The “Atelocerata” – A vanishing hypothesis?

Molecular phylogeny of arthropods with focus on primary wingless hexapods

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch–Naturwissenschaftlichen Fakultät

an der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Karen A. Meusemann
aus Bonn

Bonn, Dezember 2010

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

Die Dissertation wurde am Zoologischen Forschungsmuseum Alexander Koenig (ZFMK), Bonn durchgeführt.

1. Prüfer: Prof. Dr. Bernhard Misof
2. Prüfer: Prof. Dr. Thomas Bartolomaeus
3. Prüfer: Prof. Dr. Wolfgang Alt
4. Prüfer: Prof. Dr. Jes Rust

Tag der mündlichen Prüfung: 30. März 2012
Erscheinungsjahr: 2012

"Nothing in Biology Makes Sense Except in the Light of Evolution."

(Theodosius Dobzhansky, 1973)

All jene faszinierenden Geschöpfe, die für diese Studie
ihr Leben lassen mussten, verdienen meinen größten Respekt.

Meinen Eltern, meinen Katern und meinem besten Freund

Contents

List of Figures

List of Tables

Abbreviations

ASRV	Among site rate variation
BFT	Bayes Factor Test
bp	base pairs
BS	boostrap support
cDNA	copy DNA
CT	computer tomography
°C	*Celsius*
DNA	desoxyribonucleine acid
eGM	extended Geometry Mapping
EMBL	European Molecular Biology Laboratory, Heidelberg
EST	Expressed Sequence Tag
GB	Giga byte
HaMStR	Hidden Markov Model based Search for Orthologs using Reciprocity
HMM	Hidden Markov Model
HPC	High performance computing
IC	Information content
ITS	Internal transcribed spacer
in prep.	in preparation
LH	likelihood
LM	Likelihood mapping
LSI	Leaf stability index
MARE	MAtrix REduction (new developed software)
MCMC	Markov Chain Monte Carlo
ml	milliliter
μl	micro liter
mRNA	messenger RNA
ML	Maximum likelihood
mm	millimeter
μ	micro
mt	mitochondrial
NCBI	National Center for Biotechnology Information
PCR	Polymerase chain reaction
pers. comm.	personal coommunication
pmol	pico mol
pP	posterior probability
RAM	random access memory
RNA	Ribonucleine acid

rRNA	ribosomal RNA
RRZK	Regionales Rechenzentrum, Universität zu Köln
RT-PCR	reverse transcriptase PCR
s	second
SOS	selected optimal subset
SuGI	Sustainable Grid Infrastructure
sp.	species not determined
SPP	Schwerpunktprogramm (priority programme)
u	unit
ZFMK	Zoologisches Forschungsmuseum A. Koenig, Bonn

Summary

Arthropods encompass more than three quarters of all described living species. Among arthropods, hexapods are the most abundant group and show an enormous diversity. To understand evolutionary processes of and within hexapods, it is necessary to resolve phylogenetic relationships, especially early hexapod splits within an arthropod framework. Several contradicting hypotheses have been suggested in the last decade addressing the monophyly of hexapods, and placements and relationships of primary wingless hexapod orders. As a possible sister group of hexapods, traditionally myriapods have been suggested, uniting hexapods and myriapods to a clade "Atelocerata". Alternatively several crustacean taxa (branchiopods or copepods) have been proposed as a possible sister group of hexapods associated with the "Pancrustacea" hypothesis where hexapods and crustaceans are united into a clade. This study concentrates on the analyses of molecular data and aims to resolve deep hexapods relationships within an arthropod context using two different methodological approaches.

In the first approach, large phylogenomic (multi-gene and taxon) data sets based on nuclear protein coding genes derived from Expressed Sequence Tags (ESTs) are analyzed. This approach uses raw data sets with more than 100 taxa and more than 700 genes, and optimal data subsets with more than 100 taxa and more than 100 most informative genes to reconstruct phylogenetic trees. This is the first phylogenomic study which takes all entognathous, primary wingless hexapod orders into account. For this study, new EST data have been generated for each entognathous (Protura, Diplura and Collembola) and one ectognathous primary wingless hexapod order (Archaeognatha). A new approach based on a hill climbing algorithm is introduced to select most informative taxa and genes from raw data matrices. The aim is to select an optimized data subset with high information content. Therefore, MARE, (MAtrix REduction, `http://www.mare.zfmk.de`) has been developed by our work group. Optimized data subsets (SOS) are selected by taking information content of single genes (partitions) and the complete matrix into account without loosing to much taxa. For phylogenomic data sets addressing phylogenetic relationships, such an approach has never previously been applied. Instead, available studies rely on thresholds of available data or on maximal connected groups of data presence. Effects of selecting optimized data subsets towards high information content are examined with respect to phylogenetic reconstructions. Altogether, phylogenomic data can substantially advance our understanding of arthropod evolution and resolve several conflicts among existing hypotheses. Optimized data subsets show strong support for a sister group relationship of onychophorans and euarthropods. Within pancrustaceans, analyses yield paraphyletic crustaceans and monophyletic hexapods and robustly resolved deep hexapod relationships. Within neopteran insects, endopterygote (holometabolous) insects are monophyletic where hymenopterans branch off first with strong support. Analyses show a remarkable sensitivity to methods of analyses for the placement of myriapods. Altogether, results of this thesis show that new heuristics for the selection of optimized submatrices and other applied tools to improve data quality pay off their effort.

The second approach to resolve deep hexapod relationships within an arthropod framework relies on two well known nuclear ribosomal RNA genes (large subunit 28S and small subunit 18S rRNA).

Both genes are popular markers for studies addressing metazoan, arthropod and hexapod phylogeny. Analyses using an arthropod data set with 148 taxa of all important arthropod groups including both nuclear rRNA genes are improved by employing plausible models of sequence evolution. This rRNA study incorporates background knowledge on the evolution of nuclear ribosomal RNA gene sequences, in particular, into various steps of data processing. Mainly, automated methods have been used, an automated secondary structure guided alignment approach (RNAsalsa) and the software Aliscore for alignment masking. Further, mixed RNA/DNA models have been applied to avoid artifacts due to interdependence and covariation of paired sites of rRNA genes. Concurrently, reconstruction methods have been used that account for variation of evolutionary rates among lineages (non-stationarity). Although split-decomposition networks indicated conflicting signal in the data set, analyses modeling non-stationary statistically outperform stationary approaches. Topologies show strong support for a pancrustacean clade. The placement of some myriapod orders remains suspicious, a sister group of hexapods cannot robustly be resolved. Analyses taking non-stationarity into account unequivocally propose monophyletic Hexapoda. Relationships among entognathous primary wingless hexapods are resolved and Ectognatha are maximally supported. Within endopterygotes, hymenopterans are strongly proposed as a sister group of remaining holometabolous insects. Again, advanced methods in data quality assessment and modeling pay off its effort.

This thesis was conducted within the DFG priority program SPP 1174 “Deep Metazoan Phylogeny” and funded by the grant MI 649/6.

1. General Introduction

Arthropods encompass more than three quarters of all described living species. They were the first animals to conquer land and air. This extraordinary evolutionary success is based on an astoundingly wide array of highly adapted body organizations. Still, it is not clear, how often they conquered land. Relationships within and between the major arthropod groups, chelicerates, myriapods, crustaceans and hexapods (Euarthropoda), are still unresolved. Since arthropod phylogeny has been addressed in scientific studies, almost every possible scenario has been proposed. In the late 20^{th} century, molecular studies again reanimated the debate.

1.1. Relationships between major arthropod clades

Mandibulata: "Atelocerata" versus "Pancrustacea"

Traditionally, crustaceans, myriapods and hexapods are subsumed under "Mandibulata" (Snodgrass, 1935). The clade unites all euarthropods with mandibles (Harzsch et al., 2005; Scholtz and Edgecombe, 2006). Chelicerates are considered as sister group of Mandibulata. Mandibulates were originally classified into crustaceans and "Atelocerata" (Heymonds, 1901; Snodgrass, 1938; Ax, 1990; Bitsch and Bitsch, 2004), or "Tracheata" (Pocock, 1893; Ax, 1990; Bäcker et al., 2008). The term Atelocerata is derived from the loss of the second pair of antennae and unites myriapods and hexapods (Snodgrass, 1938). Crustaceans were considered as sister group of Atelocerata, but already Snodgrass (1935) contemplated a possible paraphyly of crustaceans. Suggested morphological characters, e.g. the tracheal system and spiracles, the presence of eversible vesicles, the structure of mandibles (see Klass and Kristensen, 2001; Koch, 2001) or the equipment of the trunk pleura of myriapods and insects with a characteristic set of concentric sclerites around the leg base and accompanying muscles (Bäcker et al., 2008) support Atelocerata. Paulus (1979) initially challenged this view based on his work on ommatidial structures of arthropod compound eyes. He described striking similarities between fine structural organizations of insects and crustaceans and indicated potential problems with the taxon Atelocerata. Likewise, Averof and Akam (1995) suggested a crustacean-like common ancestor of hexapods and crustaceans from comparative developmental and molecular studies. Additionally, neurobiological studies (Fanenbruck et al., 2004; Harzsch et al., 2005; Harzsch, 2006; Ungerer and Scholtz, 2008) supported a clade "Pancrustacea" (Zrzavý and Štys, 1997) or "Tetraconata" (Dohle, 2001).

Molecular analyses proposed Pancrustacea (Friedrich and Tautz, 1995; Regier and Shultz, 1997; Shultz and Regier, 2000; Giribet et al., 2001). Seminal work was published by Boore et al. (1995) and Boore et al. (1998): they found that crustaceans and hexapods share common mitochondrial (mt) gene arrangements with exclusion of myriapods. Pancrustacea is supported by most molecular single gene analyses (Friedrich and Tautz, 1995; Shultz and Regier, 2000; Friedrich and Tautz, 2001; Giribet et al., 2001; Hwang et al., 2001; Regier and Shultz, 2001; Pisani et al., 2004; Mallatt et al., 2004; Regier et al., 2005; Hassanin, 2006; Mallatt and Giribet, 2006; Bourlat et al., 2008; Regier et al.,

2008; Dell'Ampio et al., 2009) and, recently, by an extensive phylogenomic analysis of metazoan taxa (Dunn et al., 2008) and a multi-gene analysis of arthropods (Regier et al., 2010). However, many of these studies present reconstructions that lack a robust resolution between Chelicerata, Myriapoda and Pancrustacea (e.g. Mallatt and Giribet, 2006; Dunn et al., 2008; Regier et al., 2010).

"Mandibulata" versus "Myriochelata"

The monophyly of Mandibulata was unchallenged for a long time (Snodgrass, 1938; Wägele, 1993; Boore et al., 1995; Scholtz et al., 1998; Edgecombe et al., 2000; Giribet et al., 2001). Based on molecular studies (Hwang et al., 2001; Friedrich and Tautz, 2001; Mallatt et al., 2004; Hassanin et al., 2005; Hassanin, 2006; Dunn et al., 2008), a possible sister group relationship of Chelicerata + Myriapoda has been suggested, coined "Myriochelata" (Pisani et al., 2004) or "Paradoxopoda" (Mallatt et al., 2004). Kadner and Stollewerk (2004) reported a correspondence in neurogenesis of Myriapoda and Chelicerata, but alternatively the suggested character sets may reflect a plesiomorphic state within Euarthropoda (Stollewerk and Simpson, 2005; Stollewerk and Chipman, 2006). Mayer and Whintington (2009) studied the nervous system of velvet worms and suggested possible, apomorphic characters for Paradoxopoda from neurogenesis and embryonic germ disk. Still, characters that support Paradoxopoda are sparse (Wägele, 1993; Klass and Kristensen, 2001; Dohle, 2001; Harzsch, 2006; Bäcker et al., 2008; Minelli, 2009; Edgecombe, 2009, 2010; Shear and Edgecombe, 2010).

From a molecular perspective, it remains to be tested whether alternative reconstruction methods and additional molecular data (e.g. from myriapods and especially from apterygote hexapod lineages) support a pancrustacean clade. The resolution of these questions will shed light on the evolution of terrestrial arthropods, e.g. if arthropods conquered land only once, or if hexapods are in fact 'specialized terrestrial crustaceans'.

1.2. Hexapods and the role of apterygote lineages

Within hexapods, relationships are far from being resolved (Ogden and Whiting, 2003; Kjer, 2004; Kukalová-Peck and Lawrence, 2004; Misof et al., 2007; Whitfield and Kjer, 2008). Major controversies concern the earliest splits within hexapods, especially relationships of the so-called apterygote lineages or "basal" hexapods: Protura, Collembola, Diplura, Archaeognatha and Zygentoma (see Koch, 1997; Kristensen, 1998; Carapelli et al., 2000, 2005, 2007; Giribet et al., 2001; D'Haese, 2002a,b; Nardi et al., 2003a; Delsuc et al., 2003; Luan et al., 2003; Kjer, 2004; Giribet et al., 2004; Regier et al., 2004; Luan et al., 2005; Misof et al., 2007; Regier et al., 2008, 2010; Dell'Ampio et al., 2009). They play an essential role as a possible link to resolve major arthropod relationships. Consequently, the inclusion of these taxa in phylogenetic reconstruction is crucial. So far, they are little considered in morphological and most molecular studies.

The taxon Apterygota (Lang, 1888) constitutes an artificial group, because the primary winglessness is considered as a symplesiomorphy (Hennig, 1953). Since the mid-20^{th} century, hexapods were classified into Entognatha and Ectognatha (Hennig, 1953, 1969, 1981; Kristensen, 1991). Entognatha include Protura (coneheads), Diplura (diplurans) and Collembola (springtails) and have been traditionally considered as sister group of Ectognatha (Kristensen, 1981). All three orders show, except the hexapody and the division of the body into *caput*, *thorax* and *abdomen*, extremely aberrant characters compared with a hexapod ground plan. Ectognatha include the primary wingless orders

Archaeognatha (bristletails), Zygentoma (silverfish and firebrats) and pterygote insects. Zygentoma have been considered as sister group to Pterygota (Hennig, 1969; Kristensen, 1975).

Ectognatha show eleven abdominal segments, epimeric development and compound eyes. In contrast to Entognatha, the monophyly of Ectognatha is generally accepted from morphological and molecular point of view (e.g. Kristensen, 1975; Kjer et al., 2006; Misof et al., 2007; Szucsich and Pass, 2008; Grimaldi, 2010; Regier et al., 2010).
Apomorphic characters are for example the construction of antennae where muscle is only present in the basal part (scapus), the Johnston's organ, seperated antennae vessels of the aorta, etc.

Ectognathous, primary wingless orders

Archaeognatha (bristletails) are classified into two recent families and comprise ca. 500 species. Usually, they occur under rocks or barks. They are identifies by large compound eyes that meet dorsally (apomormphic), and large *ocelli*. The filamentous, multi-segmented antennae are at their base closely positioned. Like entognathous hexapods, the mandible is connected with the head capsule by one articulation point (monocondylic). The large, seven-segmented maxillary palps are prominent. Like Zygentoma, bristletails have two cerci and a median filament (apomorphic), but the filament is longer, mostly as long as their body. Archaeognatha can jump by a flexure of the body. Currently, they are the assumed sister group to Dicondylia (Zygentoma + pterygote insects).

Zygentoma (silverfish and firebrats) share several plesiomorphic characters with Archaeognatha. Both use indirect sperm transfer. Like Archaeognatha, most zygentoman species have a scale-covered body (Grimaldi and Engel, 2005; Grimaldi, 2010). The second articulation point of the mandibles unites Zygentoma with pterygote insects. This character complex is often considered as a synapomorphy (Staniczek, 2003), but see Koch (2001), Bitsch and Bitsch (2004) and Regier et al. (2004). Zygentoma lack the archaeognathan hump; the body is flattened and they do not jump. Compound eyes are mostly reduced or absent. They are classified into five recent families. The family Lepidotrichidae, with probably most primitive characters, is represented by only one extant species, *Tricholepideon gertschi* (Wygodzinski, 1965). Other Lepidotrichidae are only known from Baltic amber. They possess three *ocelli*, while these are absent in all other families. The monophyly of Zygentoma has been put into question by morphological, molecular and combined analyses: they have been proposed as paraphylum with *Tricholepideon* as sister taxon to remaining Dicondylia (Beutel and Gorb, 2001; Koch, 2001; Regier et al., 2004; Beutel and Gorb, 2006; Kjer, 2004; Kjer et al., 2006; Carapelli et al., 2007; Misof et al., 2007; Comandi et al., 2009).

