cazzamali, G., Chaboub, L., Chacko, J., Chavez, D., Childers, C. P., Choi, J.-H., Clark, M. E., Claudianos, C., Clinton, R. A., Cree, A. G., Cristino, A. S., Dang, P. M., Darby, A. C., de Graaf, D. C., Devreese, B.,

Dinh, H. H., Edwards, R., Elango, N., Elhaik, E., Ermolaeva, O., Evans, J. D., Foret, S., Fowler, G. R., Gerlach, D., Gibson, J. D., Gilbert, D. G., Graur, D., Grunder, S., Hagen, D. E., Han, Y., Hauser, F., Hultmark, D., Hunter, H. C., Hurst, G. D. D., Jhangian, S. N., Jiang, H., Johnson, R. M., Jones, A. K., Junier, T., Kadowaki, T., Kamping, A., Kapustin, Y., Kechavarzi, B., Kim, J., Kim, J., Kiryutin, B., Koevoets, T., Kovar, C. L., Kriventseva, E. V., Kucharski, R., Lee, H., Lee, S. L., Lees, K., Lewis, L. R., Loehlin, D. W., Logsdon, J. M., Lopez, J. A., Lozado, R. J., Maglott, D., Maleszka, R., Mayampurath, A., Mazur, D. J., McClure, M. A., Moore, A. D., Morgan, M. B., Muller, J., Munoz-Torres, M. C., Muzny, D. M., Nazareth, L. V., Neupert, S., Nguyen, N. B., Nunes, F. M. F., Oakeshott, J. G., Okwuonu, G. O., Pannebakker, B. A., Pejaver, V. R., Peng, Z., Pratt, S. C., Predel, R., Pu, L.-L., Ranson, H., Raychoudhury, R., Rechtsteiner, A., Reid, J. G., Riddle, M., Romero-Severson, J., Rosenberg, M., Sackton, T. B., Sattelle, D. B., Schluns, H., Schmitt, T., Schneider, M., Schuler, A., Schurko, A. M., Shuker, D. M., Simoes, Z. L. P., Sinha, S., Smith, Z., Souvorov, A., Springauf, A., Stafflinger, E., Stage, D. E., Stanke, M., Tanaka, Y., Telschow, A., Trent, C., Vattathil, S., Viljakainen, L., Wanner, K. W., Waterhouse, R. M., Whitfield, J. B., Wilkes, T. E., Williamson, M., Willis, J. H., Wolschin, F., Wyder, S., Yamada, T., Yi, S. V., Zecher, C. N., Zhang, L., & Gibbs, R. A. (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, 327(5963), 343–348.

Wiegmann, B. M., Trautwein, M. D., Winkler, I. S., Barr, N. B., Kim, J.-W., Lambkin, C., Bertone, M. A., Cassel, B. K., Bayless, K. M., Heimberg, A. M., Wheeler, B. M., Peterson, K. J., Pape, T., Sinclair, B. J., Skevington, J. H., Blagoderov, V., Caravas, J., Kutty, S. N., Schmidt-Ott, U., Kampmeier, G. E., Thompson, F. C., Grimaldi, D. A., Beckenbach, A. T., Courtney,

G. W., Friedrich, M., Meier, R., & Yeates, D. K. (2011). Episodic radiations in the fly tree of life. *Proceedings of the National Academy of Sciences*, 108(14), 5690–5695.

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B. G., Ingram, K. K., Falquet, L., Nipitwattanaphon, M., Gotzek, D., Dijkstra, M. B., Oettler, J., Comtesse, F., Shih, C.-J., Wu, W.-J., Yang, C.-C., Thomas, J., Beaudoing, E., Pradervand, S., Flegel, V., Cook, E. D., Fabbretti, R., Stockinger, H., Long, L., Farmerie, W. G., Oakey, J., Boomsma, J. J., Pamilo, P., Yi, S. V., Heinze, J., Goodisman, M. A. D., Farinelli, L., Harshman, K., Hulo, N., Cerutti, L., Xenarios, I., Shoemaker, D., & Keller, L. (2011). The genome of the fire ant Solenopsis invicta. *Proceedings of the National Academy of Sciences*, 108(14), 5679–5684.

Xue, J., Zhou, X., Zhang, C.-X., Yu, L.-L., Fan, H.-W., Wang, Z., Xu, H.-J., Xi, Y., Zhu, Z.-R., Zhou, W.-W., Pan, P.-L., Li, B.-L., Colbourne, J. K., Noda, H., Suetsugu, Y., Kobayashi, T., Zheng, Y., Liu, S., Zhang, R., Liu, Y., Luo, Y.-D., Fang, D.-M., Chen, Y., Zhan, D.-L., Lv, X.-D., Cai, Y., Wang, Z.-B., Huang, H.-J., Cheng, R.-L., Zhang, X.-C., Lou, Y.-H., Yu, B., Zhuo, J.-C., Ye, Y.-X., Zhang, W.-Q., Shen, Z.-C., Yang, H.-M., Wang, J., Wang, J., Bao, Y.-Y., & Cheng, J.-A. (2014). Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation. *Genome Biology*, 15, 521.

Yin, C., Liu, Y., Liu, J., Xiao, H., Huang, S., Lin, Y., Han, Z., & Li, F. (2014). ChiloDB: A genomic and transcriptome database for an important rice insect pest Chilo suppressalis. *Database*, 2014.

You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., Zhan, D., Baxter, S. W., Vasseur, L., Gurr, G. M., Douglas, C. J., Bai, J., Wang, P., Cui, K., Huang, S., Li, X., Zhou, Q., Wu, Z.,

Chen, Q., Liu, C., Wang, B., Li, X., Xu, X., Lu, C., Hu, M., Davey, J. W., Smith, S. M., Chen, M., Xia, X., Tang, W., Ke, F., Zheng, D., Hu, Y., Song, F., You, Y., Ma, X., Peng, L., Zheng, Y., Liang, Y., Chen, Y., Yu, L., Zhang, Y., Liu, Y., Li, G., Fang, L., Li, J., Zhou, X., Luo, Y., Gou, C., Wang, J., Wang, J., Yang, H., & Wang, J. (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics*, 45(2), 220–225.

Zhan, S., Merlin, C., Boore, J. L., & Reppert, S. M. (2011). The Monarch Butterfly Genome Yields Insights into Long-Distance Migration. *Cell*, 147(5), 1171–1185. Danaus Plexippus.

Zhang, H.-L., Huang, Y., Lin, L.-L., Wang, X.-Y., & Zheng, Z.-M. (2013). The Phylogeny of the Orthoptera (Insecta) as Deduced from Mitogenomic Gene Sequences. *Zoological Studies*, 52, 37.