Entognathous, primary wingless orders

Protura are tiny animals and characterized by their cryptic life style. Described by Silvestri (1907), they were classified into hexapods, but they show many unique and untypical characters for hexapods. Protura are probably the most aberrant hexapod group. The term "Protura" is derived from "proto-" [gr.: first, original] and "ura" [gr.: tail] and refers to the lack of advanced or specialized structures at the back of the abdomen. "Coneheads" [engl.] is derived from the cone-shaped head. Proturans have three well developed pairs of legs which each consist of six podomers, similar to 'typical' hexapods. "Beintastler" [germ.] originates from the usage of their forelegs: usually, they take over a sensory or antennal-like function. The abdomen consists of twelve abdominal segments. The first

three abdominal segments carry one pair of rudimentary limbs. Berlese described this order in 1909 as "Myrientomata", an in-between between hexapods and myriapods. Proturans show an anameric development like some myriapod lineages (Grimaldi and Engel, 2005; Carapelli et al., 2006; Dallai et al., 2010). The early larval stages comprise nine and ten abdominal segments. They do not molt after sexual maturation in contrast to Diplura, Collembola and Zygentoma (Grimaldi and Engel, 2005). They lack compound eyes and ommatidia. The genitals are quite different from all other hexapods. Proturans lack antennae. On the head, they possess temporal organs, the *pseudoculi*. Their function is still unknown. In Westheide and Rieger (e.g. 2007) and Klass (2007) these organs are termed "Tömösváry organs", but an assignment with the Tömösváry organs of myriapods or the postantennal organs (PAO) of springtails is questioned by several authors, (e.g. Tuxen, 1964). The ultrastructure of the sperm axoneme also deviates from other hexapods: the microtubuli show unusual patterns (Bacetti and Dallai, 1973; Dallai et al., 1990, 1992; Dallai, 1994; Dallai and Afzelius, 1999; Machida et al., 2002; Dallai et al., 2010).The absence of many hexapod diagnostic characters has repeatably been used to argue against an inclusion of Protura into Hexapoda (Tuxen, 1964).

Diplura are small in size and live in soil, leaf litter and compost heap. As proturans, they lack ommatidia and compound eyes. They possess only ten abdominal segments, including one pair of styli and eversible vesicles on the first seven segments (e.g. Ikeda and Machida, 1998; Klass and Kristensen, 2001). Diplura are classified into Campodeoidea, Japygoidea and Projapygoidea (Pagés, 1959). Campodeoidea have multi-segmented cerci and are omnivorous. Japygoidea show unsegmented forcipate cerci and are predatory using their maxillae or their forceps to capture prey. The microtubuli pattern in sperm axonemes of Campodeoidea resembles Insecta *sensu stricto* (Jamieson et al., 1999). In contrast, Japygoidea, show a sperm axonem microtubuli pattern similar to proturans and springtails. Also ovariole structures of Campodeoidea differ from those of Japygoidea. A possible paraphyly of diplurans is still discussed, but comprehensive analyses of morphological and molecular characters support a monophyly of Diplura (e.g. Koch, 1997; Luan et al., 2005; Gao et al., 2008). Fossil record is present, but its applicability was put into question because of the low quality of preservation (Grimaldi and Engel, 2005; Szucsich and Pass, 2008).

Collembola form the most species-rich order. They often exhibit extremely high population densities, which makes them the worlds most abundant hexapods. Springtails are cosmopolitans. They have a temporal organ at the base of primary four-segmented antennae (postantennal organ, PAO). Such organs are known from myriapod symphylans and centipedes and maybe proturans (*pseudoculi*), but see Tuxen (1964). The PAO probably functions as chemo- or hygroreceptor. In contrast to diplurans and proturans, collembolans show maximally eight ommatidia (*pseudocelli*). Except for Symphypleona, Collembola lack tracheae. Springtails have well developed legs with five podomers: instead of a tibia and a tarsus, there is a *tibiotarsus* The abdomen has six segments; the first segment carries the ventral tube, a multi-functional organ for physiological regulations. The name "Collembola" ("kola" [gr.: glue] and "embolon" [gr.: piston, peg]) refers to this ventral tube which has adhesive properties. The primary function is most likely for excretion and maintaining water balance. The third and fourth abdominal segment carry the *retinaculum* and *furculum*. These interlocking structures enable the springtail a hardly controlled catapult into the air, an effective means of escaping predation. The monophyly of Collembola Lubbock 1873 is accepted (Hennig, 1953; Bitsch and Bitsch, 2000, 2004; Giribet et al., 2004; Luan et al., 2005; Dell'Ampio et al., 2009). The fossil *Rhyniella praecursor* (Hirst, 1926), a springtail similar to the family Isotomidae from the Early Devonian, is one of the

oldest documented hexapods (Grimaldi and Engel, 2005; Grimaldi, 2010). Analyses of mitochondrial protein coding genes suggested to exclude Collembola from hexapods. Consequently, the monophyly of Hexapoda and Entognatha was questioned (Nardi et al., 2003a,b; Carapelli et al., 2005, 2007).

Monophyly of Hexapoda, Entognatha and relationships of apterygote lineages

Morphological support for monophyletic Hexapoda is weak (see Klass and Kristensen, 2001; Carapelli et al., 2007; Szucsich and Pass, 2008; Grimaldi, 2010). The most obvious apomorphy is the three-tagmated body organization into head, the three-segmented thorax with three pairs of locomotory limbs and the abdomen with eleven segments. The abdomen carries different numbers of rudimentary limbs.

Except from the tagmatic body and the hexapody, Entognatha differ in most characters from this 'hexapod ground plan' (Klass and Kristensen, 2001). For example, the number of abdominal segments deviates in all entognathous orders. Instead of six, Collembola only have five podomers with its characteristic *tibiotarsus*. Abdominal appendages of Collembola differ from other hexapods. Protura and Diplura lack any pigment cells, ommatidia and compound eyes (see above). Sperm axoneme structures from proturans are completely aberrant from other hexapods (e.g. Machida, 2006; Dallai et al., 2010). Most problems occur by the question how to polarize character states, either as possible synapomorphies or plesiomorphies (see discussion in Szucsich and Pass, 2008). A correct assignment requires knowledge of the sister group, but still, this question has not been resolved. The only proposed 'alternative hypothesis' to the monophyly of Hexapoda is that "hexapods are not monophyletic" (e.g. Nardi et al., 2003a,b; Carapelli et al., 2005, 2007). This negation, however, lacks convincing arguments (see Szucsich and Pass, 2008).

Apart from the sparse availability of molecular data, molecular studies provide contradicting results concerning the monophyly of Hexapoda. Friedrich and Tautz (1995, 2001) and Giribet et al. (2001) obtained Hexapoda with moderate support. Most contradictions arose from analyses of mitochondrial (mt) data sets (Nardi et al., 2003a). Hexapods were reconstructed as polyphyletic within crustaceans. This scenario lacks any morphological support. A reanalysis of this data suggested monophyletic Hexapoda, although with weak support (Delsuc et al., 2003). Mitochondrial studies of Cook et al. (2005) and Hassanin (2006) did not support hexapod monophyly. Several additional, fully characterized mt genomes of basal hexapods have been published (Podsiadlowski, 2006; Podsiadlowski et al., 2006; Carapelli et al., 2007, 2008). However, in all these studies Protura are missing. Giribet et al. (2004) rejected the monophyly of hexapods based on molecular mt + nuclear markers and combined analyses. In their scheme, Collembola and Ectognatha were related to crustaceans; Protura + Diplura (Nonoculata) were proposed as sister group to Symphyla (Myriapoda). Only few studies included all primary wingless hexapod orders which mostly based on nuclear ribosomal RNA genes. These studies strongly support Hexapoda (Luan et al., 2004, 2005; Regier et al., 2005; Mallatt and Giribet, 2006; Dell'Ampio et al., 2009; von Reumont et al., 2009; Mallatt et al., 2010). Yet, a phylogenomic approach covering all entognathous orders is missing.

The validity of the clade Entognatha has been questioned since Hennig (1953). The traditional "Ellipura" hypothesis (Börner, 1910) uniting Protura + Collembola, is based on morphological features. The "Nonoculata" hypothesis arose from molecular data (Luan et al., 2005) and unites Protura + Diplura. Assuming Entognatha as a valid taxon, either monophyletic (Hennig, 1981) or paraphyletic

diplurans (Štys and Bilinski, 1990) are suggested as sister group to Ellipura. Alternatively, Collembola are proposed as sister group to Nonoculata (Luan et al., 2003, 2004, 2005; Mallatt and Giribet, 2006; Dell'Ampio et al., 2009; von Reumont et al., 2009; Mallatt et al., 2010).

Several relationships have been reconstructed in molecular and combined analyses that reject Entognatha. For example, Giribet et al. (2004) suggested a close relationship of Collembola and Ectognatha, Nonoculata were placed within crustaceans. Nardi et al. (2003a) reconstructed Collembola as sister group to remaining pancrustaceans (see above). Studies on embryogenesis and sperm ultrastructure suggested the exclusion of Protura from Entognatha and even Hexapoda (Machida, 2006; Dallai et al., 2010). A sister group relationship of Diplura and Ectognatha was proposed based on attachment structures (Beutel and Gorb, 2001, 2006).

The clade Ellipura mainly traces back to observed similarities considering entognathy (Hennig, 1981; Kristensen, 1981, 1998; Klass and Kristensen, 2001; Bitsch and Bitsch, 2004; Beutel and Gorb, 2006; Klass, 2007; Szucsich and Pass, 2008). The mouthpart appendages of Protura, Diplura and Collembola are recessed within a gnathal pouch on the head capsule with oral folds (*plica oralis*) building its lateral parts. Oral folds meet at a ventral midline, the *linea ventralis*. This ventral cuticular groove extends backwards from the labium / neck to the thorax and is only present in collembolans and proturans. Cephalic folds are connected with labrum and labium, building the gnathal pouch. The base of maxillae and mandibles of one side are separated. In contrast, Diplura have a less developed pouch, the base of maxillae and mandibles is not separated; the oral folds extend to the labium, but remain differentiated from it by a longitudinal, small sclerite, the *admentum*. The proposed synapomorphy of entognathy was put into question based on morphological studies. Koch (1997) argued that both states of entognathy cannot be traced to a common ground pattern. Therefore, the entognathous condition was discussed as homoplasy; Diplura might have independently acquired this feature associated with the entognathous peculiar feeding behavior (see Carapelli et al., 2006). Szucsich and Pass (2008) alternatively discussed the entognathous condition as a plesiomorphic state and showed that analyses of character sets can result in contradicting hypotheses. Oral folds are also present in Archaeognatha and *Tricholepideon* (Staniczek, 2000), but very short-formed. This has been either interpreted as 'pseudo-entognathy' or as a plesiomorphic state of hexapods. Embryological differences concerning the entognathy are known between diplurans and collembolans, but yet, nothing is known from proturans. However, if those characers are in general usefull for homology has to be out in question. While the entognathous condition *per se* as a synapomorphic character has also been questioned, the *linea ventralis* and its evolutionary origin is probably the only positive synapomorphy of Protura and Collembola (Klass and Kristensen, 2001; Szucsich and Pass, 2008).

Mainly reduced or absent characters also questioned the validity of Entognatha, for example 'reduced' ommatidia and compound eyes, 'reduced' malpighian tubules or the 'reduced' abdomen. However, character absence delivers only weak arguments for synapomorphy or symplesiomorphy statements (see discussion in Szucsich and Pass, 2008). When a pattern of present and absent characters can be recognized, this might be again more useful for phylogenetic statements.

Reduced eyes as possible synapomorphy of Ellipura can be considered as a weak argument, because ommatidia and compound eyes are completely absent in Protura. The term "Nonoculata" is derived from the absence of eyes in both Protura and Diplura. From a morphological view, the absence of

eyes and ommatidia was discussed as only known synapomorphy (character polarity: loss of eyes). Szucsich and Pass (2008) state several uncertainties about the appropriateness of this character: a loss as common origin conflicts with the dipluran fossil *Testajapyx* (Kukalová-Peck, 1987) (see below). Secondly, the loss of eyes reveals a high probability of homoplasy. Within arthropods eyes have been independently lost many times. Szucsich and Pass (2008) point out "a ventrical articulation of the coxa with the sternite of the respective body segment" as a possible synapomorphy, but indicate problems with this character. Nonoculata is not further supported by morphological or developmental reconstructions. In combined morphological and molecular "total evidence" analyses, Giribet et al. (2004) proposed a sister group relationship of Nonoculata + Symphyla within polyphyletic hexapods. However, Nonoculata have been recovered unequivocally, mostly in molecular (rRNA) analyses (Wheeler et al., 2001; Luan et al., 2003, 2004, 2005; Kjer, 2004; Giribet et al., 2005; Kjer et al., 2006; Mallatt and Giribet, 2006; Misof et al., 2007; Gao et al., 2008; Dell'Ampio et al., 2009; Mallatt et al., 2010, etc.), challenging the traditional Ellipura – concept. In these studies, Hexapoda are strongly supported and Entognatha show moderate support. Collembola are mostly inferred as sister group to Nonoculata. However, the clade Nonoculata is discussed as a result of systematic bias (Luan et al., 2003, 2004, 2005; Kjer, 2004; Giribet et al., 2005; Kjer et al., 2006; Mallatt and Giribet, 2006; Misof et al., 2007; Gao et al., 2008; Dell'Ampio et al., 2009; Mallatt et al., 2010). Analyses of von Reumont et al. (2009) show also strong support for Nonoculata (see chapter 3): since non-stationary processes were included in the modeling approach, it is suggested that systematic bias caused by compositional base heterogeneity can be excluded for this taxon.

Klass (2007) and Grimaldi (2010) state that "Ellipura are characterized by a reduced tracheal system". In Collembola, a neck stigma is only present in Symphypleona; in Protura, meso- and metathoracal stigmata are only present in Eosentomidae. Therefore, the polarization of this character remains questionable.

Beutel and Gorb (2001, 2006) reconstructed Ellipura, based on absent characters: both, Protura and Collembola have no cerci. Additionally, the presence of cerci was used to suggest a sister group relationship of Diplura + Ectognatha. However, the formulation of the lack of cerci on an 11th abdominal segment in springtails and proturans is questionable as independent character: collembolans possess only six abdominal segments and proturans twelve. The different number of abdominal segments in all three orders is from an ontogenetical perspective not reduced. Mainly, Diplura + Ectognatha (alternatively coined "Cercophora" (Staniczek and Bechly, 2007)), have been proposed based on external morphology (e.g. Kukalová-Peck, 1987, 1991; Koch, 1997, 2001; Kristensen, 1998; Bitsch and Bitsch, 2000, 2004; Wheeler et al., 2001). Several character complexes, e.g. a similar reduction of abdominal limbs or their constitution (Kukalová-Peck, 1991), lack comparative studies. The fossil record of Diplurans is very poor (see Staniczek and Bechly, 2007). Thus, *Testajapyx thomasi* from the upper carbon, published by Kukalová-Peck (1987) reanimated the debate. It shows ectognathous mouthparts, compound eyes and palpi, but appendages forceps like Japygoidea supporting Diplura + Ectognatha implying various parallel developments. Its assignment, however, has been questioned because of poor data quality (Kristensen, 1998; Grimaldi and Engel, 2005; Grimaldi, 2010).

Supporting Ellipura, the reduction of antennae (springtails have maximally four antennomeres) has also been proposed, but Protura have no antennae at all (Tuxen, 1964; Nosek, 1967, 1973; Szucsich and Pass, 2008). In Westheide and Rieger (2007), reduced antennae in Protura are homologized with the temporal organs without any evidence. The temporal organs (*pseudoculi*) have been also proposed

as reduced eyes and compared with ommatidia (*pseudocelli*) in Collembola (Grimaldi, 2010). A homology of *pseudoculi* (Protura), the PAO (Collembola) and the Tömösváry organs in myriapods and probably malacostracans has also been proposed (Klass and Kristensen, 2001; Klass, 2007; Westheide and Rieger, 2007). If true, these temporal organs might be a synapomorphy of mandibulates, but have been lost in many groups, not only in Diplura and Ectognatha. Still, proturans are not studied well enough to allow plausible conclusions (which is already stated by Tuxen, 1964).

The occurrence of unpaired, pretarsal claws in Protura and Collembola (Boudreaux, 1979) was suggested as synapomorphy. Diplura and Ectognatha have dorsolateral paired claws (Kristensen, 1981). Szucsich and Pass (2008) pointed out the presence of an additional median unpaired claw-like structure in Japygidae and Zygentoma (already recognized by Snodgrass, 1935). Depending on the coding strategy, contradicting hypotheses can be drawn: a synapomorphic loss of paired claws for Ellipura or a synapomorphic appearance in Diplura + Ectognatha. An outgroup comparison does not help, because undivided claw-like praetarsi or paired claws can be found in many euarthropod groups. Thus, this character complex might be "not useful in phylogenetic inference" (Szucsich and Pass, 2008). Nevertheless, it is still used as synapomorphic character (cf. Grimaldi, 2010; Dallai et al., 2010).