Zhao, C., Escalante, L. N., Chen, H., Benatti, T. R., Qu, J., Chellapilla, S., Waterhouse, R. M., Wheeler, D., Andersson, M. N., Bao, R., Batterton, M., Behura, S. K., Blankenburg, K. P., Caragea, D., Carolan, J. C., Coyle, M., El-Bouhssini, M., Francisco, L., Friedrich, M., Gill, N., Grace, T., Grimmelikhuijzen, C. J. P., Han, Y., Hauser, F., Herndon, N., Holder, M., Ioannidis, P., Jackson, L., Javaid, M., Jhangiani, S. N., Johnson, A. J., Kalra, D., Korchina, V., Kovar, C. L., Lara, F., Lee, S. L., Liu, X., Löfstedt, C., Mata, R., Mathew, T., Muzny, D. M., Nagar, S., Nazareth, L. V., Okwuonu, G., Ongeri, F., Perales, L., Peterson, B. F., Pu, L.-L., Robertson, H. M., Schemerhorn, B. J., Scherer, S. E., Shreve, J. T., Simmons, D., Subramanyam, S., Thornton, R. L., Xue, K., Weissenberger, G. M., Williams, C. E., Worley, K. C., Zhu, D., Zhu, Y., Harris, M. O., Shukle, R. H., Werren, J. H., Zdobnov, E. M., Chen, M.-S., Brown, S. J.,

Stuart, J. J., & Richards, S. (2015). A Massive Expansion of Effector Genes Underlies Gall-Formation in the Wheat Pest Mayetiola destructor. *Current Biology*, 25(5), 613–620.

E

Supplemental material to chapter 4

Supplemental Figures:

Supplemental Tables:

E.1 Supplemental Methods

E.1.1 Apoid wasp transcriptomes

We *de novo* sequenced whole body transcript libraries of 24 apoid wasp species in the context of the international 1KITE project (Table E.1). Adult wasps were collected via hand-netting and immediately preserved in RNAlater. RNA extraction, cDNA synthesis, and sequencing library preparation followed the methodology outlined by (Misof et al., 2014). Briefly, RNA was extracted using a standard phenol/guanidine isothiocyanate-based extraction method and tested for quality before processing for library construction. cDNA libraries were constructed by shearing and amplification of mRNA that was isolated using magnetic beads. A random hexamer primer was added and the double-stranded cDNA then underwent end-repair, a single 'A' base addition and adapter ligation. Library size selection was performed by gel electrophoresis and excision of the 250 ± 20 bp band. The product was indexed and PCR amplified to obtain paired-end cDNA. After cDNA fragment size verification, the cDNA libraries were sequenced on an Illumina HiSeq2000 platform following standard protocols. For each library, roughly

2.5 Gbp of raw data was sequenced with 150 bp paired-end reads. After filtering steps to ensure high quality raw data libraries, transcripts were assembled with SOAPdenovo-trans-31kmer v1.01 (Xie et al., 2014) with moderately strict parameters (`-e3`). The assembled transcript libraries were finally screened for vector and adapter contamination using a local VecScreen installation (`http://www.ncbi.nlm.nih.gov/tools/vecscreen`) and the UniVec database build 7.0 (`http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec`). Assembled transcripts at least 200 bp in length were checked for cross-contamination with reads from samples sequenced on the same Illumina lane. In brief, BLAST hits with lengths > 179 and identity of at least 98% were compared for their *k*-mer coverage values (as computed during assembly with SOAPdenovo-trans). From a cluster of highly similar sequences only the sequence with highest k-mer coverage was kept, and only if the coverage was at least 2x higher than the second best, otherwise all were discarded. The remaining contigs were submitted to the NCBI Transcriptome Shotgun Assembly (TSA) database, where they were again screened for potential contaminants from vector nucleotide sequences as well as for sequences that might originate from non-target species contamination.

Sequencing data were deposited at the Sequence Read Archive (SRA) and the Transcriptome Shotgun Assembly (TSA) database of NCBI GenBank (accession numbers see Additional file 2) and are available at NCBI via the Umbrella BioProject ID PRJNA183205 ("The 1KITE project: evolution of insects").

E.1.2 Orthograph

Orthograph dependencies

Orthograph requires the software packages HMMER3 (Eddy, 2011), NCBI BLAST+ (Camacho et al., 2009), MAFFT (Katoh & Standley, 2013), and Exonerate (Slater & Birney, 2005) as well as either MySQL or SQLite (for specific version see Table E.2).

Ortholog reference set

The user must provide Orthograph with a set of reference OGs to which transcripts are mapped. For this purpose, Orthograph requires the amino acid and corresponding nucleotide sequences of all protein-coding genes in the user-selected reference OGS. It additionally needs information about which genes in these genomes are orthologous. Information on orthology relations of genes in the reference genomes (*i.e.*, what genes form OGs) can be obtained from databases such as OrthoDB (`http://orthodb.org`), InParanoid (`http://inparanoid.sbc.su.se`), OrthoMCL DB (`http://orthomcl.org`), and OMA (`http://omabrowser.org`). Alternatively, a reference ortholog set has to be inferred using an orthology prediction approach that works on fully sequenced genomes, such as the respective tools for the databases (OrthoDB (Kriventseva et al., 2015), OrthoMCL (Li et al., 2003), InParanoid (Sonnhammer & Östlund, 2015), OMA (Altenhoff et al., 2015)) or the pipeline OrthoFinder (Emms & Kelly, 2015). To construct a reference set of OGs for identifying orthologous *de novo*-sequenced transcripts of 24 apoid wasps (see below), we exploited OrthoDB 5 (Waterhouse et al., 2011), a database that delineates orthologs among published genomes using a graph-based clustering strategy. For the

analysis of the 24 *de novo*-sequenced transcript libraries of apoid wasps, we used a reference ortholog set that contains OGs from six species of Hymenoptera: *Acromyrmex echinatior*, *Apis mellifera*, *Camponotus floridanus*, *Harpegnathos saltator*, *Linepithema humile*, and *Nasonia vitripennis*. The OGS versions and download URLs are listed in Table E.3. These taxa were selected because a) their genomes are well sequenced, fully annotated, published, and publicly available, and b) they represent major lineages of Hymenoptera comparatively closely related to apoid wasps. The hierarchical level for clustering orthologous genes in the OrthoDB query was set to the node Apocrita (*N. vitripennis*/rest of Hymenoptera). We requested genes in the above six reference species to be always present in single copy. Given these settings, OrthoDB 5 identified 5,561 OGs fulfilling these criteria. The resulting OrthoDB table was subsequently filtered to only contain information about the selected taxa. Since Orthograph needs to relate identifiers in the OrthoDB table to sequences in the OGS, headers in the OGS files were modified so that they match the header naming scheme in the OrthoDB table. For testing the functionality and performance of Orthograph, we used a different set (see below).

Scan for candidate transcripts using profile hidden Markov models

Orthograph creates a multiple sequence alignment (MSA) from the individual amino acid sequences that are part of a given OG using MAFFT L-INS-I (Katoh & Standley, 2013). From each of the resulting MSAs, Orthograph constructs a profile hidden Markov model (pHMM) using HMMER3 with default parameters, resulting in one pHMM per OG. Orthograph uses these pHMMs to search the transcript library (or any other pool of coding sequences that can also include short non-coding sequence sections, such as introns) for candidate orthologs on

amino acid level in all six possible reading frames. Orthograph allows the user to specify an alternative genetic code translation table when dealing with species that use a different genetic code. All search results are stored in a relational database for later evaluation; note that no orthology delineation is performed at this point. As relational database management system, the user can chose between MySQL and SQLite. The first is a reasonable choice when running in a network environment with one computer acting as a database server; the latter when running Orthograph on a HPC cluster.