The anameric development of Protura, aberrant microtubuli patterns of their sperm axoneme and the presence of a functional primary embryonic membrane (*serosa*) clearly contradict Ellipura, and moreover, Entognatha and Hexapoda (Dallai, 1991; Dallai and Afzelius, 1999; Machida, 2006; Dallai et al., 2010). The microtubuli pattern of the sperm axoneme differs within proturans and within diplurans. Considering Diplura, several authors point out a resemblance of the dipluran flagellum to Ectognatha: the pattern of the central microtubuli shows an additional ring (9+9+2) (Klass, 2007; Szucsich and Pass, 2008, referring to Jamieson (1987)). Jamieson (1987) only observed this for Campodeoidea while Japygoidea show a collembolan-like structure (9+2, Jamieson, 1987; Jamieson et al., 1999). On the other hand, Campodeoidea were proposed to be closer related to Ellipura: Campodeoidea show non-metameric, sac-like ovaries (Štys and Bilinski, 1990; Štys et al., 1993; Bilinski, 1994; Štys and Zrzavý, 1994). In contrast, Japygoidea show a metameric arrangement of ovarioles (Štys et al., 1993) like Ectognatha, which was proposed as a potential synapomorphy (Bilinski, 1994). The genital character complexes, taken as possible synapomorphy, reject both Diplura and Entognatha. From an embryological perspective, a sister group relationship of Diplura + Ectognatha is supported. Both have a secondary membrane (*amnion*). Protura and Collembola only possess a *serosa*. In contrast to other hexapods, the proturan *serosa* differentiates into the dorsal body wall (Machida, 2006). The *serosa* in Collembola, Diplura and Ectognatha might have lost its function (Machida et al., 2002; Machida, 2006). Machida (2006) cautiously suggested to exclude proturans from hexapods. On the other hand, the availability of proturan embryological data is still too sparse for any conclusion (Machida, 2006). The *serosa* is also present in myriapods and crustaceans wherefore the interpretation of this character as a plesiomorphy remains possible.

Monophyly of entognathous orders

The monophyly of Protura and of Collembola is widely accepted. In contrast, the monophyly of Diplura was seriously questioned based on comparative genital structure studies (see above, Jamieson, 1987; Jamieson et al., 1999; Štys and Bilinski, 1990; Štys et al., 1993; Bilinski, 1994; Štys and Zrzavý, 1994). However, Koch (1997) argued that entognathous conditions of Campodeoidea and Japygoidea

provided enough evidence for the monophyly of Diplura, because the form of entognathy is much more similar compared with Protura and Collembola. Other morphological studies obtained contradicting results (e.g. Bitsch and Bitsch, 2000, 2004; Giribet et al., 2004). Only few studies included all dipluran subgroups. Szucsich and Pass (2008) re-analyzed the dipluran leg articulation and muscles (Manton, 1977) with the inclusion of Projapygoidea. They pointed towards a cautious interpretation of many morphological or developmental characters and suggested comparative reanalyses with modern methods. Using data from literature only bears the danger of missing and misinterpreting characters.

From a molecular perspective, the monophyly of entognathous orders was not challenged. Even diplurans were recovered monophyletic, mostly with strong support (e.g. Wheeler et al., 2001; Kjer, 2004; Luan et al., 2005; Kjer et al., 2006; Mallatt and Giribet, 2006; Misof et al., 2007; Gao et al., 2008; Mallatt et al., 2010). In contrast, order internal relationships have been – and are still – highly discussed, for example the position of the Projapygidae, a subgroup of Diplura, and relationships within Collembola.

Altogether, all non-molecular studies do neither strongly support Ellipura, nor Nonoculata, nor Diplura + Ectognatha. Contradictions between studies show that appropriate criteria for homologization are missing. Currently, we do not have a clear picture from morphological, developmental and neuroanatomical studies, neither of the major euarthropod relationships, nor of the monophyly of hexapods, its possible sister group and, especially, not of relationships among basal hexapod splits.

With respect to molecular studies, there is no clear support for Hexapoda or Entognatha (Giribet et al., 2001; Wheeler et al., 2001; D'Haese, 2002b; Nardi et al., 2003a; Giribet et al., 2004; Carapelli et al., 2005, 2007). Diplura + Ectognatha has not been recovered. Most analyses strongly support Nonoculata (e.g. Luan et al., 2003, 2004, 2005; Kjer, 2004; Kjer et al., 2006; Mallatt and Giribet, 2006; Misof et al., 2007; Gao et al., 2008; Dell'Ampio et al., 2009; von Reumont et al., 2009; Mallatt et al., 2010). However, frequently observed phenomena, for example non-stationarity within data sets, have been ignored or handled by exclusion of taxa.

1.3. Aim and course of this study

Support for the monophyly of Hexapoda, Entognatha, Ellipura or Nonoculata and a proposed sister group relationship of Diplura and Ectognatha is weak. From a molecular perspective, the monophyly of Hexapoda and Entognatha has seriously been put into question. The observed strong support for Nonoculata has been debated as a result of systematic bias. Relationships within Protura, Diplura and Collembola are unresolved. Moreover, major arthropod relationships are not resolved: the placement of myriapods is unclear, and there is no consensus concerning the sister group of hexapods. Present molecular data covering all entognathous orders are mainly restricted to rRNA genes. A phylogenomic approach is missing, because data availability for all primary wingless hexapod lineages is sparse. The inclusion of these orders is crucial to infer deep hexapod and euarthropod relationships.

In this study, two different molecular approaches, a phylogenomic approach based on EST data (chapter 2) and analyses of the nuclear ribosomal RNA markers (chapter 3) are applied to solve the following questions:

- Are Hexapoda monophyletic?

- Which taxon is a potential sister group of Hexapoda?
- Are Entognatha monophyletic?
- Ellipura *versus* Nonoculata?
- Are Ectognatha supported?

Within an 'arthropod' framework, following hypotheses are addressed:

- Atelocerata *versus* Pancrustacea?
- Mandibulata *versus* Myriochelata?

For both approaches, it will be considered which clades are robustly resolved and which clades remain problematic. Methodological aspects will be critically considered in detail. It is addressed, if applying advanced methods pay off with respect to data- and analyses quality.

Course of the study and methodological aspects

A taxon sampling in the sense of Philippe et al. (2005a) has been considered for the present study. Innovative, recently developed methods and sophisticated software have been incorporated for both approaches. The idea is to improve data quality and to correct misleading effects from a methodological point of view. Approaches and tools are introduced in detail in the *Background* and *Methods* of respective chapters.

Chapter 2, Arthropod Phylogenomics, presents the largest known arthropod phylogenomic EST data set. New EST data for all monocondylic, primary wingless hexapod orders have been generated. Additionally, new euarthropod EST projects are included from cooperation partners. An overview of history, techniques, phylogenetic use of EST data and their availability for euarthropods focused on primary wingless hexapods is given in the *Background* section. Closely interwoven with phylogenomic data sets are upcoming methodological challenges to handle large data amounts. Methodological challenges are introduced, for example, orthology prediction of genes, percentage and distribution of data in large data matrices, matrix saturation, overlap of genes and taxa. Applied tools are outlined, especially the reduction heuristics software MARE (Misof et al., *in prep.*), recently developed by B. Misof and our work group. MARE considers signal in genes to optimally select a data submatrix with high total average information content and as much taxa as possible. As well as other tools, it aims to improve data quality prior to tree reconstruction by minimizing noise. Taxon sampling, laboratory work, data processing and analyses setup, the generation of data sets and phylogenetic reconstructions are explained in the *Methods* section. Different arthropod data subsets have been generated (for example, a data subset without proteome taxa and a data subset with exclusively ribosomal respectively non-ribosomal genes) to evaluate a possible impact of different character selection on phylogenetic inference. Additionally, an endopterygote data set and a published metazoan data set (Dunn et al., 2008) have been used to apply the new reduction heuristics. In the *Results* section, the full phylogenomic arthropod data set, its selected optimal subset (SOS) and further data subsets are analyzed. The impact of the new reduction heuristics is comparatively examined with respect to

informativeness, saturation and resolution of inferred trees. Additional data sets are analyzed with respect to the reduction heuristics performance and its impact on the tree robustness. The *Discussion* assesses methodological aspects and observed effects of different data subsets. Phylogenetic relationships are discussed focusing on hexapods and primary wingless orders. Relationships between major euarthropod groups are also addressed.

Chapter 3, Arthropod Phylogeny inferred from nuclear rRNAs, presents a large arthropod data set based on two nuclear rRNA markers, including all primary wingless hexapod orders. Nearly for the past 40 years, ribosomal RNA genes have been used for phylogenetic inference and their properties are well known. They comprise regions which build stems (paired regions) where characters do not evolve independently. Background knowledge has been incorporated in a) the alignment procedure (RNAsalsa Stocsits et al., 2009) and b) in models for phylogenetic reconstruction. Observed non-stationarity (= time-heterogeneity) within the data set has been modeled as well during phylogenetic reconstruction. For this thesis, comparative analyses of the rRNA data set have been conducted: in different combinations (i) secondary structure information and (ii) non-stationarity has been modeled or ignored. In the *Background* section, rRNA markers and their use in phylogenetic inference are introduced. Next follows a short overview on data availability of (nearly) complete nuclear rRNA markers for 'basal hexapods' and myriapods. Previously used approaches are summarized with their strengths and problems. Applied software that considers secondary structure information in alignment and phylogenetic reconstruction and tools for modeling non-stationary processes are introduced. The *Methods* section gives detailed information about taxon sampling, laboratory work and comparative analyses. In the *Results* section, topologies are analyzed considering 'standard' and sophisticated methods (standard DNA *versus* mixed models, ignoring *versus* modeling non-stationarity). The *Discussion* focuses on current hypotheses addressing primary wingless hexapod lineages (including intra-order relationships) with respect to different analyses. Euarthropod relationships are also addressed.

Chapter 4 outlines some future perspectives.

A part of chapter 2 has been published in

Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, 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**(11):2451–2464. doi: 10.1093/molbev/msq130.

A part of chapter 3 has been published in

von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits R, Luan YX, Wägele JW, Pass G, Hadrys H, Misof B (2009): Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. BMC Evolutionary Biology (2009) 9:119. doi: 10.1186/1471-2148-9-119.

This thesis was conducted within the priority program SPP 1174 "Deep Metazoan Phylogeny". The project was funded by the German Science Foundation (Deutsche Forschungsgemeinschaft, DFG), grant MI 649/6.

2. Arthropod Phylogenomics

2.1. Background

2.1.1. ESTs and their use for phylogenetic inference

Inferring hexapod, arthropod or metazoan phylogeny has been based either on single gene analyses (e.g. Aguinaldo et al., 1997; Peterson et al., 2004; Anderson and Swofford, 2004; Luan et al., 2005; Mallatt and Giribet, 2006) or on multi-gene analyses (e.g. Baurain et al., 2007; Bourlat et al., 2008; Regier et al., 2008; Wiegmann et al., 2009; Aleshin et al., 2009; Regier et al., 2010). The establishement for many genes has been extremely time-consuming. The early 1990s brought along the development of sequences obtained from so called "Expressed Sequence Tags" (ESTs) which provide multiple gene data with less effort. ESTs are small pieces of copy DNA (cDNA), usually 200–500 nucleotides long, generated by sequencing one or both ends of an expressed gene (Adams et al., 1991). Copy DNA is received by reverse transcription of cellular messenger RNA (mRNA). This approach was developed by Putney et al. (1983) to detect previously unknown protein coding genes. After construction of cDNA libraries, ESTs have been generated by randomly sequencing clones (Fig. A.1). Bits of cDNA represent genes expressed in certain cells, tissues, or organs. These "tags" provide a rich source of information and are used, for example, for identification of novel genes, gene mapping, comparative genomics and functional characterization of gene products. Large-scale EST sequencing was championed by Craig Venter of the Institute for Genomic Research in Gaithersburg, MD, and was readily taken up by the private sector. Data collections rapidly increased demanding new techniques of databasing. In 1992, the dbEST database (NCBI, `http://www.ncbi.nlm.nih.gov/dbEST/`) was released (Boguski et al., 1993). To date, more than 66,792,500 entries are published (access November 2010). Detailed information can be accessed from NCBI Fact sheets or from Parkinson and Blaxter (2009). A brief historical summary is given in Boguski (1995). In the present study, ESTs were obtained via standard Sanger-sequencing of randomly selected cDNA-clones in 2007 and 2008. In the meantime, this method has been gradually superseded by next generation pyro-sequencing (Hudson, 2008).

The use of ESTs for phylogenetic inference started in the early 21st century. For example, Bapteste et al. (2002) studied phylogenetic affinities of amoebas and related taxa based on ESTs. Later, EST based studies have been published focusing on arthropod subgroups, for example on early pterygote or endopterygote insects (Simon et al., 2009; Savard et al., 2006). ESTs have also been used to infer deep splits like ecdysozoans (Roeding et al., 2007), chordates (Delsuc et al., 2008) or lophotrochozoans (Helmkampf et al., 2008). Philippe et al. (2004, 2005b) used large EST-based alignments to infer metazoan phylogeny and related eukaryotic taxa. Dunn et al. (2008) and Philippe et al. (2009) as well published large scale phylogenies for metazoans. All these studies provided promising prospects that ESTs bear sufficient phylogentic information to resolve the deep metazoan phylogeny (Philippe and Telford, 2006). The fast and extensive growth of EST data comes along with new demands for phylogenetic inference. Addressed questions are for example: How can orthologous genes be properly

identified? How should missing information be treated? What percentage of missing information can be incorporated without leading to biased phylogenetic inference? Are supertree or supermatrix approaches appropriate for phylogenetic inference? Is the signal heterogeneous throughout a data matrix? How should it be handled in large data sets? Which impact do have single species on topological inference? How much overlap between genes and taxa is necessary within a large data set? These are some of the questions that point towwards the development of new methods dealing with large-scale data sets.

Data accumulation demands new methods

Massive data accumulation demanded developments in theoretical and bioinformatical aspects. For example, a denser taxon sampling is recommended as it might break down long branches (Philippe et al., 2005a; Wiens, 2006). Also, genome-scale approaches might overcome contradictory results of single gene analyses (Jeffroy et al., 2006; Philippe and Telford, 2006). With phylogenomic approaches, the impact of stochastic or sampling error is reduced (Philippe and Telford, 2006; Brinkmann and Philippe, 2008). As a consequence, statistical support should become reliable but, large data may also increase systematic error in cases of model misspecification. Thus, accumulation of data is not sufficient to guarantee robust tree reconstruction. Instead, several additional elements must be part of the analysis pipeline, e.g. careful selection of orthologous genes, the consideration of data quality, reduction of the data gappiness and model fitting (Roeding et al., 2007; Dunn et al., 2008; Hartmann and Vision, 2008; Ebersberger et al., 2009; Philippe et al., 2009). This has recently been shown in phylogenomic analyses (e.g. Roeding et al., 2007; Dunn et al., 2008) and (Philippe et al., 2009).

The distinction between orthologous and paralogous genes (Fitch, 1970) is essential in molecular analyses, especially for phylogenomic data sets (Roure et al., 2007). Prediction of orthologous genes, however, is a difficult task (Koonin, 2005). Orthologous genes originate from shared ancestral genes from a last common ancestor and are related via specification events. They reflect the organismal phylogeny. Paralogous genes arise from gene duplication events (Fitch, 2000). They should be excluded from phylogenomic analyses as they imply the presence of multiple copies of a given gene per species, and some do not reflect organismal phylogeny. Recent studies started to distinguish between co-orthologs, in- and outparalogs (Sonnhammer and Koonin, 2002). Genes in one lineage that are collectively orthologous to one or more genes in another lineage, due to a lineage-specific gene duplication, are coined co-orthologs. Inparalogs are paralogous genes resulting from a lineage-specific gene duplication after a speciation event. Outparalogs are paralogous genes resulting from a duplication prior to a speciation event. Latter can never be orthologous and infer exclusively gene trees (O'Brien et al., 2005). Instead, species trees are crucial for phylogenetic inference (Rannala and Ziheng, 2008). Several published strategies for identification of orthologs have been evaluated by Chen et al. (2007). Most recent approaches have been developed by Schreiber et al. (2009) and Ebersberger et al. (2009).

The problem of gappiness in EST data has been addressed by several authors which propose different ways for handling this problem (Wiens, 1998, 2003, 2005, 2006; Philippe et al., 2004, 2005a; Yan et al., 2005; Sanderson, 2007; Hartmann and Vision, 2008; Wiens and Moen, 2008; Cotton and Wilkinson, 2009). Either supertree (e.g. Sanderson, 1998; Bininda-Emonds, 2004) or supermatrix methods (e.g. De Queiroz and Gatesy, 2006) have been proposed which both have their strengths and weaknesses. Supertree techniques rely on a separate tree search for each gene (Bininda-Emonds

et al., 2002; Wilkinson et al., 2005; De Queiroz and Gatesy, 2006; Steel and Rodrigo, 2008; Cotton and Wilkinson, 2009). Resulting gene trees are summarized with supertree consensus techniques. The supertree approach can fail as single genes might not contain sufficient signal to resolve targeted relationships, leading to unresolved supertrees. The problem of how to combine trees into a supertree is not satisfyingly solved, despite recent progress (Holland et al., 2007; Steel and Rodrigo, 2008; Cotton and Wilkinson, 2009). In simulation studies, the supermatrix approach consistently outperformed the supertree approach (De Queiroz and Gatesy, 2006). Its advantage is the fact that tree inference relies on maximally collected data/signal (De Queiroz and Gatesy, 2006). The reliability of trees can be addressed, for example with bootstrapping or posterior probabilities on the complete data set (Felsenstein, 2004; De Queiroz and Gatesy, 2006). On the other hand, supermatrices often show data sparseness. Only 10% of the sequences or even less might actually be present within such a supermatrix (e.g. Driskell et al., 2004; Thomson and Shaffer, 2010). Nowadays, supermatrices of more than 100 taxa and genes are concatenated from available data (Sanderson and Driskell, 2003; Philippe et al., 2004, 2009; Delsuc et al., 2005; De Queiroz and Gatesy, 2006; Hausdorf et al., 2007; Dunn et al., 2008; Bourlat et al., 2008; Galtier and Daubin, 2008; Regier et al., 2008; Smith et al., 2009; Schierwater et al., 2009). The majority of these data sets shows a large percentage of missing data. In the present study, the supermatrix strategy is applied.

Given a sparse data availability, simulations have shown that 20–30% saturation of a supermatrix might suffice to reconstruct phylogenetic relationships correctly (Wiens, 2003). However, these simulations have been performed assuming a similar strength of phylogenetic signal of each marker which is clearly unrealistic. From an empirical point of view, heterogeneity of signal should be assumed and incorporated in simulation studies that face proper methods to handle large data sets. In order to improve the reliability of tree reconstructions, predefined thresholds of data availability are frequently used to select a subset of taxa and genes maintaining a supermatrix approach (e.g. Dunn et al., 2008; Philippe et al., 2009). Dunn et al. (2008) showed that removing least saturated taxa and increasing saturation of their supermatrix above 50%, improved overall bootstrap support, but predefined thresholds appear arbitrarily chosen. Several publications tested the impact of missing data (e.g. Wiens, 2003, 2006; Sanderson and Driskell, 2003; Wiens and Moen, 2008). Related to a high percentage of missing data, methods have been established to search for a maximal or sufficient gene overlap between taxa. Suggestions have been made to solve these problems, e.g. with maximal-biclique enumeration algorithms (Alexe et al., 2002; Sanderson et al., 2003) that identify complete subsets of taxa and genes where each sequence is available. Since many data matrices are lowly saturated (e.g. $< 10\%$, see Sanderson and Driskell, 2003; Driskell et al., 2004; Thomson and Shaffer, 2010), these algorithms often lead to very small subsets of genes and taxa. Alternatively, quasi-biclique approaches (Yan et al., 2005) allow a predefined threshold of missing data while the connectivity between taxa and genes is guaranteed. This usually leads to enlarged data subsets where much more information is retained. Data matrices are interpreted as bipartite graphs, and algorithms seek to optimize the connectivity of bicliques or guarantee the connectivity of quasi-bicliques during the reduction to submatrices. Currently, the application of both approaches on large data sets is restricted due to NP-complete computational problems (Dawande et al., 2001; Dias et al., 2007).

Up to now, neither studies with predefined thresholds nor algorithms dealing with gene overlap between taxa have considered a) signal heterogeneity within the data set or b) information content (IC) of single genes and taxa to select 'optimal' data matrices. All approaches work exclusively

with presence|absence matrices. This thesis is the first empirical study that applies a new approach of reduction heuristics of supermatrices to select an optimal data (sub)set with high information content. The applied approach, MARE (MAtrix REduction), has been developed by our work group (Misof et al., *in prep.*). It incorporates information content reflecting potential signal of each gene within a large data matrix. Subsequently, reduction heuristics implements this information content which is based on geometry quartet mapping (Eigen et al., 1988; Nieselt-Struwe, 1997; Nieselt-Struwe and von Haeseler, 2001; Grünewald et al., 2007). An optimal subset of taxa and genes (SOS) with high information content (IC) is selected. Thereby as much taxa as possible are retained towards a balanced taxon-gene ratio.

Phylogenomics to infer arthropod phylogeny

Most phylogenomic approaches considering arthropod relationships have been focused on subtrees, for example on Pancrustacea (Timmermans et al., 2008; Aleshin et al., 2009) or on endopterygote insects (Wiegmann et al., 2009). Others considered arthropod relationships only in a wider context, for example addressing Ecdysozoa (Roeding et al., 2007). Studies with a substantial arthropod taxon sampling mainly exist for single gene analyses (e.g. rRNA genes, Mallatt and Giribet, 2006; Dell'Ampio et al., 2009). The largest arthropod data sets were published by Regier et al. (2008) and Regier et al. (2010). In Regier et al. (2008), a phylogeny was inferred from 56 euarthropod taxa plus six tardigrades. Amplification of all 68 gene regions was successful for only 13 panarthropod taxa. Thus, the data matrix considering all taxa covered 68 gene fractions for these 13 panarthropods plus three genes for remaining 49 species. The data set comprises 41 kb protein coding genes (nucleotide level), with 71% missing data. Regier et al. (2010) complemented their previous data set and relied on selected 62 nuclear protein coding genes for 75 arthropod taxa. However, proturans are missing in their analyses. At amino acid level, the data set is relatively small (around 13,000 amino acid positions). Some splits like the proposed sister group relationship of hexapods and 'Xenocarida' (= Remipedia + Cephalocarida), remain ambiguous. For both studies of Regier et al., data were obtained by gene amplification with specially designed targeted primers of prior selected, single genes from cDNA. The authors have worked on this data set since 2001 and such an approach seems very time consuming with a great laboratory effort (`http://www.biology.duke.edu/cunningham/DeepArthropod.html`).

Here, a phylogenomic EST-based approach is presented with the currently largest arthropod data set. The present data set covers all major arthropod clades and comprises 233 taxa of which 222 species are euarthropods. Data of four new EST projects from apterygote hexapods (all entognathous orders and bristletails) have been successfully generated and are included in this study. It is crucial to cover at least all entognathous orders when addressing questions of basal euarthropod splits (see section 2.1.2). Data of nine new EST projects covering velvet worms, myriapods, sea spiders, euchelicerates, non-malacostracan crustaceans and pterygote insects were kindly provided by cooperation partners within the SPP 1174 (Burmester/Roeding, Biozentrum Grindel, University of Hamburg; Wägele/v. Reumont, ZFMK Bonn and Hadrys/Simon, ITZ Hannover, (Tab. A.1). Recently developed methods for orthology prediction (HaMStR, Ebersberger et al., 2009) and alignment masking (Aliscore, Misof and Misof, 2009; Kück et al., 2010) have been used. New reduction heuristics (MARE) has been applied. The full data set comprises 233 taxa and 775 genes. The selected optimal subset (SOS) spans more than 37,000 amino acid positions, covers 129 genes and more than 100 euarthropod species.

2.1.2. Available and new EST data for arthropod phylogenomics

Internationally, EST analyses with a phylogenetic background are under way for example in England, Laboratory of A. Vogler, Sillwood College (pers. comm.) and the US, Whiting Lab, Provo, Utah (pers. comm.), hoping to derive a robust Insect Tree of Life. In the last two years, the number of arthropod EST projects increased substantially, but with a strong bias towards endopterygote insects. The data availability of apterygote hexapods, non-endopterygote insects, non-malacostracan crustaceans and myriapods is still poor. Within primary wingless hexapods, three of four orders, Protura, Diplura and Archaeognatha, are missing. Further on, the number of sequenced contigs is very heterogeneous (dbEST, NCBI). Five EST projects have been published on apterygote hexapods (access January 2009): four springtails, *Orchesella cincta* (28 ESTs), *Folsomia candida* (8,703 ESTs), *Onychiurus arcticus* (16,379 ESTs), *Cryptopygus antarcticus* (1,180 ESTs) and one silverfish, *Tricholepisma aurea* (425 ESTs). Published EST data on springtails comprise either highly derived species (*Folsomia candida*) or species adapted to an ecologically extreme habitat (*O. arcticus* or *C. antarcticus*). Only two EST projects have been published on myriapods, (*Julida* sp. APV-2005, Diplopoda, 453 ESTs and *Scutigera coleoptrata*, Chilopoda, 2,400 ESTs).

The arthropod tree of life cannot be reliably reconstructed without covering primary wingless hexapods or myriapods. Especially, data of Protura, Diplura and Collembola must be characterized if hypotheses concerning hexapods are addressed. Data sets which cover all entognathous hexapods, however, are very sparse and restricted to rRNA genes (e.g. Luan et al., 2005; Gao et al., 2008; Dell'Ampio et al., 2009; von Reumont et al., 2009; Mallatt et al., 2010). Likewise crucial is an increased data availability of myriapods.

For the present study, specimens of four representative species, each covering a primary wingless hexapod order (Fig. 2.1), were sampled to generate EST libraries: *Acerentomon franzi* (Protura), *Campodea* cf. *fragilis* (Diplura), *Anurida maritima* (Collembola) and *Lepismachilis y-signata* (Archaeognatha). Data of all four species have been successfully included in the EST analyses (Tab. 2.1). Sequences have been deposited in EMBL and GenBank (NCBI).

2.1.3. Orthology prediction – HaMStR

Paralogous genes arise from gene duplication and are not appropriate for phylogenetic inference because they potentially are in conflict with species trees. Since the prime goal of building phylogenetic trees is to decipher the evolutionary relationships among organisms based on their shared common ancestry, only orthologous sequences should be used. Several different approaches to identify orthologous sequences have been proposed (Li et al., 2003; Schreiber et al., 2009; Ebersberger et al., 2009). The technique of reciprocal BLAST has received the most attention (Chen et al., 2007). Dunn et al. (2008) also used a sort of reciprocal BLAST to identify orthologous sequences. In the present study, HaMStR (Hidden Markov Model based Search for Orthologs using Reciprocity) has been applied (Ebersberger et al., 2009). HaMStR uses trained profile Hidden Markov Models (HMMs Durbin et al., 1998) and a reciprocal blast criterion (Altschul et al., 1997) to identify orthologous sequences. HaMStR is embedded in an analysis pipeline and was established by the CIBIV, Vienna within the priority program SPP 1174.

Figure 2.1.: Species used for EST projects. Specimens were sampled and preserved for RNA extraction in 2007 and 2008. Photographs do not reflect the original specimens size. **a** Three specimens of *Acerentomon franzi*, Protura (ca. ~2.5 x 0.5 mm), © J. Dambach & K. Meusemann. **b** Specimen of *Campodea fragilis*, Diplura (~5 x 2 mm), © University of Bratislava, Slovakia, Zoological Department (`http://zoology.fns.uniba.sk/poznavacka/images/31_Campodea_fragilis.jpg`). **c** Specimen of *Anurida maritima*, Collembola (~3 x 2 mm), © Steve Hopkin (`http://www.stevehopkin.co.uk`). **d** Specimen of *Lepismachilis y-signata*, Archaeognatha (~1.8 x 0.5 cm), © A. Staudt (`http://www.delattinia.de/GM/GM_Nied.htm`).

2.1.4. Reduction heuristics – MARE: a new approach for gene- and taxa selection

A sparse matrix saturation is mostly observed within large supermatrices which yield many genes and many taxa (e.g. Driskell et al., 2004). This prompts several technical questions: How can genes and taxa be properly selected? Which percentage of data relative to missing information is necessary to avoid systematic error? Possible solutions might be dependent on the data set and its traits, like taxon choice, signal heterogeneity within the matrix, distribution of (missing) data, etc. For simulations with 50 x 50 supermatrices (10–30% saturation and heterogeneous signal among genes) it has been shown

that correct trees cannot be reconstructed below 30% saturation (Misof et al., *in prep.*). Most large data sets show less saturation prior to selection (e.g. Driskell et al., 2004; Dunn et al., 2008; Simon et al., 2009). None of published phylogenomic study provides information, about signal of selected genes. To handle this issue, Misof et al. (*in prep.*) developed an alternative to existing approaches. The idea behind the new approach is to reduce effects of under-sampling of taxa and genes, and filter genes with no (or poor) signal and taxa before tree reconstruction. With this strategy, a subset of the concatenated supermatrix is selected to receive a maximally informative set of taxa and genes. Prior to selection, potential information content of each single gene (partition) is calculated (Fig. 2.2) within a superalignment. This is conducted with generalized geometry quartet mapping (Nieselt-Struwe and von Haeseler, 2001) at an amino acid level. Each gene obtains a value of information content between 0.0 and 1.0, reflecting the relative number of resolved quartet trees. A data availability matrix (presence|absence matrix) is then transformed into a matrix of potential information content of each taxon and gene by multiplying the availability (0|1) with scores of information content (Fig. 2.3). The information content of each gene is calculated as the average value over all taxa including missing data. The total average information content P of a supermatrix is calculated as the sum of the relative information content of all genes in relation to the number of taxa. Secondly, a simple hill climbing procedure is used to select an optimal subset (SOS) of taxa and genes with high total average information content (IC). Reduction starts with dropping either the taxon (row) or the gene (column) with the lowest average IC, generating a new matrix. In case of ties, genes are excluded. Taxa or genes with the lowest average IC will be stepwise discarded from the matrix, receiving a submatrix with increased total average information content P'. In order to reach an optimum, the underlying optimality function takes into account that size reduction and low total average IC are penalized (see below). Thereby, the connectivity between taxa is monitored by a minimum number of overlapping genes and taxa, previous to every reduction step. In terms of the graph theory, a connected graph must be drawn between each subset (cluster) of nodes (genes). Nodes (or clusters) are only connected, if the overlap criterion is fulfilled. The optimality function favors reduction of matrices to high total average IC. The selected optimal subset (SOS) is a reduced supermatrix, corresponding to one possible solution of the quasi-biclique approach. The total average IC and the matrix saturation (= number of present gene entries in relation to the total size of the matrix) was increased tremendously for the SOS of the original 'arthropod' supermatrix of this study and comprised a fair balanced taxon-gene ratio. In a final step, the original superalignment is rewritten based on the SOS. Constraints (taxa, genes, or both) can be defined and will not be dropped from the matrix. The average information content of taxa can optionally be weighted (option -t). The algorithm has been implemented in the software MARE (MAtrixREduction), written in *C++*. Details on the algorithm will be published elsewhere (Misof et al., *in prep.*). The heuristics algorithm of MARE is time efficient and easily applicable to matrices of $>$ 100 x 100, in contrast to quasi-biclique approaches. This preprocessing of the data helps to considerably reduce the effort spent in tree reconstructions. It also opens a route to assess whether it is worthwhile to include new taxa or genes in a pre-existing supermatrix based on their contribution to the total average information content of the supermatrix without using tree reconstructions. Simulations have shown that the chance to reconstruct the correct tree increases tremendously after matrix reduction (Misof et al., *in prep.*). This will be tested in the present study on real data sets. Additionally, data subsets covering endopterygote insects and the data set of Dunn et al. (2008) have been used to study performance of reduction heuristics and its impact on tree

reconstructions compared with original data sets.

MARE: pre-alpha version

A pre-alpha version of MARE (three Perl scripts) was used , available from upon request from mail2mare@gmx.de. Information content (IC) for each gene and taxon is calculated based on $\binom{n}{4}$ but maximal 20,000 taxa quartets using the BLOSUM62 substitution matrix. Genes containing less than four sequences and taxa containing less than 1/3 of a single gene sequence are considered as absent. Currently, the optimality function $f(P)$ penalizes size reduction of an original matrix M and low total average information content P of a reduced matrix M' as follows:

$$f(P) = 1 - |(\lambda - P^{\alpha \times (1-P)})| \text{ if } P < 1$$

with α as a scaling factor (default $\alpha = 3$), λ as the size ratio between reduced M' and original matrix M (matrix size = # taxa x # genes). P is maximized, if $P = 1$ reduction stops. Total connectivity is monitored: one gene must share a minimum number of two taxa with another gene, building a cluster. Each cluster must at least have an overlap of two taxa with another cluster, thus one connected graph can be drawn throughout the matrix. After each reduction step, the current size ratio λ, the average information content $\bar{p}$ of taxa and genes of M' are recalculated, M' is resorted and the total average information content P of M' is recalculated as well. During reduction, taxa can be weighted by 5/3, invoking the -t option. The total average information content P and the matrix saturation (genes with an IC < 0.04 are considered as missing) are provided as output for the original data matrix and the selected optimal data subset (SOS). Finally, the original superalignment is rewritten based on the selected optimal submatrices.