Establishing BRH criterion using BLAST+

A BLAST database is generated from all amino acid sequences of all reference proteomes. Orthograph uses the predicted amino acid sequence section of a candidate transcript (or other coding sequences) that returned a match during the pHMM search as query for a search against the above reference proteome database using protein BLAST of the NCBI BLAST+ program suite. All retrieved search results are subsequently stored in the database for later evaluation. Note that in contrast to the algorithm in HaMStR, Orthograph attempts no orthology delineation at this point.

Extension of clusters of orthologous genes

Orthograph retrieves the results from all pHMM searches from the database sorted by descending alignment bit score. Sorting by bit score increases the likelihood of retrieving the biologically most relevant hit by using sequence similarity as a criterion for putative sequence homology. For each candidate transcript, the search results are tested for reciprocity: if the subsequent re-

verse BLAST search using the candidate transcript as query matches a target sequence from the OGS that is part of the OG that formed the basis for this particular pHMM, the BRH criterion is fulfilled. In this case, an ortholog relationship between the target transcript section and the OG is assumed and the target transcript section is assigned to the OG unless it overlaps with a previous assignment. If it overlaps with a previous assignment, *i.e.* two different sequences fulfilling the BRH criterion on overlapping regions of the OG, a paralogous relationship is assumed and the transcript section is recorded accordingly. To avoid protein domain walking, Orthograph does not consider transcript sections of fewer than 30 amino acids in length for forther processing. This cutoff can be changed by the user, if necessary.

Frameshift-corrected ORF inference

To infer ORFs and to correct for frameshift errors, which may be present in NGS products, Orthograph employs the alignment program Exonerate (Slater & Birney, 2005). It is used to compute a pairwise alignment of the amino acid sequence of the most similar reference taxon and the orthologous transcript section on nucleotide level to infer the corresponding coding DNA sequence. As a result, Orthograph provides corresponding amino acid and nucleotide sequences for the orthologous transcripts. Orthograph can extend the ORF beyond the pHMM alignment coordinates by inferring ORFs from the entire transcript sequence. More than 50% (default value that can be changed by the user) of the resulting ORF must be part of the sequence region for which orthology has been inferred (Figure E.1). This is done to obtain a longer ORF while retaining orthology information for the majority of its length.

E.1.3 Reanalysis of publicly available data

Sensitivity and accuracy when searching for single-copy orthologs

From the OrthoDB 7 database (Waterhouse et al., 2013), a set of OGs was obtained for four species of Hymenoptera and an outgroup beetle. The hierarchical level was set to the split (Hymenoptera/rest of Holometabola) and we requested that genes in *A. mellifera*, *C. floridanus*, *H. saltator*, *N. vitripennis*, and *T. castaneum* occur in single-copy, while copy number in all other taxa was left unspecified. This query returned 4,625 OGs. The resulting table was filtered to contain only entries from the above five species. This table was re-filtered twice to obtain two different ortholog sets: one that was missing entries from *A. mellifera*, and one that excluded entries from *H. saltator*. Note that we included only the longest isoform per gene from the OGS libraries, irrespective of the species. The sets were imported into Orthograph. We ran the analysis using default parameters. Evaluation was performed using custom-made Bash scripts.

Identification of splice variants or isoforms

To identify splice variants or isoforms, we used the ortholog set derived from five reference species with 4,625 OGs from the analysis for testing Orthograph performance when searching for single-copy orthologs. We included sequences from all five species in the set. Additionally, we downloaded the *C. floridanus* OGS transcripts from the Hymenoptera Genome Database (Munoz-Torres et al., 2011). The sequence headers were reformatted to match the format used in the OrthoDB table. The ortholog set was imported in the Orthograph database, and we ran the analysis using default parameters. The results were evaluated using custom-made Bash scripts.

Identification of inparalogs

We complemented the ortholog set from the analysis for testing Orthograph sensitivity and accuracy when searching for single-copy orthologs with amino acid sequences from *A. cephalotes* by a modified query to OrthoDB 7, demanding presence, but without copy-number restriction for genes from *A. cephalotes*. We obtained the OGS of *A. cephalotes*, version 1.2, from `http://www.hymenopteragenome.org/atta/?q=genome_consortium_datasets` (Suen et al., 2011). Phylogenetic split as well as copy-number restrictions for the other taxa were kept as described above. This query returned 301 OGs. The resulting table was filtered to contain only entries from the six selected species *A. mellifera*, *C. floridanus*, *H. saltator*, *N. vitripennis*, *T. castaneum*, and *A. cephalotes*. The set was imported into the Orthograph database, and we ran the analysis using default parameters. The results were evaluated using custom-made Bash and Perl scripts.

E.1.4 Non-redundant mapping of transcripts

The dataset from Struck *et al.* Struck et al. (2014) was obtained from the Dryad database (`http://datadryad.org/bitstream/handle/10255/dryad.62820/Struck_Platyzoa2014.tgz`). The ortholog set used by Struck *et al.* Struck et al. (2014) was obtained from the HaMStR website at `http://deep-phylogeny.org/hamstr/download/datasets/hmmer3/lophotrochozoa_hmmer3.tar.gz`. The reference OGSs provided in the online material from *Struck et al.* Struck et al. (2014) were reformatted and imported into Orthograph. We ran the analysis with para-

meters that closely resemble the settings in HaMStR used by *Struck et al.* Struck et al. (2014). Evaluation of the annotation result was performed using custom-made Bash and Perl scripts.

E.1.5 Computational performance

We tested the computational performance of Orthograph by running analyses on a workstation computer with an Intel Core i7 quad-core processor (3.4 GHz) and 8 GB of RAM. We used the same set of 5,561 single-copy orthologs that was used by Mayer et al. Mayer et al. (2016). For testing the multi-threaded performance, we used a HPC machine with two 6-core Intel Xeon processors (2.67 GHz) capable of running 24 parallel threads total. Orthograph was run on a medium-sized transcriptome of the 24 apoid wasp transcriptomes (*Chalybion californicum*, with 34 Mbp) with default settings and using 1 to 16 parallel threads.

E.2 Supplemental Figures

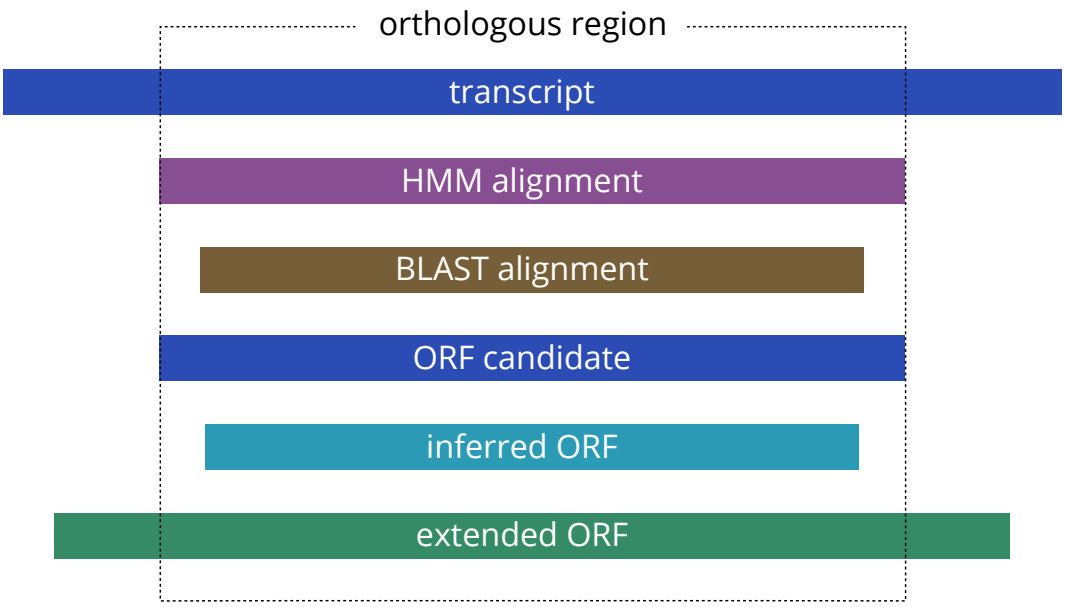

Figure E.1: Alignment regions in Orthograph. On the transcript, there is a candidate ortholog region that was identified using a HMM alignment. The reverse search result using BLAST confirms orthology for the candidate region. For ORF inference, the transcript subsequence that was identified as putatively orthologous using the HMM search is used. The resulting ORF may then be extended by using the entire transcript sequence, resulting in ORF coordinates that exceed the orthologous region.

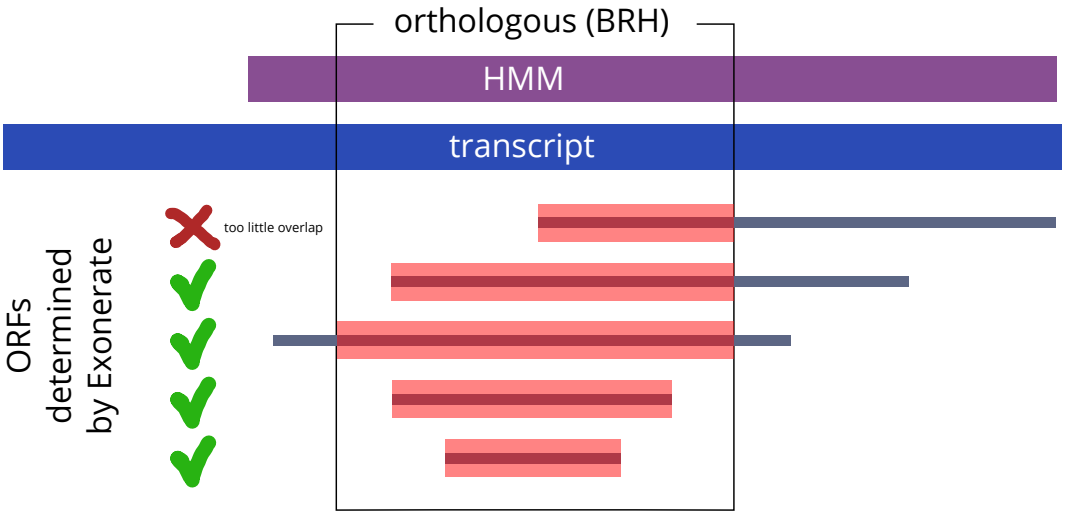

Figure E.2: ORF extension criteria in Orthograph. Inferred ORFs that do not overlap at least 50 % of the orthologous region are discarded due to insufficient confidence in orthology status. As long as the majority of the ORF length is inside the orthologous region on the transcript, ORFs are accepted.

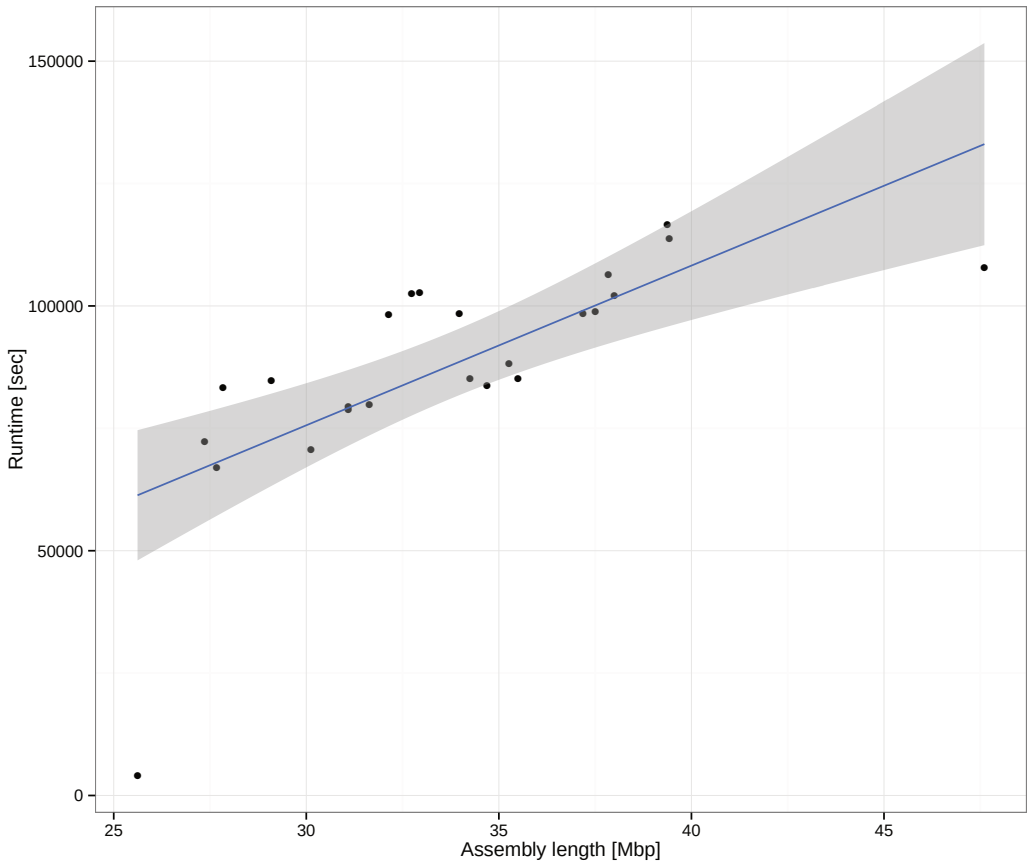

Figure E.3: Orthograph runtime is significantly correlated to total transcriptome assembly length (Spearman rank correlation, $S = 326, p \ll 0.001$) when running with a single thread. Dots indicate measurements for individual transcriptome assemblies. Blue line: linear regression model; gray area: confidence interval.