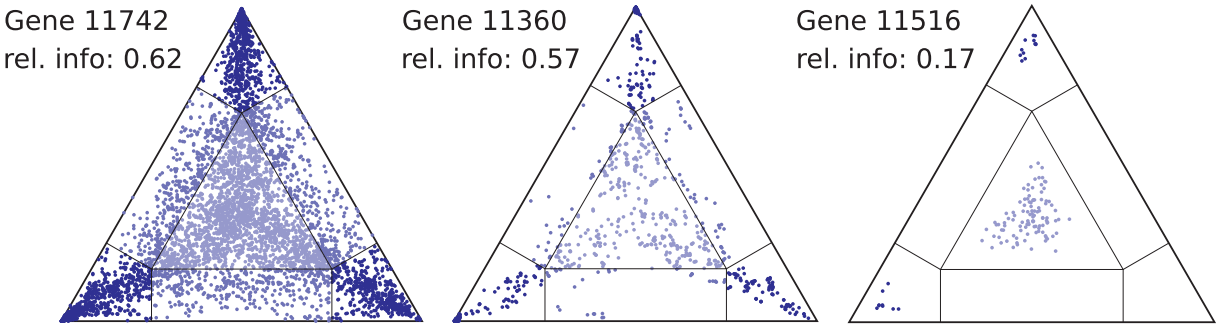

Figure 2.2.: Potential information content of three genes from the original arthropod data set AP_1 (Tab. 2.3). The information content (IC) is visualized by 2D simplex bipartite graphs. The potential IC of a gene is defined as the relative tree–likeness of the data using geometry mapping (Nieselt-Struwe and von Haeseler, 2001). Relative tree-likeness corresponds to the relative frequency of simplex points within the outer areas of (partially) resolved trees. Genes containing less than four sequences and taxa containing less than 1/3 of a gene sequence are considered as absent.

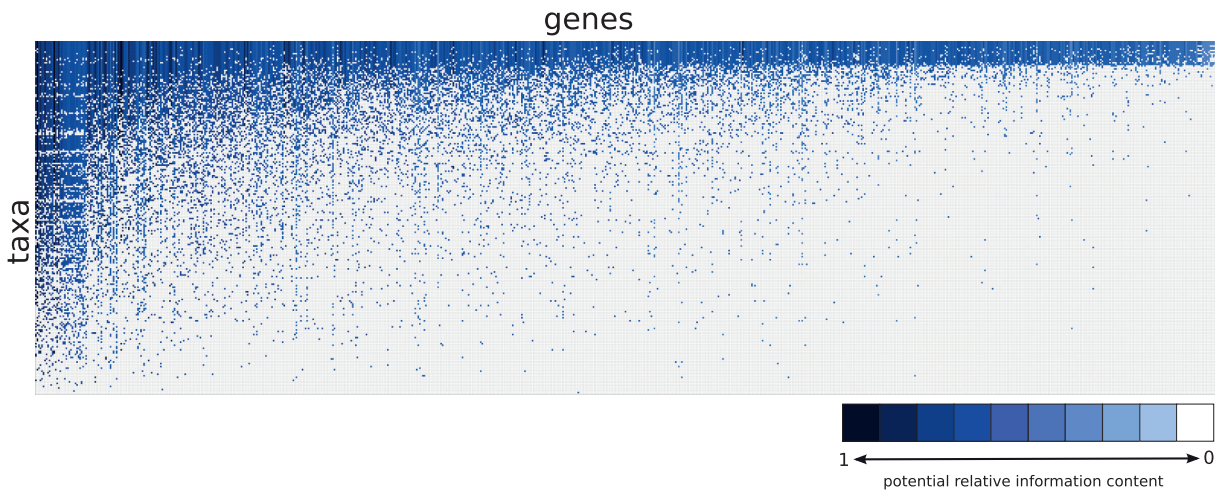

Figure 2.3.: Original data matrix with potential information content of each gene and taxon, original matrix (arthropod data set `AP_1` with 233 taxa and 775 genes, see Tab. 2.3). An availability matrix (presence|absence) is transformed into a matrix with information content (IC) by multiplying availability (0|1) with scores of informativeness. Taxa / genes with a high IC are located on top, left; Taxa / genes with a low IC are located on bottom, right. X-axis: genes, Y-axis: taxa. Information content of genes ranges from 0.0–1.0 (10 units), color coded from dark blue ($>$ 0.9–1.0) to white (IC of $\geq$ 0–0.1 or missing data). Genes with a IC $<$ 0.04 are considered as absent. Total average IC of the matrix: P 0.1, matrix saturation: 17.6%.

2.2. Materials and Methods

2.2.1. Taxon sampling and preservation of apterygote hexapods

Specimens from four primary wingless hexapod orders were sampled for EST projects in 2006, 2007 and 2008: Protura, Diplura, Collembola and Archaeognatha. Sampling sites were located in Germany, Austria and the Netherlands. Several hundred specimens of Diplura and Collembola, 20 bristletails and more than 1,000 specimens of Protura were required to extract enough total RNA. Diplura and Collembola were collected using a small exhauster. Proturans were captured using Berlese traps. Bristletails were manually sampled using collection vials. To ensure that specimens belong to the same species, 30 individuals were preserved in ethanol and determined afterwards. Sample localities were as small as possible due to syntopy of species of Protura and Diplura. Stress at the time of specimen-preservation, which can evoke a strong shift in gene expression towards stress-response proteins, was avoided as possible. A strong expression of particular genes is not appropriate for phylogenetic inference, since as much protein coding genes that usually are expressed should be covered. Unlike DNA, RNA rapidly degrades due to enzymatic RNAses. To avoid degradation, liquid nitrogen was used for deadening specimens. The dying process is much faster than using RNAlater; all degrading enzymes are immediately inhibited. Nitrogen was used, at least, for empirical reasons. In an earlier approach, a preservation of diplurans in RNAlater was unsuccessful: the extracted RNA was highly degraded and not appropriate for cDNA library construction.

Protura, Hexapoda

Proturans are eudaphic and live in soil and leaf litter. One of the largest representative is *Acerentomon franzi* Nosek, 1965 (Acerentomidae, Protura) with a length of ca. 2.5 mm and width of ca. 0.5 mm. Specimens were sampled within a 3-month project in fall 2007 at the University of Vienna, Austria. Sampling was supported by the work group of Prof. G. Pass and Dr. M. Walzl. Leaf litter and upper

soil layers were sampled in Carinthia, Austria. Living specimens were extracted via 40 Berlese traps (Fig. 2.4). Around 2,000 soil samples were checked for *Acerentomon franzi*. Every specimen was checked on the belonging to *A. franzi* with binoculars. Then, specimens were carefully transferred into 1.7 ml DNAse/RNAse free Eppendorf safe lock tubes with a single brush hair. Subsequently, they were shock frozen in liquid nitrogen. 1,066 adult specimens and more than 2,000 juvenile specimens were finally stored at −80 °C. Only adult specimens were used for RNA extraction.

Diplura, Hexapoda

Campodea fragilis Meinert, 1865 (Campodeidae, Diplura) is ca. 3–5 mm long and characterized by its thin, flexible and nearly translucent body. Antennae and tail appendages are long and multi-segmented and easily break during sampling. In summer 2007, ca. 600 specimens of *Campodea* cf. *fragilis* were collected out of leaf litter and loamy soil of an old deciduous forest in Friesdorf (near Bonn, Germany) with an extra small exhauster. Living specimens were transported to the lab, preserved in liquid nitrogen and stored at −80 °C.

Collembola, Hexapoda

In August 2006, around 400 specimens of *Anurida maritima* (Guérin-Méneville, 1836) (Neanuridae, Collembola) were collected at Texel, Netherlands, at a small bay near the ferry port (Fig. 2.5). *A. maritima* is a cosmopolitan springtail of the intertidal zone, often found in aggregations up to several hundred individuals on rocky shores or tidal marshes. They move in rhythm with the tidal cycle. Specimens were collected with extra small exhauster. Living specimens were transferred to the lab and kept alive at 4 °C until shock freezing. They were stored at −80 °C until RNA was extracted.

Archaeognatha, Hexapoda

20 specimens of the bristletail *Lepismachilis y-signata* Kratochvil 1945 (Archaeognatha), were sampled using small collecting vials near Breitenfurt, Austria. This species, 5–10 mm long, belongs to the family Machilidae. It is characterized by its eponymous dark, y-shaped pattern in the compound eyes. Specimens were kept alive in a wetted box until they were shock frozen. Specimens were kindly provided by Nikola Szucsich, Group Pass, University of Vienna, Austria.

2.2.2. Laboratory work

Preservation of specimens was conducted at the molecular lab of the ZFMK and at the Department of Evolutionary Biology, University of Vienna, Austria. Specimens of Protura, Diplura, Collembola and Archaeognatha (section 2.1.2) were preserved in liquid nitrogen and stored at −80 °C. Extraction of total and messenger RNA, construction of cDNA libraries, cloning and sequencing was conducted at the MPIMG, AG Reinhardt/Kube, Berlin, Germany. Total and mRNA was prepared with standard kits (Tab. 2.1). cDNA libraries were constructed using CloneMiner (Invitrogen) or Creator SMART (Clontech). For all species, except *A. maritima*, primer extension cDNA libraries were constructed. For *A. maritima*, a long distance cDNA library was generated. From cDNA libraries, ESTs were sequenced using randomly selected cDNA-clones with Sanger sequencing (Sanger et al., 1977), from the 5' end on a capillary sequencer system ABI 3730XL (Applied Biosystems, Darmstadt, Germany) using BIGDYE chemistry, Applied Biosystems (for details of all new EST data refer to Tab. A.1).

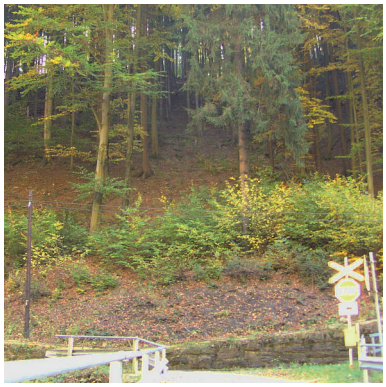

Figure 2.4.: Sample locality and extraction of *Acerentomon franzi* (Protura). From left to right: Sample locality in Carinthia, Austria, northern escarpment of a deciduous forest. Berlese-traps with leaf litter: leaf litter slowly desiccates from top to bottom; specimens creep downwards and fall into a plastic box with gypsum; boxes are placed in plastic bags damped with water. Box with a single specimen of *A. franzi*.



Figure 2.5.: Sample locality of *Anurida maritima* (Collembola). From left to right: Ferry port on Texel, Netherlands. Bay near the ferry port: specimens of *Anurida* have been sampled using a small exhauster, ca. one hour after high tide. Anurida-specimens creeping on Algae.

Table 2.1.: New EST projects of the present study. No. of EST raw data: number of EST raw sequences before processing and clustering.

Species, Group	RNA extraction	cDNA library construction	No. of EST raw data
Acerentomon franzi Hexapoda, Protura	Absolutely RNA (Stratagene)	CloneMiner (Invitrogen)	4,600
Campodea cf. *fragilis* Hexapoda, Diplura	Trizol (Invitrogen), Absolutely RNA (Stratagene)	CreatorSMART (Clontech), CloneMiner (Invitrogen)	8,375
Anurida maritima Hexapoda, Collembola	Trizol (Invitrogen)	CreatorSMART (Clontech)	4,391
Lepismachilis y-signata Hexapoda, Archaeognatha	Absolutely RNA (Stratagene)	CloneMiner (Invitrogen)	4,895

2.2.3. Sequence processing, contig assembling and prediction of orthologous genes

Sequence processing, quality check, EST contig assembling and prediction of orthologous genes took place at the CIBIV, University of Vienna, Austria. Published EST data were downloaded from dbEST (NCBI) or the NCBI Trace Archive. Own EST data, EST data from cooperation partners and published ESTs for euarthropods plus onychophorans, tardigrades and selected species of nematodes, annelids and mollusks (Tab. A.2) were mined and processed in four steps: preprocessing, processing, orthology prediction and annotation (Fig. A.2). (1) New EST sequences were screened for vectors and poly-A tails using the software LUCY (Chou and Holmes, 2001). All sequences were screened for contamination by comparison against UniVec (`http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html`) with Crossmatch (Green, 1996) and SeqClean (`http://compbio.dfci.harvard.edu/tgi/software/`). SeqClean screened for poly-A tails as well. Sequences with less than 100 nucleotides were excluded from subsequent processing. Repetitive elements were masked with RepeatMasker (Smit et al., 1996-2004) using Repbase (Jurka et al., 2005). Between 4,373 and 8,253 EST sequences of apterygote hexapods have been processed from cDNA libraries (Tab. 2.2). (2) Overlapping EST reads were assembled into contigs for each species using TGICL (Pertea et al., 2003). Own ESTs were quality clipped with LUCY and clustered a second time to obtain and keep longer contigs. For three species (Tab. A.2), published contigs were directly taken from the Gene Index project (`http://compbio.dfci.harvard.edu/tgi/tgipage.html`). Finally, all contigs were translated at an amino acid level in all reading frames prior to orthology prediction with HaMStR (Ebersberger et al., 2009). (3) For HaMStR, a set of 13 reference proteomes (*Daphnia pulex*, *Tribolium castaneum*, *Bombyx mori*, *Aedes aegypti*, *Apis mellifera*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Capitella* sp., *Lottia gigantea*, *Homo sapiens*, *Tetraodon nigroviridis* and *Xenopus tropicalis*) was compiled from InParanoid (Remm et al., 2001; Berglund et al., 2008, `http://inparanoid6.sb.su.se`). From these 'primer taxa', multiple sequence alignments of 'core' orthologs were generated. These alignments were used to train profile Hidden Markov Models (pHMMs) to search protein hits in each taxon. A reciprocal BlastP (Altschul et al., 1997) decided about a survival of a hit. For a subsequent re-blast step, the presumably evolutionary closest primer taxon (proteome species) was chosen. In total, 244 species were 'hamstred' of which 28 species comprised a complete proteome. 775 genes were identified as putative orthologs. (4) EST contigs were annotated using BlastX (Altschul et al., 1997) against NCBI's non-redundant protein database. Protein sequences of the 25 best hits per contig were aligned with GeneWise (Birney et al., 2004). This software takes frameshifts into account by a special scoring procedure. Each contig was annotated according to the protein sequence with the highest GeneWise score. All ESTs sequences of present projects have been deposited to EMBL and GenBank (Tab. 2.2).

Among 'primer' taxa, three vertebrate species had been included to train pHMMs for orthology prediction. Vertebrates and multiple *Drosophila* species were not relevant for addressed phylogenetic questions. Subsequently, vertebrates and eight *Drosophila* 'proteome' species were excluded from further processing for computational reasons. The complete data set (`AP_1`) comprised 233 taxa and 775 orthologous genes (Tab. A.2, A.3). This was the reference data set for all further arthropod and endopterygote data sets used in this study (Tab. 2.3).

Table 2.2.: Overview of new EST data of apterygote orders. No. of EST proc.: number of processed ESTs reads; No. of EST contigs: number of assembled EST contigs; No. of orthologs: number of putative orthologous genes identified by HaMStR.

Species	Order	No. of EST proc.	Accession numbers	No. of EST contigs	No. of orthologs
Acerentomon franzi	Protura	4,565	FN186135-FN190445	1,995	99
Campodea cf. *fragilis*	Diplura	8,253	FN203025-FN211277	6,407	150
Anurida maritima	Collembola	4,373	FN190447-FN194819	3,504	131
Lepismachilis y-signata	Archaeognatha	4,854	FN219557-FN224410	2,288	123

2.2.4. Alignment and alignment masking

Arthropod and endopterygote data sets

At amino acid level, all genes were separately aligned with the software MAFFT *L-INSI* (Katoh and Toh, 2008) for respective data sets. Within data set `AP_1`, all proteome taxa (17 species) were present in all 775 genes. As proteome species represented a considerable percentage of the overall matrix saturation in data set `AP_1`. A data set without proteome species was generated (`AP_3_oP`). Therefore, proteome species were excluded from every single alignment with a Perl script, 22 genes with only proteome species were excluded. (`AP_3_oP`) comprised 216 taxa and 753 genes. For analyses, genes were realigned.

Out of data set `AP_1`, two endopterygote data sets were designed. They comprised all endopterygote taxa plus a) ensiferans (Orthoptera) as outgroup (together 111 taxa), or b) all ensiferans + *Daphnia pulex* (Branchiopoda, Crustacea) as outgroup, comprising 112 taxa (Tab. 2.3). All genes were realigned after discarding non-considered species from alignments. (`AP_3_oP`) and endopterygote data sets were used to examine performance of reduction heuristics (MARE) and and subsequent phylogenetic inference. Additionally with respect to the endopterygote data sets, it was aimed to infer the position of Hymenoptera which is still matter of debate (see Castro and Dowton, 2005; Savard et al., 2006; Wiegmann et al., 2009).

Alignment masking

In the present study, Aliscore (Misof and Misof, 2009; Kück et al., 2010, `http://aliscore.zfmk.de`) was used to identify randomly similar sections in alignments. Randomized similar sections were identified from all single gene alignments at amino acid level based on the BLOSUM62 matrix, default window size and maximal number of pairwise comparisons, separately for each data set. Only sequences with more than one half of the respective alignment length were included in Aliscore analyses. All sections scored as randomly similar were discarded with ALICUT (Kück, 2009, available from `http://www.utilities.zfmk.de`). Masked gene alignments from the respective data sets were concatenated into masked superalignments with a Perl script (Fig. A.3). Additionally, the script provided charset files with range information for each gene. Superalignments and charset files served as input for MARE.