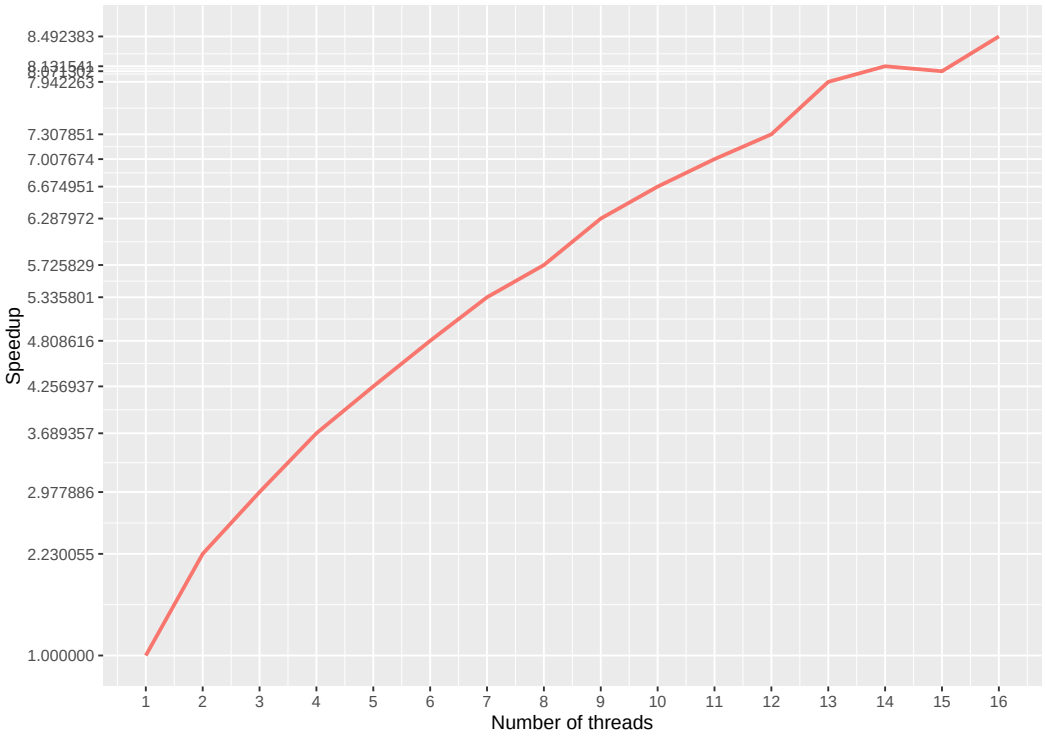

Figure E.4: Orthograph profits from multiple CPU threads. The x axis shows the number of CPU threads; the y axis shows the relative speedup compared to single-threaded performance on a transcriptome assembly of 34 Mbp. Using 16 threads reduces Orthograph runtime to 11.7 % of single-threaded runtime.

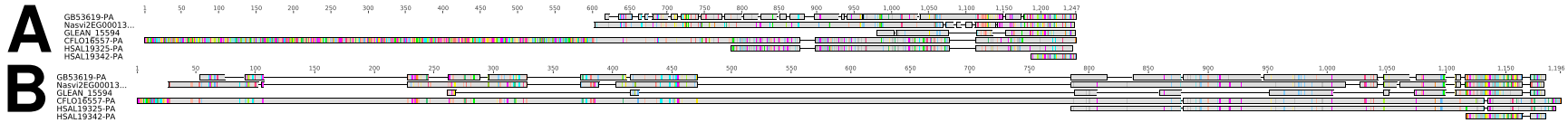

Figure E.5: Multiple sequence alignment (MSA) of an ortholog group (OG) as an examplary assignment of a gene from the *H. saltator* reference gene set (RGS) to the "wrong" OG. A: Alignment using the ClustalW algorithm (Thompson et al., 1994); B: Alignment using the MUSCLE algorithm (Edgar, 2004). According to OrthoDB, the protein HSAL19342-PA belongs to the OG with the ID EOG7KHN7Q. Orthograph, however, identified the protein HSAL19325-PA as orthologous to this OG due to a high similarity to one of the proteins in the OG (404 amino acids alignment overlap, 64.6 % identical sites). The sequence HSAL19325-PA has been added to the MSA to demonstrate that it is in large parts more similar to a sequence from *C. floridanus* (CFLO16557-PA) and therefore yields a higher alignment bit score than the correct – according to OrthoDB – ortholog HSAL19342-PA. In contrast, the protein that is recorded in OrthoDB as part of the OG is shorter and displays little similarity (61 amino acids alignment overlap, 31.1 % identical sites). This leads to a higher bit score in the reverse search for the longer and more similar – but not orthologous according to OrthoDB – sequence. In turn, the BRH criterion for the correct, but shorter ortholog (according to OrthoDB) was not fulfilled. This demonstrates that using different alignment algorithms can impede successful orthology assignment. Grey areas indicate conserved regions, colored bars indicate sequence-specific different amino acid positions. Graphic created using Geneious v7.1 (`http://www.geneious.com`).

E.3 Supplemental Tables

Table E.1: Species for which 1KITE transcriptomes were analyzed.

Order	Family	Subfamily	Genus	Species
Hymenoptera	Crabronidae	Bembicinae	Alyssontini	*Alysson spinosus*
Hymenoptera	Crabronidae	Bembicinae	Bembicini	*Bembix rostrata*
Hymenoptera	Crabronidae	Bembicinae	Bembicini	*Gorytes laticinctus*
Hymenoptera	Crabronidae	Bembicinae	Bembicini	*Harpactus elegans*
Hymenoptera	Crabronidae	Bembicinae	Bembicini	*Sphecius convallis*
Hymenoptera	Crabronidae	Bembicinae	Bembicini	*Stizoides tridentatus*
Hymenoptera	Crabronidae	Bembicinae	Nyssonini	*Nysson niger*
Hymenoptera	Crabronidae	Crabroninae	Crabronini	*Crabro peltarius*
Hymenoptera	Crabronidae	Crabroninae	Crabronini	*Crossocerus quadrimaculatus*
Hymenoptera	Crabronidae	Crabroninae	Larrini	*Tachysphex fulvitarsis*
Hymenoptera	Crabronidae	Crabroninae	Oxybelini	*Oxybelus bipunctatus*
Hymenoptera	Crabronidae	Crabroninae	Trypoxylini	*Trypoxylon figulus*
Hymenoptera	Crabronidae	Dinetinae	-	*Dinetus pictus*
Hymenoptera	Crabronidae	Pemphredoninae	Pemphredonini	*Diodontus minutus*
Hymenoptera	Crabronidae	Pemphredoninae	Pemphredonini	*Pemphredon lugens*
Hymenoptera	Crabronidae	Pemphredoninae	Psenini	*Psenulus fuscipennis*
Hymenoptera	Crabronidae	Philanthinae	Cercerini	*Cerceris arenaria*
Hymenoptera	Crabronidae	Philanthinae	Philanthini	*Philanthus triangulum*
Hymenoptera	Sphecidae	Ammophilinae	-	*Podalonia hirsuta*
Hymenoptera	Sphecidae	Sceliphrinae	Sceliphrini	*Chalybion californicum*
Hymenoptera	Sphecidae	Sceliphrinae	Sceliphrini	*Sceliphron curvatum*
Hymenoptera	Sphecidae	Sphecinae	Prionychini	*Prionyx kirbii*
Hymenoptera	Sphecidae	Sphecinae	Sphecini	*Isodontia mexicana*
Hymenoptera	Sphecidae	Sphecinae	Sphecini	*Sphex funerarius*

Table E.2: Software packages required by Orthograph. It has been developed and tested with these versions. Older versions are not supported.