2.2.5. Selection of taxa and genes: MARE

Reduction heuristics (MARE) was applied with different settings on arthropod data sets `AP_1` and `AP_2`, on data set `AP_3_oP` without proteome taxa, on endopterygote data sets and on the data set of Dunn et al. (2008). The idea was to obtain selected optimal data subsets (SOS) with increased total average information content and matrix saturation (see Tab. 2.3).

Creating optimized data subsets and MARE applications

In this study, the pre-alpha version of MARE was used (see section 2.1.4). Initially, information content for each gene and taxon was calculated. Respective for each data set, a taxa *versus* gene matrix with scores of informativeness was generated. Secondly, reduction heuristics for respective data sets was performed to select optimal data subsets (SOS). For some data sets, taxa were weighted (-t option, see Tab. 2.3). Constraint taxa were defined: the copepod *Tigriopus* (Crustacea) and the centipede *Scutigera* (Myriapoda) were defined as constraints for data set `AP_1`. Copepods are proposed as a possible sister group to Hexapoda (Mallatt and Giribet, 2006; von Reumont et al., 2009). *Scutigera* was the only representative of centipedes. Therefore, matrix reduction was constraint to retain both species as important taxa in selected subsets. Additionally, the velvet worm *Epiperipatus* sp. (Onychophora) was defined as constraint for data set `AP_2`. The idea was to examine the impact of including another velvet worm in phylogenetic inference, since data availability is sparse for this phylum. Finally, original superalignments were rewritten based on the selected optimal submatrices. SOS superalignments were used for tree reconstruction (see Fig. A.3).

Additionally, MARE was used for calculation of the information content (IC) and matrix saturation without reduction (option -j): original matrices are resorted from highest to lowest IC and saturation. This was done for all original data sets to examine heterogeneity of signal among genes, percentage and distribution of present|absent data and a possible impact on the reduction performance of MARE. Distribution of present|absent data strongly differ in published data matrices (compare Driskell et al., 2004; Dunn et al., 2008), ranging from a Gaussian or a random to power-law distribution (most genes are covered by few taxa or vice versa). The distribution of present|absent data might impact i) reduction performance of MARE and ii) tree inference. The IC among genes is often heterogeneous within a data matrix. Remarkable differences of the IC between different genes and different taxa can be observed in real data sets (Driskell et al., 2004).

Additional data sets

Based on the SOS of `AP_1`, three additional data subsets, `AP_op_un`, `AP_1_ri` and `AP_1_nri`, were designed (see Tab. 2.3).

For `AP_op_un` all gene alignments, but unmasked, of the SOS of `AP_1` were concatenated to an unmasked superalignment (unmasked SOS). For this unmasked SOS alignment, information content and matrix saturation were recalculated. The unmasked and masked SOS of `AP_1` were compared using split analyses (section 2.2.6 and 2.3.2).

`AP_1_ri` and `AP_1_nri` were created by separation of all masked ribosomal protein coding (`ri`) and non-ribosomal protein coding genes (`nri`) from the `AP_1`-SOS. After concatenation of respective (ribosomal and non-ribosomal) genes, total average information content and matrix saturation were calculated for both data subsets. `AP_1_ri` and `AP_1_nri` were used i) for tree reconstruction and b)

to visualize potential topological inconsistencies between ribosomal and non-ribosomal gene tree in a consensus network (see section 2.3.2, Fig. 2.23).

Table 2.3.: EST data sets used for MARE. The calculation of the information content (IC) for each gene was conducted for listed data sets. taxa: number of taxa within the data set; genes: number of genes present in the data set. To select an SOS, two settings were used: default and taxa weighting (-t option). *Scu*: *Scutigera coleoptrata*, Chilopoda (Myriapoda); *Tig*: *Tigriopus californicus*, Copepoda (Crustacea), defined as constraint taxa for data set `AP_1`. Handling `AP_2`, accessorily *Epi*, *Epiperipatus* sp. (Onychophora) was defined as constraint. For endopterygote data sets (`En_oP` and `En_oP_Da`), with respectively without *Daphnia pulex*, five proteome species were excluded to receive a submatrix with a fairly balanced taxon-gene ratio: *Tribolium castaneum*, *Bombyx mori*, *Apis mellifera*, *Drosophila melanogaster* and *Anopheles gambiae*. Those species showed the highest IC and saturation within data set `En`. Abbreviations: phylog.: phylogenetic; SO: selected optimal; rib.: ribosomal; endopt.: endopterygote.

ID	data set	taxa	genes	MARE red.	MARE options	constraints	phylog. analyses	remarks alignment
AP_1	full arthropod data set	233	775	x	default, -t	*Scu, Tig*	ML, Bayesian	masked
AP_op_un	SO arthropod data subset	117	129	-	-	-	-	unmasked
AP_1_ri	SO arthropod data subset: rib. genes	117	32	-	-	-	ML	masked
AP_1_nri	SO arthropod data subset: non-rib. genes	117	97	-	-	-	ML	masked
AP_2	full arthropod data set 2	233	775	x	default, -t	*Scu, Tig, Epi*	ML	masked
AP_3_oP	full arthropod data set, - proteome taxa	216	753	x	default, -t	-	ML	masked
En	full endopt. data set + Ensifera	111	775	x	default, -t	-	-	masked
En_Da	full endopt. data set + Ensifera + *Daphnia*	112	775	x	default, -t	-	-	masked
En_oP	full endopt. data set + Ensifera,	106	775	x	-t	-	ML	masked
	- 5 proteome taxa	106	775					
En_oP_Da	full endopt. data set + Ensifera + *Daphnia*, - 5 proteome taxa	107	775	x	-t	-	ML	masked
Dunn	full data set of Dunn et al. (2008)	77	150	x	default, -t	-	ML	-

2.2.6. Split analyses

Split networks were computed with SplitsTree 4 (Huson and Bryant, 2006). Neighbornets (Bryant and Moulton, 2004) with uncorrected p-distances were calculated for the SOS alignment of `AP_1` and its unmasked version (`AP_op_un`). Both network structures were compared, since a network graph gives an indication of noise, signal-like patterns and conflicts within an multiple sequence alignment.

2.2.7. Phylogenetic reconstructions and consensus networks

For phylogenetic inference, maximum likelihood (ML) trees were calculated with RAxML (Stamatakis, 2006b,a; Ott et al., 2007). Bayesian analyses were conducted with PhyloBayes (Lartillot et al., 2008). Consensus networks (Holland and Moulton, 2003) were generated with SplitsTree 4.8 (Huson and Bryant, 2006) to visualize possible inconsistencies between single topologies. A consensus network was constructed i) from all single Bayesian runs (SOS, `AP_1`) and ii) from both resulting topologies of the ribosomal (`AP_1_ri`) and non-ribosomal data subset (`AP_1_nri`). The thresholds of conflict was set to 0.01, and network computing implemented averaged edge weights.

Arthropod data sets

ML analysis was performed with RAxML parallelized Pthreads 7.0.0 (Stamatakis, 2006b,a; Ott et al., 2007) on data set `AP_1` and its SOS. The SOS-alignment of `AP_1` (117 taxa; 101 arthropods including onychophorans, two waterbears, selected nematodes and annelids, and three mollusks; 129 genes) spanned 37,476 amino acid positions. ML tree search and rapid bootstrapping was applied on the SOS within one step (-f a) with 1,000 bootstrap replicates. The `AP_1` alignment (233 taxa; 775 genes) spanned 350,356 amino acid positions. Here, ML tree search and rapid bootstrapping in one step was not possible due to limited Random Access Memory (RAM). Instead, ten single ML tree searches and separate bootstrapping (100 replicates) were performed. The bootstrap procedure took almost four months. Due to restricted computational power, the calculation of branch lengths was not possible. The ML tree with the best likelihood value was chosen to plot bootstrap support (Fig. 2.19). All ML tree searches were calculated with the PROTMIX model (Stamatakis, 2006b,a), applying the WAG substitution matrix (Whelan and Goldman, 2001). Rate heterogeneity was considered in all analyses. Leaf stability indices (LSI, Thorley and Wilkinson, 1999) were assessed from collected bootstrap trees for the `AP_1`-SOS and for the `AP_3_oP`-SOS with Phyutility (Smith and Dunn, 2008). A threshold was set to 0.95, taxa with an LSI < 0.95 were considered as 'unstable'. ML analysis was repeated after exclusion of 'unstable' taxa from the `AP_1`-SOS alignment to evaluate their impact on the tree topology and tree robustness.

Bayesian trees were calculated from the `AP_1`-SOS with PhyloBayes 2.3c (Lartillot et al., 2008) running the CAT mixture model (Lartillot and Philippe, 2004) and default options. 25 Markov Chain Monte Carlo (MCMC) chains ran for 20,000 cycles each, sampling every cycle. Chains were computed on a Linux Blade system (Hewlett Packard, 64 GB RAM / node, performing one chain per node); analyses took almost five months. Bayesian analyses were enabled by the SuGI (Sustainable Grid Infrastructure, project leader V. Achter, RRZK, University of Cologne). Parameter values were checked for convergence by plots using the statistical software package R (R Development Core Team, 2008, v2.9). The burn-in was set to 5,000 cycles for all chains. A majority rule consensus tree was drawn from each chain with the *bpcomp* tool implemented in PhyloBayes. The discrepancy observed across

all bipartitions of all pairwise compared chains and of all 'triple' chain combinations (three chains are considered at once) was checked on the basis of calculated *maxdiff*-values. Computation of *maxdiff*-values is also conducted with the *bpcomp* tool. Harmonic means of log-likelihood values of each chain (burn-in excluded) were calculated. Finally, three chains were included to infer the final Bayesian majority rule consensus (mrc) tree. They showed the lowest *maxdiff*-value and the best log-likelihood values (harmonic means) of all 'triple-chain combinations' (section 2.3.3, Tab. 2.7). The mrc tree was drawn with *bpcomp*. A consensus network was computed (Holland and Moulton,

2003) with SplitsTree 4.8 (Huson and Bryant, 2006) from all PhyloBayes chains (section 2.3.3).

For data sets `AP_1_ri`, `AP_1_nri` and both SOS of `AP_2` and `AP_3_oP` (taxa weighted), maximum likelihood analyses were conducted (Tab. 2.3 and section 2.3). Again, analyses were performed with RAxML Pthreads 7.0.0 (-f a; PROTMIX, WAG; 1,000 bootstrap replicates). All trees were rooted with mollusks.

Endopterygote data sets

Maximum likelihood analyses were performed with RAxML Pthreads 7.0.0 (-f a; PROTMIXWAG; 1,000 bootstrap replicates) for endopterygote selected optimal subsets (see Tab. 2.5 and section 2.3.3). ML analyses for respective full data sets, `En_oP` and `En_oP_Da` (Tab. 2.3), were conducted with identical settings, but 100 bootstrap replicates. Trees were rooted with i) Ensifera (Orthoptera) or ii) *Daphnia* (Branchiopoda, Crustacea). Computation time for the phylogenetic reconstruction for complete data sets took about six weeks on Linux blade systems of the ZFMK (section 2.2.7).

Published data sets: Dunn et al. (2008)

For the data set of Dunn et al. (2008), ML trees were inferred full data set (77 taxa, 150 genes), for the SOS data subset (53 taxa, 33 genes, taxa unweighted) and for the SOS with the -t option applied (70 taxa, 29 genes). For each analysis, 1,000 bootstrap replicates and ML search were conducted in one step (-f a) with RAxML Pthreads 7.0.0 on Linux Blade systems (see section 2.2.7).

2.3. Results

Alignment masking with Aliscore (Misof and Misof, 2009; Kück et al., 2010) and ALICUT (Kück, 2009) `http://utilities.zfmk.de` increased signal in arthropod and endopterygote data sets. After applying MARE, all selected data subsets considerably showed a higher total average information content (IC) and matrix saturation. Inferred trees from selected optimal data subsets (SOS) showed an improved resolution for nearly all clades. This was true, at least, for all SOS that display a power-law distribution of missing data (Li and Liu, 2007) in respective original data matrices. Thus, present analyses suggest that the strategy to select a subset with higher total average IC and matrix saturation was successful. Several deep splits in previous studies thought to be resolved (Dunn et al., 2008; Roeding et al., 2007), e.g. the position of sea spiders (Pycnogonida) or the position of myriapods, however, showed remarkable sensitivity to available data and reconstruction methods. The placement of Myriapoda, in this study only represented by centipedes and millipedes, remained unresolved.

2.3.1. Alignments masking and taxa/gene selection

Alignment masking

In general, signal was increased and noise was reduced with Aliscore (Misof and Misof, 2009; Kück et al., 2010) and ALICUT. From originally 826,633 amino acid positions of data set `AP_1`, 42.38% (350,356 aa positions) retained after alignment masking. The arthropod data set `AP_3_oP` (proteome species excluded) include 16 genes with only one species, where alignment masking was not possible. The unmasked supermatrix of `AP_3_oP` comprise 311,020 amino acid positions. This is less than half the alignment length of the unmasked alignment of data set `AP_1` (proteome species included). Thus, the enormous length of the unmasked `AP_1` alignment is mainly caused by proteome species. For data set `AP_3_oP`, 73.44% (228,401 positions) retained after alignment masking, a substantial higher percentage compared with data set `AP_1`. In both endopterygote data sets (Tab. 2.3), the percentage of excluded positions was similar. The masked alignment of data set `En` stands 72.95% (411,489 of originally 564,001 amino acid positions) of its original superalignment length. For data set `En_Da`, 73.25% from originally 569,270 characters were included. Alignment masking was not applied on the data set of Dunn et al. (2008). The authors used Gblocks (Castresana, 2000) to mask the alignments. Since only the masked alignments were provided, a comparison between unmasked and masked alignments to examine its impact on signal or phylogenetic inference was not possible.

Selection of taxa and genes using new reduction heuristics

A remarkable increase of total average information content and matrix saturation has been recorded for all by MARE selected submatrices. For some data sets, taxa weighting (- t option) was necessary to obtain a fairly balanced taxon-gene ratio. In first order, as much taxa as possible and, in second order, as much genes as possible should be selected. Number of taxa and genes, total average information content and matrix saturation for data sets and respective SOS are listed in Tab. 2.4 and 2.5.

Simulations have shown that, at least, a total average information content of a data matrix of ca. 0.3 theoretically is sufficient to infer a plausible resolved phylogenetic reconstruction (Misof and Meyer, pers. comm). This criterion was fulfilled for all SOS in this study (Tab. 2.4, 2.5).

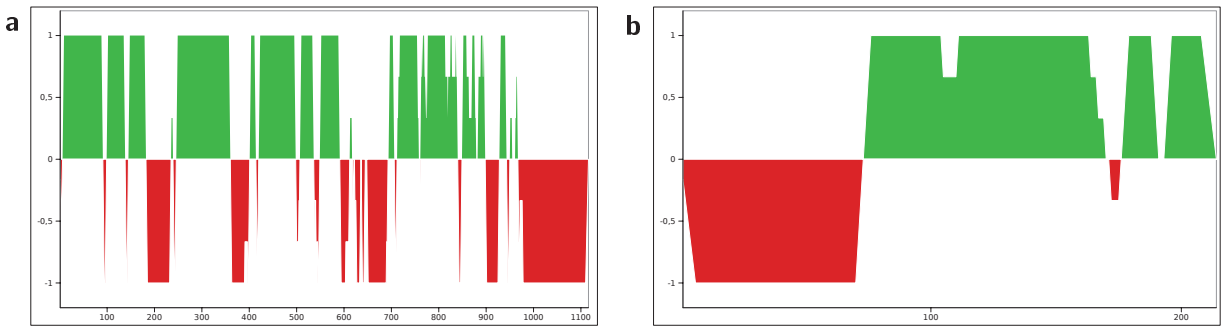

Figure 2.6.: Aliscore consensus profiles for two gene alignments from data set AP_1. Positions and alignment length are given on the x-axes, y-axes show the score value of each position within an alignment. Red (values < 0) represent random similarity, green (values ≥ 0) non-random similarity. **a** Consensus profile of gene 11352. 38.17% (426 from originally 1,116 positions) were excluded from the alignment. **b** Consensus profile of gene 12040. 76 positions (35.51%) were discarded.

Table 2.4.: Results of MARE before and after applying reduction heuristics on arthropod data sets (see Tab. 2.3). Bold and blue: selected optimal subset (SOS) used for phylogenetic inference. Bold: original data set used for phylogenetic inference. total average IC: total average information content of the matrix; -t: option taxa weighting (5/3) in MARE; *Scu*: *Scutigera coleoptrata*, Chilopoda (Myriapoda); *Tig*: *Tigriopus californicus*, Copepoda (Crustacea); *Epi*: *Epiperipatus* sp. (Onychophora).

data set	**features**	**original data set**	**SOS** default	-t	**remarks**
AP_1	total average IC	**0.103**	**0.43**	0.43	taxa weighting [-t]
	matrix saturation	**17.6%**	**62.3%**	62.3%	had no effect on the
	number of taxa	**233**	**117**	117	data subset selection,
	number of genes	**775**	**129**	129	constraints: *Scu*, *Tig*
AP_2	total average IC	0.103	**0.43**	-	constraints: *Scu*, *Tig*, *Epi*
	matrix saturation	17.6%	**62.3%**	-	
	number of taxa	233	**118**	-	
	number of genes	775	**127**	-	
AP_3_oP	total average IC	0.087	0.625	**0.4**	all proteome
	matrix saturation	12.7%	81.4%	**55.8%**	species excluded
	number of taxa	216	32	**100**	
	number of genes	753	113	**125**	

The diagram (Fig. 2.8) reflects the reduction process to select the SOS from data set AP_1. With every reduction step, the size-ratio of the matrix λ (size of the reduced matrix M' / size of original matrix M, see 2.1.4) decreases (blue graph). Concurrently, the total average information content (IC) P increases (red graph). Low total average IC of the matrix is penalized; the optimality function $f(p)$ (orange graph, see 2.1.4) reaches a maximum after 764 reduction steps displaying SOS with 8.1% of the original matrix size.