Package	Version	Download from
Perl	5.14	http://www.perl.org
SQLite	3.8.2	http://sqlite.org/download.html
MySQL	5.6.17	http://dev.mysql.com/downloads/mysql/
MAFFT	7.023b	http://mafft.cbrc.jp/alignment/software/
HMMer	3.1b1	http://hmmer.janelia.org/software/
NCBI BLAST+	2.2.28+	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Exonerate	2.2.0	http://www.ebi.ac.uk/~guy/exonerate/

Table E.3: Official gene sets for the reference ortholog set generation.

Species	Version	Citation
Acromyrmex echinatior	1.2	Nygaard et al. (2011)
Apis mellifera	1.1	Honeybee Genome Sequencing Consortium (2006)
Atta cephalotes	1.2	Suen et al. (2011)
Camponotus floridanus	3.3	Bonasio et al. (2010)
Harpegnathos saltator	3.3	Bonasio et al. (2010)
Linepithema humile	1.2	Smith et al. (2011)
Nasonia vitripennis	1.2	Werren et al. (2010)
Tribolium castaneum	3.0	Tribolium Genome Sequencing Consortium (2008)

Acromyrmex echinatior: `http://hymenopteragenome.org/acromyrmex/?q=genome_consortium_datasets`
Apis mellifera : `http://hymenopteragenome.org/beebase/?q=download_sequence`
Atta cephalotes : `http://hymenopteragenome.org/atta/?q=genome_consortium_datasets`
Camponotus floridanus: `http://hymenopteragenome.org/camponotus/?q=genome_consortium_datasets`
Harpegnathos saltator: `http://hymenopteragenome.org/harpegnathos/?q=genome_consortium_datasets`
Linepithema humile : `http://hymenopteragenome.org/linepithema/?q=genome_consortium_datasets`
Nasonia vitripennis : `http://hymenopteragenome.org/nasonia/?q=sequencing_and_analysis_consortium_datasets`
Tribolium castaneum : `http://beetlebase.org/?q=download_settings`

Table E.4: Species, 1KITE library IDs (see http:// 1 kite.org/ 1 kite_species.php), number of assembled transcripts, total assembly size, N50 values, and NCBI GenBank accession numbers. Note that the assemblies were filtered to contain only contigs longer than 199 bp.

Species	1KITE library ID	Tax. ID	BioProject	BioSample accession	Sample acc.	Exp. accession
Alysson spinosus	INSytvTBDRAAPEI-9	1507100	252289	SAMN02870203	SRS651858	SRX642976
Bembix rostrata	INSswpTBNRAAPEI-44	1507104	252270	SAMN02870220	SRS651839	SRX642957
Cerceris arenaria	INSytvTBFRAAPEI-12	1507109	252291	SAMN02870235	SRS651861	SRX642978
Chalybion californicum	INSytvTBQRAAPEI-57	411700	252298	SAMN02870236	SRS651868	SRX642985
Crabro peltarius	INSswpTBJRAAPEI-37	1507127	252268	SAMN02870270	SRS651838	SRX642955
Crossocerus quadrimaculatus	INSswpTBPRAAPEI-46	1126388	252271	SAMN02870271	SRS651841	SRX642958
Dinetus pictus	INSjdsTAYRAAPEI-43	1507342	252320	SAMN02870280	SRS651890	SRX643007
Diodontus minutus	INSjdsTBMRAAPEI-88	1294192	252322	SAMN02870281	SRS651892	SRX643009
Gorytes laticinctus	INSytvTBERAAPEI-11	1126390	252290	SAMN02870305	SRS651860	SRX642977
Harpactus elegans	INSswpTAFRAAPEI-16	1507137	252247	SAMN02870308	SRS651818	SRX642935
Isodontia mexicana	INSswpTBDRAAPEI-30	288402	252264	SAMN02870321	SRS651834	SRX642951
Nysson niger	INSswpTBGRAAPEI-34	1507151	252266	SAMN02870351	SRS651836	SRX642953
Oxybelus bipunctatus	INSjdsTBIRAAPEI-75	1507154	252321	SAMN02870362	SRS651891	SRX643008
Pemphredon lugens	INSytvTBBRAAPEI-95	1507158	252288	SAMN02870371	SRS651859	SRX642975
Philanthus triangulum	INSswpTBTRABPEI-62	280486	252273	SAMN02870374	SRS651843	SRX642960
Podalonia hirsuta	INSswpTBRRAAPEI-56	1088627	252272	SAMN02870381	SRS651842	SRX642959
Prionyx kirbii	INSytvTBSRAAPEI-74	330847	252299	SAMN02870385	SRS651869	SRX642986
Psenulus fuscipennis	INSswpTATRAAPEI-13	1507163	252256	SAMN02870386	SRS651827	SRX642944
Sceliphron curvatum	INSswpTAZRAAPEI-19	1507168	252261	SAMN02870396	SRS651832	SRX642949
Sphecius convallis	INSnfrTBORAAPEI-14	420963	252349	SAMN02870401	SRS651919	SRX643036
Sphex funerarius	INSytvTAIRAAPEI-18	1507169	252279	SAMN02870403	SRS651849	SRX642966
Stizoides tridentatus	INSytvTARRAAPEI-44	1507174	252284	SAMN02870412	SRS651854	SRX642971
Tachysphex fulvitarsis	INSswpTAKRAAPEI-21	1507176	252251	SAMN02870419	SRS651822	SRX642939
Trypoxylon figulus	INSytvTAWRAAPEI-88	1124897	252286	SAMN02870436	SRS651856	SRX642973