Figure 2.9 shows the connectivity between taxa for the AP_1-SOS. Total connectivity is set to a minimum number of two overlapping genes and taxa (see section 2.1.4). The tree displays how

many genes one taxon *A* shares with another taxon *B*; it does not reflect phylogenetic relationships. *Campodea* cf. *fragilis* (Diplura), for example, has most genes (41) in common with *Heliothis virescens* (Lepidoptera). The mayfly *Baetis* has the largest gene overlap with the springtail *Folsomia candida*, they share 22 genes. The springtail *Anurida maritima* shares most genes (20 genes) with the 'taxa group' (*Baetis,Folsomia candida*). Also the proturan *Acerentomon franzi* shows its largest gene overlap (35 genes) with a butterfly + a penaeid shrimps (*Plutella xylostella,Marsupenaeus japonicus*). With the penaeid shrimp only, *Anurida* has 33 genes in common. *Blattella* (Blattodea) shows its maximal gene overlap (21 genes) with *Calanus* (Crustacea). In total, this is the smallest maximal gene overlap between two taxa in this data subset. Most proteome species show a maximal gene overlap with other proteome taxa.

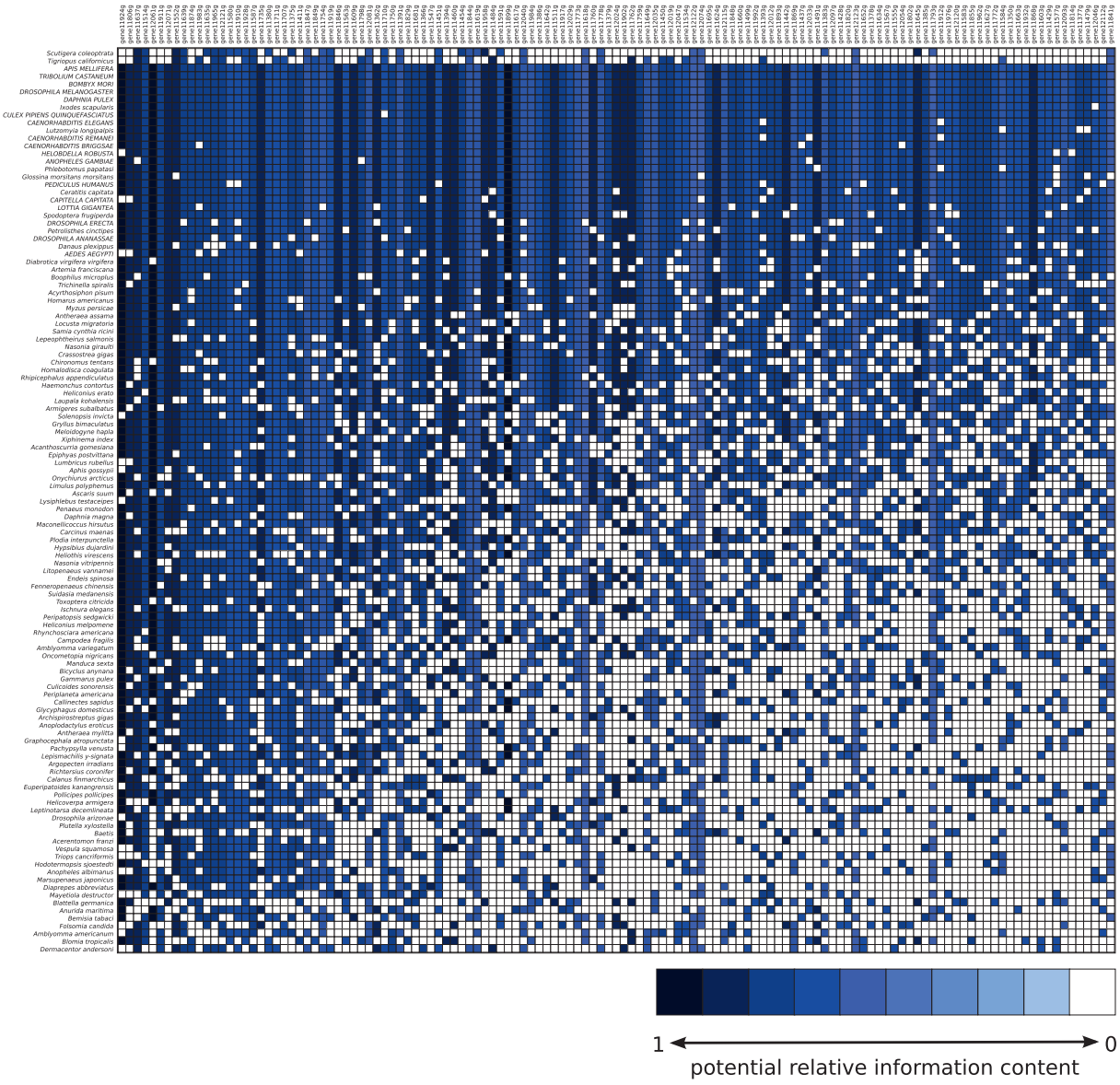

Figure 2.7.: Selected optimal subset (SOS) of data set `AP_1`. The submatrix comprises 117 taxa (rows) and 129 genes (columns). The matrix is sorted; high information content (IC) and matrix saturation is located on the left, top; low information content and matrix saturation is located on right, bottom. The first and second row display the constraint taxa *Tigriopus* (Copepoda, Crustacea) and *Scutigera* (Diplopoda, Myriapoda). X-axis: genes, Y-axis: taxa. Potential information content ranges from 0.0–1.0 (10 units) and is color coded from dark blue (> 0.9–1.0) to white (IC of $\geq$ 0–0.1 or missing data). Genes with a IC < 0.04 were considered as absent. Total average IC of the matrix: 0.43, overall saturation: 62.3%.

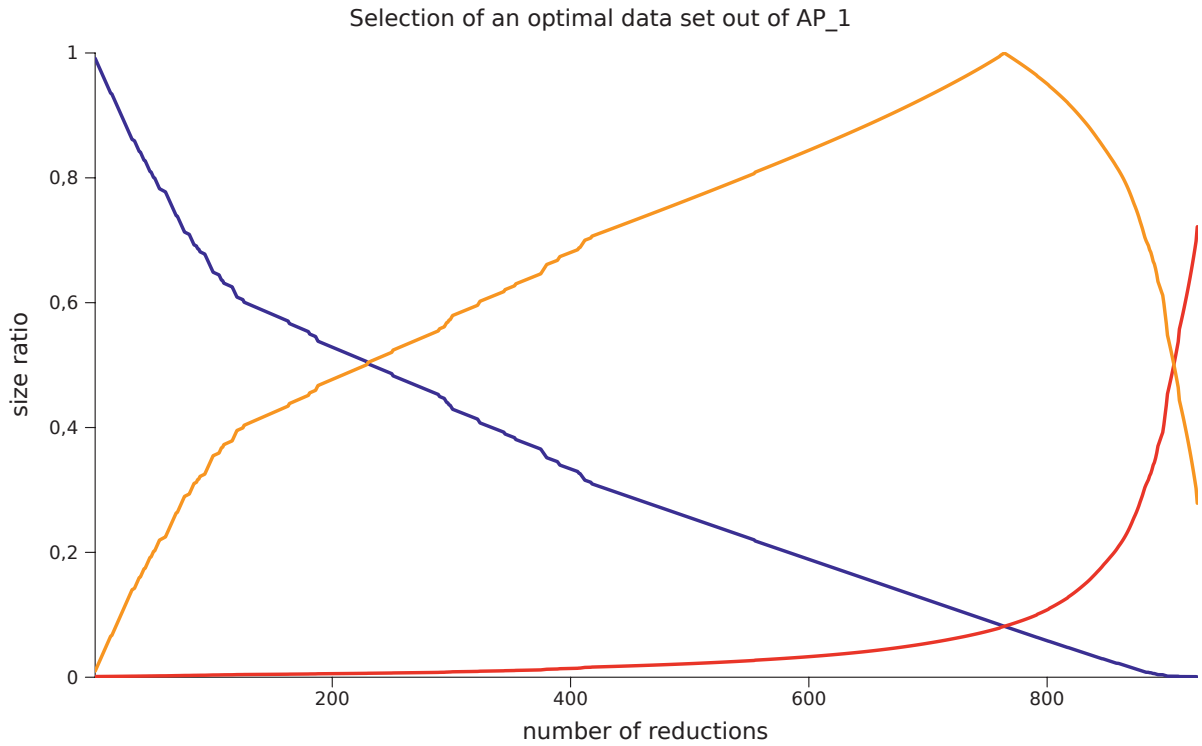

Figure 2.8.: Process of reduction heuristics for selecting an SOS out of data set AP_1. The size ratio λ (blue) of decreases while the total average information content P' of the matrix increases (red). The optimality function $f(P)$ (orange), (see 2.1.4) reaches its maximum (intersection point) after 764 reduction steps.

The unmasked SOS, AP_op_un, was compared with the masked SOS of AP_1 to evaluate the impact of alignment masking on information content and matrix saturation. The concatenated unmasked SOS alignment has a length of 81,713 amino acid positions with a total average information content (IC) of 0.244 (Fig. 2.10). This is substantially lower compared with the masked SOS (total average IC: 0.43, see Fig. 2.7). For 16 genes, the IC is = 0. The matrix saturation of AP_op_un is again lower (41.9%) compared with the masked SOS (63.2%). The effect of alignment masking, a reduction of noise, is also visualized in split networks (see section 2.3.2 and Figs. 2.16-2.17).

Ribosomal and non-ribosomal data subsets AP_1_ri and AP_1_nri

Data subsets AP_1_ri and AP_1_nri were created by splitting the SOS of AP_1 into ribosomal and non-ribosomal protein coding genes. AP_1_ri consists of 32 ribosomal genes (Tab. A.4) and AP_1_nri consists of 97 non-ribosomal protein coding genes. Both data sets include all taxa of the SOS of AP_1. Data subset AP_1_ri shows a total average information content of 0.522 and a matrix saturation of 82.9%. This is sizeably higher compared with the SOS of AP_1. Data subset AP_1_nri shows a slightly lower total average information content (0.4) and matrix saturation 55.5% than the SOS of AP_1. The potential signal among single genes highly heterogeneous in both data subsets.

Phylogenetic trees were inferred for the maximum likelihood approach (see 2.2.7) to examine topological differences. Mainly, the position of myriapod species and relationship between Protura and Diplura was addressed. The idea was to examine if e.g. Nonoculata is supported by ribosomal and non-ribosomal protein coding gene sets. This clade has been recently inferred by studies dealing with ribosomal RNA genes (e.g. Luan et al., 2004, 2005; Mallatt and Giribet, 2006; Gao et al., 2008;

Dell'Ampio et al., 2009). In spite of high support for this clade in most studies, it has been discussed by several authors as biased due to GC-richness in rRNA sequences (e.g. Luan et al., 2005; Gao et al., 2008; Szucsich and Pass, 2008; Dell'Ampio et al., 2009). The GC-richness might be not restricted to structural RNAs, but may also be present in ribosomal protein coding genes (Dell'Ampio, pers. comm.). For `AP_1_ri`, the information content of genes varies from 0.42–0.86. The distribution of present|absent data resembles a Gaussian distribution (Fig. 2.11). For `AP_1_nri`, the information content of genes ranges from 0.46–0.92 and shows an exponential distribution of present|absent data (Fig. 2.11). Some taxa have only few genes in common: for example, within the non-ribosomal data subset `AP_1_nri`, the centipede *Scutigera* and the millipede *Archispirostreptus* share only two genes, both with a moderate information content (0.53 and 0.46).

Figure 2.9.: Distance tree of the SOS of (AP_1) which displays connectivity, meaning gene overlap or 'gene clusters' between species identified by MARE. Proteome species (Capital letters): gray; own EST taxa (*Acerentomon*, *Campodea*, *Anurida*, *Lepismachilis*): blue.

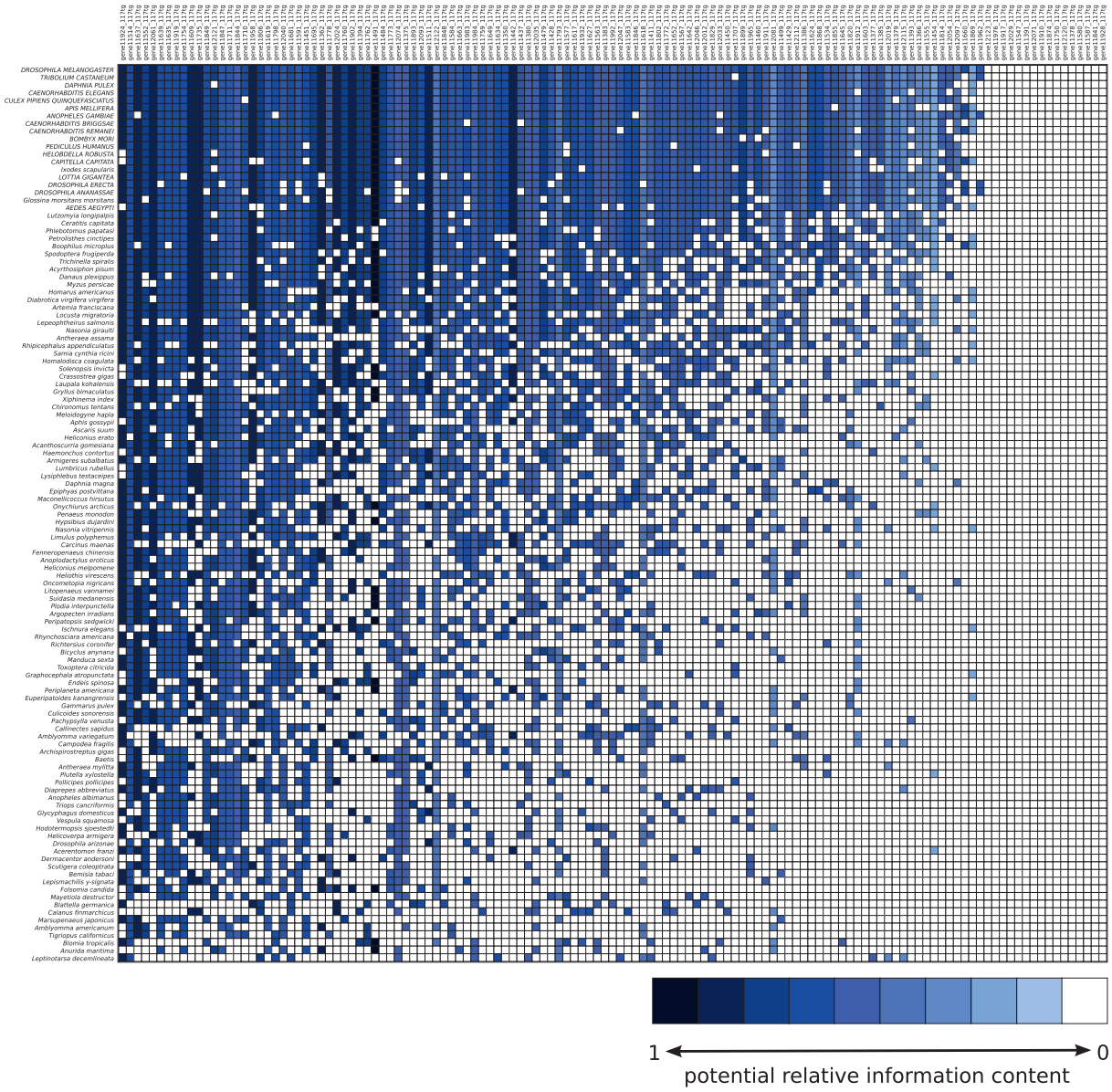

Figure 2.10.: Matrix of the unmasked SOS AP_op_un. Labeling and color code are specified in Fig. 2.7. Total average information content: 0.244, matrix saturation: 41.9%.

Data set AP_2

For data set AP_2, additionally the velvet worm *Epiperipatus sp.* was defined as constraint (Tab. 2.3). The total average information content and matrix saturation of its SOS was similar to the SOS of AP_1 (Tab. 2.4). Apart from *Epiperipatus* sp., identical taxa were selected. Instead of 129 genes, 127 genes were selected. All selected genes already had present in the SOS of AP_1. The SOS alignment of data set AP_2 has a length of 37,045 amino acid positions. This is only 431 less characters than the SOS alignment of AP_1. Gene number 11377, which codes for an arginine/serine-rich splicing factor, and gene number 11814, coding for the enzyme lactoylglutathione lyase, had been dropped from the matrix (Tab. A.3).