Species	Run accession	TSA project accession	TSA version	Transcripts	Total length	N50
Alysson spinosus	SRR1503092	GBUA00000000	GBUA01000000	40,680	47,606,733	1,568
Bembix rostrata	SRR1503073	GBQR00000000	GBQR01000000	33,341	37,839,804	6,031
Cerceris arenaria	SRR1503094	GBNS00000000	GBNS01000000	24,719	34,252,864	3,305
Chalybion californicum	SRR1503101	GBOM00000000	GBOM01000000	21,323	33,977,878	3,834
Crabro peltarius	SRR1503071	GBWG00000000	GBWG01000000	17,826	27,839,732	4,932
Crossocerus quadrimaculatus	SRR1503074	GBWH00000000	GBWH01000000	16,354	27,670,170	5,280
Dinetus pictus	SRR1503123	GBLS00000000	GBLS01000000	20,195	35,261,360	5,479
Diodontus minutus	SRR1503125	GBMA00000000	GBMA01000000	22,820	39,373,028	3,107
Gorytes laticinctus	SRR1503093	GBNR00000000	GBNR01000000	20,336	30,119,789	7,540
Harpactus elegans	SRR1503051	GBNF00000000	GBNF01000000	22,245	35,499,888	4,814
Isodontia mexicana	SRR1503067	GBPY00000000	GBPY01000000	34,622	38,000,489	2,600
Nysson niger	SRR1503069	GBNN00000000	GBNN01000000	22,496	29,091,955	4,151
Oxybelus bipunctatus	SRR1503124	GBLU00000000	GBLU01000000	22,233	37,187,137	2,311
Pemphredon lugens	SRR1503091	GBQH00000000	GBQH01000000	24,675	39,425,911	716
Philanthus triangulum	SRR1503076	GBWI00000000	GBWI01000000	21,735	27,360,360	4,209
Podalonia hirsuta	SRR1503075	GBPX00000000	GBPX01000000	21,108	32,136,789	4,075
Prionyx kirbii	SRR1503102	GBQI00000000	GBQI01000000	21,703	31,095,319	8,540
Psenulus fuscipennis	SRR1503060	GBNH00000000	GBNH01000000	24,423	31,095,907	1,024
Sceliphron curvatum	SRR1503065	GBNL00000000	GBNL01000000	22,934	32,739,311	6,440
Sphecius convallis	SRR1503152	GBOB00000000	GBOB01000000	19,967	25,618,375	3,922
Sphex funerarius	SRR1503082	GBQD00000000	GBQD01000000	26,189	37,503,328	1,476
Stizoides tridentatus	SRR1503087	GBQO00000000	GBQO01000000	27,724	34,689,230	5,790
Tachysphex fulvitarsis	SRR1503055	GBPR00000000	GBPR01000000	17,308	32,940,922	8,852
Trypoxylon figulus	SRR1503089	GBWO00000000	GBWO01000000	19,174	31,640,527	4,800

References

Altenhoff, A. M., Škunca, N., Glover, N., Train, C.-M., Sueki, A., Piližota, I., Gori, K., Tomiczek, B., Müller, S., Redestig, H., Gonnet, G. H., & Dessimoz, C. (2015). The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*, 43(D1), D240–D249.

Bonasio, R., Zhang, G., Ye, C., Mutti, N. S., Fang, X., Qin, N., Donahue, G., Yang, P., Li, Q., Li, C., Zhang, P., Huang, Z., Berger, S. L., Reinberg, D., Wang, J., & Liebig, J. (2010). Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. *Science (New York, N.Y.)*, 329(5995), 1068–1071.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 421.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10), e1002195.

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5, 113.

Emms, D. M. & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157.

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114), 931–949.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780.

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simao, F. A., Pozdnyakov, I. A., Ioannidis, P., & Zdobnov, E. M. (2015). OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, 43(D1), D250–D256.

Li, L., Stoeckert, C., & Roos, D. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9), 2178–2189.

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A Software Package for Multispecies Target DNA Enrichment Probe Design. *Molecular Biology and Evolution*, 33(7), 1875–1886.

Misof, B., Liu, S., Meusemann, K., Peters, R., Donath, A., Mayer, C., Frandsen, P., Ware, J., Flouri, T., Beutel, R., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang,

Y., Jermiin, L., Kawahara, A., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von, R. B., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M., Wiegmann, B., Wilbrandt, J., Wipfler, B., Wong, T., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K., & Zhou, X. (2014). Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science*, 346, 763–7.

Munoz-Torres, M. C., Reese, J. T., Childers, C. P., Bennett, A. K., Sundaram, J. P., Childs, K. L., Anzola, J. M., Milshina, N., & Elsik, C. G. (2011). Hymenoptera Genome Database: Integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Research*, 39(Database), D658–D662.

Nygaard, S., Zhang, G., Schiøtt, M., Li, C., Wurm, Y., Hu, H., Zhou, J., Ji, L., Qiu, F., Rasmussen, M., Pan, H., Hauser, F., Krogh, A., Grimmelikhuijzen, C. J. P., Wang, J., & Boomsma, J. J. (2011). The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. *Genome Research*, 21(8), 1339–1348.

Slater, G. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *Bmc Bioinformatics*, 6(1), 31.

Smith, C. D., Zimin, A., Holt, C., Abouheif, E., Benton, R., Cash, E., Croset, V., Currie, C. R., Elhaik, E., Elsik, C. G., Fave, M.-J., Fernandes, V., Gadau, J., Gibson, J. D., Graur, D., Grubbs, K. J., Hagen, D. E., Helmkampf, M., Holley, J.-A., Hu, H., Viniegra, A. S. I., Johnson, B. R., Johnson, R. M., Khila, A., Kim, J. W., Laird, J., Mathis, K. A., Moeller, J. A., Muñoz-Torres, M. C., Murphy, M. C., Nakamura, R., Nigam, S., Overson, R. P., Placek, J. E., Rajakumar, R., Reese, J. T., Robertson, H. M., Smith, C. R., Suarez, A. V., Suen, G., Suhr, E. L., Tao, S., Torres, C. W., van Wilgenburg, E., Viljakainen, L., Walden, K. K. O., Wild, A. L., Yandell, M., Yorke, J. A., & Tsutsui, N. D. (2011). Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). *Proceedings of the National Academy of Sciences*, 108(14), 5673–5678.

Sonnhammer, E. L. L. & Östlund, G. (2015). InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43(Database issue), D234–239.

Struck, T. H., Wey-Fabrizius, A. R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., Klebow, S., Iakovenko, N., Hausdorf, B., Petersen, M., Kuck, P., Herlyn, H., & Hankeln, T. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. *Molecular Biology and Evolution*, 31(7), 1833–1849.

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., Bouffard, P., Caldera, E. J., Cash, E., Cavanaugh, A., Denas, O., Elhaik, E., Favé, M.-J., Gadau, J., Gibson, J. D., Graur, D., Grubbs, K. J., Hagen, D. E., Harkins, T. T., Helmkampf, M., Hu, H., Johnson, B. R., Kim, J., Marsh, S. E., Moeller, J. A., Muñoz-Torres, M. C., Murphy, M. C., Naughton, M. C., Nigam, S., Overson, R., Rajakumar, R., Reese, J. T., Scott, J. J., Smith, C. R., Tao, S.,

Tsutsui, N. D., Viljakainen, L., Wissler, L., Yandell, M. D., Zimmer, F., Taylor, J., Slater, S. C., Clifton, S. W., Warren, W. C., Elsik, C. G., Smith, C. D., Weinstock, G. M., Gerardo, N. M., & Currie, C. R. (2011). The genome gequence of the leaf-cutter ant Atta cephalotes reveals insights into Its obligate symbiotic lifestyle. *PL0S Genet*, 7(2), e1002007.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673–4680.

Tribolium Genome Sequencing Consortium (2008). The genome of the model beetle and pest Tribolium castaneum. *Nature*, 452(7190), 949–955.

Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., & Kriventseva, E. V. (2013). OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, 41(Database issue), D358–365.

Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J., & Kriventseva, E. V. (2011). OrthoDB: The hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Research*, 39(Database issue), D283–288.

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., The Nasonia Genome Working Group, Beukeboom, L. W., Desplan, C., Elsik, C. G., Grimmelikhuijzen, C. J. P., Kitts, P., Lynch, J. A., Murphy, T., Oliveira, D. C. S. G., Smith, C. D., v. d. Zande, L., Worley, K. C., Zdobnov, E. M., Aerts, M., Albert, S., Anaya, V. H., Anzola, J. M., Barchuk, A. R., Behura, S. K., Bera, A. N., Berenbaum, M. R., Bertossa, R. C., Bitondi, M.

M. G., Bordenstein, S. R., Bork, P., Bornberg-Bauer, E., Brunain, M., Cazzamali, G., Chaboub, L., Chacko, J., Chavez, D., Childers, C. P., Choi, J.-H., Clark, M. E., Claudianos, C., Clinton, R. A., Cree, A. G., Cristino, A. S., Dang, P. M., Darby, A. C., de Graaf, D. C., Devreese, B., Dinh, H. H., Edwards, R., Elango, N., Elhaik, E., Ermolaeva, O., Evans, J. D., Foret, S., Fowler, G. R., Gerlach, D., Gibson, J. D., Gilbert, D. G., Graur, D., Grunder, S., Hagen, D. E., Han, Y., Hauser, F., Hultmark, D., Hunter, H. C., Hurst, G. D. D., Jhangian, S. N., Jiang, H., Johnson, R. M., Jones, A. K., Junier, T., Kadowaki, T., Kamping, A., Kapustin, Y., Kechavarzi, B., Kim, J., Kim, J., Kiryutin, B., Koevoets, T., Kovar, C. L., Kriventseva, E. V., Kucharski, R., Lee, H., Lee, S. L., Lees, K., Lewis, L. R., Loehlin, D. W., Logsdon, J. M., Lopez, J. A., Lozado, R. J., Maglott, D., Maleszka, R., Mayampurath, A., Mazur, D. J., McClure, M. A., Moore, A. D., Morgan, M. B., Muller, J., Munoz-Torres, M. C., Muzny, D. M., Nazareth, L. V., Neupert, S., Nguyen, N. B., Nunes, F. M. F., Oakeshott, J. G., Okwuonu, G. O., Pannebakker, B. A., Pejaver, V. R., Peng, Z., Pratt, S. C., Predel, R., Pu, L.-L., Ranson, H., Raychoudhury, R., Rechtsteiner, A., Reid, J. G., Riddle, M., Romero-Severson, J., Rosenberg, M., Sackton, T. B., Sattelle, D. B., Schluns, H., Schmitt, T., Schneider, M., Schuler, A., Schurko, A. M., Shuker, D. M., Simoes, Z. L. P., Sinha, S., Smith, Z., Souvorov, A., Springauf, A., Stafflinger, E., Stage, D. E., Stanke, M., Tanaka, Y., Telschow, A., Trent, C., Vattathil, S., Viljakainen, L., Wanner, K. W., Waterhouse, R. M., Whitfield, J. B., Wilkes, T. E., Williamson, M., Willis, J. H., Wolschin, F., Wyder, S., Yamada, T., Yi, S. V., Zecher, C. N., Zhang, L., & Gibbs, R. A. (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, 327(5963), 343–348.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou,

X., Lam, T.-W., Li, Y., Xu, X., Wong, G. K.-S., & Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12), 1660–1666.

Declaration of authorship

I herewith declare that I have written this thesis independently and myself. I did not use any other sources than those listed. All places where the exact words or analogous text were taken from sources are indicated. I assure that this thesis has not been submitted for examination elsewhere.

April 13, 2020

Curriculum vitae

Malte Petersen

Born on 19. August 1983 in Hamburg-Bergedorf

Institute address

Zoologisches Forschungsmuseum Alexander Koenig (ZFMK), Adenauerallee 160, 53113 Bonn

Academic education

2013–2018	PhD candidate at the ZFMK
2005–2013	Undergraduate studies in Biology at the Rheinische Friedrich-Wilhelms-Universität Bonn
2004–2005	Undergraduate studies at the Universität Hamburg
2003	Abitur, Hansa-Gymnasium Hamburg-Bergedorf

```
#!/usr/bin/perl
(my$d=q[AA                GTCAGTTCCT
  CGCTATGTA                 ACACACACCA
    TTTGTGAGT                 ATGTAACATA
      CTCGCTGGC               TATGTCAGAC
        AGATTGATC           GATCGATAGA
          ATGATAGATC      GAACGAGTGA
            TAGATAGAGT GATAGATAGA
              GAGAGA GATAGAACGA
                TC GATAGAGAGA
                 TAGATAGACA G
               ATCGAGAGAC AGATA
             GAACGACAGA TAGATAGAT
           TGAGTGATAG    ACTGAGAGAT
         AGATAGATTG        ATAGATAGAT
       AGATAGATAG            ACTGATAGAT
     AGAGTGATAG              ATAGAATGAG
   AGATAGACAG                ACAGACAGAT
  AGATAGACAG                 AGAGACAGAT
  TGATAGATAG                ATAGATAGAT
  TGATAGATAG              AATGATAGAT
   AGATTGAGTG           ACAGATCGAT
     AGAACCTTTCT    CAGTAACAGT
       CTTTCTCGC TGGCTTGCTT
         TCTAA CAACCTTACT
           G ACTGCCTTTC
           TGAGATAGAT CGA
         TAGATAGATA GACAGAC
       AGATAGATAG  ATAGAATGAC
     AGACAGAGAG      ACAGAATGAT
   CGAGAGACAG          ATAGATAGAT
  AGAATGATAG             ACAGATAGAC
  AGATAGATAG               ACAGACAGAT
  AGACAGACTG                 ATAGATAGAT
   AGATAGATAG                 AATGACAGAT
     CGATTGAATG               ACAGATAGAT
       CGACAGATAG             ATAGACAGAT
         AGAGTGATAG           ATTGATCGAC
           TGATTGATAG       ACTGATTGAT
             AGACAGATAG  AGTGACAGAT
               CGACAGA TAGATAGATA
                 GATA GATAGATAG
                    ATAGACAGA G
                  AGATAGATAG ACA
                GTCGCAAGTTC GCTCACA
])=~s/\s+//g;%a=map{chr $_=>$i++}65,84,67,
71;$p=join$;,keys%a;while($d=~/([$p]{4})/g
){next if$j++%96>=16;$c=0;for$d(0..3){$c+=
$a{substr($1,$d,1)}*(4**$d)}$perl.=chr $c}
             eval $perl;
```

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration is an obfuscated Perl program that prints the text "Just another genome hacker". Unfortunately, the source for this program is no longer known to me. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.