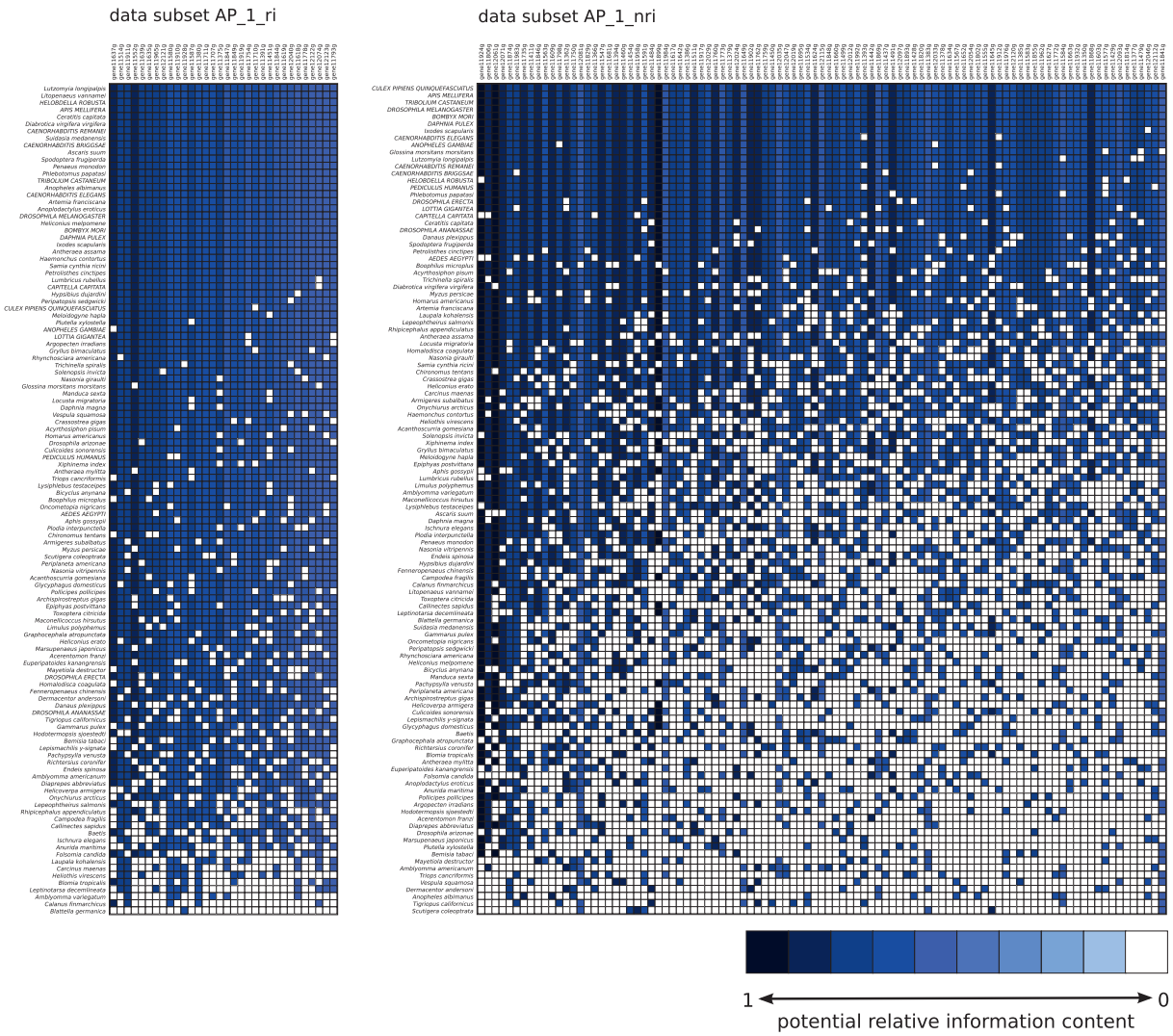

Figure 2.11.: Matrix of data subset `AP_1_ri` (32 genes) and data subset `AP_1_nri` (97 genes). Labeling and color code are specified in Fig. 2.7. `AP_1_ri`: total average information content: 0.522, matrix saturation: 82.9%. `AP_1_nri`: total average information content: 0.4, matrix saturation: 55.5%.

Data set AP_3_oP

Data set `AP_3_oP` covers 216 taxa and 753 genes; 23 genes are not considered because they comprised less than 4 taxa. Total average information content (0.087) and matrix saturation (12.7%) are smaller compared with `AP_1` (Tab. 2.4) while the information content of single genes remarkably differs. 57 genes have a information content = 0 (Tab. A.5). Present|absent data are power-law distributed and potential signal throughout the matrix shows a high heterogeneity. The SOS covers 113 genes, but only 32 taxa (total average information content: 0.625, matrix saturation: 81.4%). Applying taxa weighting, the SOS obtained an total average information content of 0.4 and a matrix saturation of 55.8% with 125 taxa and 100 genes (Fig. 2.12, Tab. A.5). The alignment spans 29,534 amino acid positions. The information content of genes ranges from 0.95–0.41. For phylogenetic inference, latter SOS was used due to its fairly balanced taxon-gene ratio. It shares 97 genes with SOS of `AP_1`. The SOS includes three genes that are not present in the SOS of `AP_1` (signal sequence receptor: gene 11676; translocon-associated protein gamma subunit: gene 11811; pyruvate kinase: gene 12018, marked in blue, Tab. A.5) and 24 additional taxa (Tab. A.6). *Tigriopus*, *Scutigera* and *Epiperipatus* retained in the SOS without defining them as constraints. The SOS still shows a power-law distribution of present|absent data and a high heterogeneity of potential signal. For 80 genes, the information content was higher than in the SOS of `AP_1`.

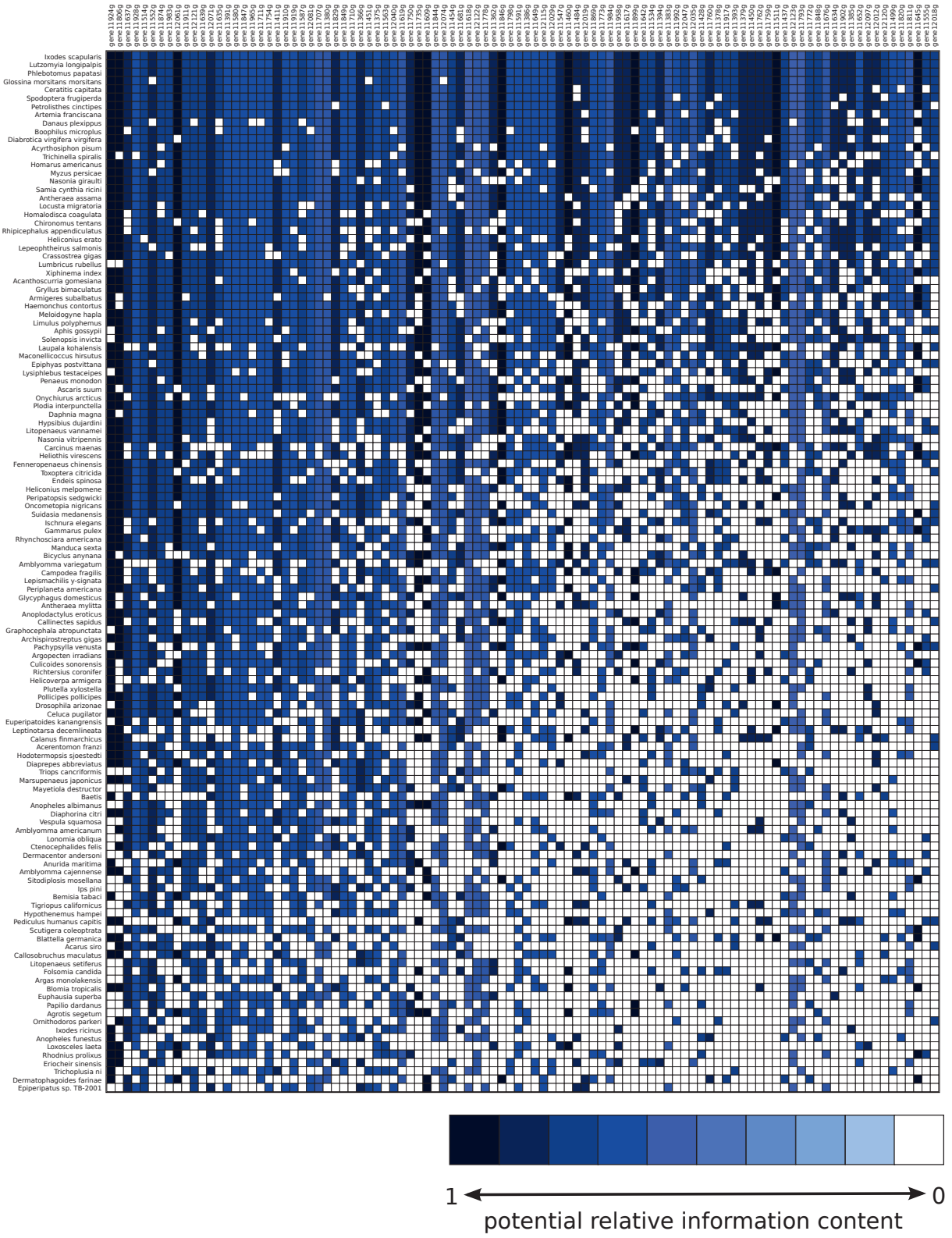

Figure 2.12.: SOS matrix of `AP_3_oP` (`-t` option). The SOS covers 125 taxa and 100 genes. Labeling and color code are specified in Fig. 2.7. Total average information content: 0.4, matrix saturation: 55.8%.

Endopterygote data sets

All original endopterygote data sets show a high heterogeneity of potential signal among genes and a strong power-law distribution of present|absent data.

Data sets `En` and `En_Da`

Endopterygote data sets `En` and `En_Da` show both a similar total average information content (≈ 0.11) and matrix saturation (≈ 15–16%). Both optimal subsets selected by MARE (either with or without taxa weighting) were rather inappropriate for phylogenetic inference due to an unbalanced taxon-gene ratio. The SOS of `En` consisted of 9 taxa and 365 genes, respectively 30 taxa and 519 genes (`-t` option applied, Tab. 2.5). Likewise, too much taxa had been dropped of data set `En_Da` either without or with taxa weighted. Nevertheless, the total average information content and matrix saturation remarkably increased for both data sets. To gain a considerable taxon-gene ratio, five proteome taxa had been excluded (`En_oP` and `En_oP_Da`, see Tab. 2.3).

Data sets `En_oP` and `En_oP_Da`

`En_oP` and `En_oP_Da` show a total average information content (IC) of ≈ 0.09 and a matrix saturation of ≈ 12%. Both SOS (generated with default settings) revealed an total average IC and a matrix saturation that were increased by a factor of 10 (total average IC: ≈ 0.9, matrix saturation: 100% / 98.2%, Tab. 2.5). The heterogeneity of potential signal among genes was very low in both SOS. Again, the default setting of MARE resulted in subsets with an unbalanced taxon-gene ratio (4 taxa, 345 genes for `En_oP` and 5 taxa, 226 genes for `En_oP_Da`). Additionally, both SOS exclusively consisted of proteome species which covered all genes (Fig. 2.13). Proteome species within these data matrices may attach great weight during selection, in favor to keep them. These subsets were not used for phylogenetic inference, since it was aimed to obtain subsets with much taxa as possible. The usage of the `-t` option resulted in subsets with a more balanced taxon-gene ratio, albeit not ideal. The SOS of `En_oP` covers 29 taxa and 63 genes (total average IC: 0.67; matrix saturation: > 80%) and displays a low heterogeneity of potential signal among genes. The SOS of `En_oP_Da` includes 29 taxa and 112 genes (total average IC: 0.6; matrix saturation: 78%) and shows a moderate signal heterogeneity among genes (Fig. 2.14). Both SOS were used for phylogenetic inference (Tab. 2.5).

Data set `En_oP_Da` (`-t` option) demonstrates the possible impact on the performance of reduction heuristics due to proteome species covering all or a large percentage of genes. Including *Daphnia pulex* provides an SOS that shows nearly twice as much genes compared with the SOS where previously *Daphnia pulex* had been excluded (Fig. 2.14). Therefore, proteome species should treated cautiously, especially, if data matrices have a high percentage of missing data which is power-law distributed.

Figure 2.13.: Selected data subsets of `En_oP` and `En_oP_Da` with default settings. Labeling and color code are specified in Fig. 2.7. Above: SOS of `En_oP`. Four proteome taxa retained (*C. pipiens quinquefasciatus*, *D. ananassae*, *D. erecta*, *A. aegypti*). The SOS includes 345 genes (total average information content: 0.922; matrix saturation: 100%. Below: SOS of `En_oP_Da`. Five proteome taxa retained (*D. pulex*, *D. ananassae*, *C. pipiens quinquefasciatus*, *D. erecta*, *A. aegypti*). The SOS includes 226 genes (total average information content: 0.902; matrix saturation: 98.2%).

Table 2.5.: Results of MARE on endopterygote data sets (see Tab. 2.3). Bold, blue: SOS used for phylogenetic inference. Bold: original data sets used for phylogenetic inference. Total average IC: total average information content of the matrix; -t: option taxa weighting (5/3) in MARE.

data set	**features**	**original data set**	**data subset** default	-t	**remarks**
En	total average IC	0.110	0.761	0.415	
	matrix saturation	15.7%	97.5%	58.1%	
	number of taxa	111	9	30	
	number of genes	775	365	519	
En_Da	total average IC	0.108	0.674	0.398	
	matrix saturation	16.4%	96.7%	58.7%	
	number of taxa	112	10	31	
	number of genes	775	511	533	
En_oP	total average IC	**0.09**	0.922	**0.67**	5 proteome taxa
	matrix saturation	**11.8%**	100%	**82.9%**	excluded
	number of taxa	**106**	4	**29**	
	number of genes	**775**	345	**63**	
En_oP_Da	total average IC	**0.091**	0.902	**0.6**	5 proteome taxa
	matrix saturation	**12.7%**	98.2%	**78%**	excluded
	number of taxa	**107**	5	**29**	
	number of genes	**775**	226	**112**	

Data set of Dunn et al. (2008)

The data set of Dunn et al. (2008) was already optimized by the authors. For selection of taxa and genes, they used a particular threshold based on presence|absence information. In first order, MARE was applied on this data set to evaluate total average information content and matrix saturation. Additionally, differences were examined with respect to node resolution, topology and support of arthropod clades between the original data set and the SOS (default and -t). The original data set includes 77 taxa and 150 genes but, this data matrix shows a low total average information content (0.175, Fig. 2.15). The matrix saturation is moderate (50.19%). Information content of genes ranges from 0.06 to 0.8 (average: 0.34) and is heterogeneous among genes. The distribution of present|absent data is similar to a Gaussian distribution. The SOS (default options) includes 53 taxa and 33 genes (total average information content: 0.408, matrix saturation: 80.16%). The SOS selected with taxa weighting (-t option) includes 70 taxa and 29 genes (total average information content: 0.384, matrix saturation: 77.29%).

2.3.2. Split analyses

Neighbornet analyses were calculated from the masked SOS alignment (117 taxa, 129 genes) of `AP_1` and from the unmasked SOS alignment `AP_opt_un`. Noise within the unmasked SOS, producing cob-webs, is slightly diminished in the masked SOS. Both neighbornet graphs (Fig. 2.16), however, illustrate a dense network. The inner parts show little treeness (Fig. 2.17), which indicates a high degree of conflicting signal. The neighbornet graph of the unmasked SOS alignment (Figs. 2.16a and 2.17a) shows a fuzzier pattern, especially within deep splits. The graph calculated from the

masked SOS alignment shows a more distinct pattern (Fig. 2.16b and 2.17b), e.g. for pterygote insects (except hemipterans). This applies also for Diptera (Hexapoda), Branchiopoda (Crustacea) or Myriapoda. Within the masked SOS, common patterns for large clades, for example ectognathous hexapods, are more distinct than the unmasked graph. This might indicate an increased signal. Both graphs, moreover, indicate the presence of some typical problems of studies that address deep phylogeny: (1) Some taxa, for example Euchelicerata or Hemiptera appear in different sections of the network. Diplura and Protura cluster with remaining apterygote hexapods, Archaeognatha (*Lepismachilis*) and Collembola, but a distinct pattern for Entognatha is not clearly perceptible. (2) Single species, *Pediculus*, several species within Euchelicerata and complete groups like Nematoda have long branches. Consequently, these taxa may be misplaced due to signal erosion or homoplasies. These taxa may have an impact on the position of other taxa due to long branch artifacts. Their placement in trees should be discussed with caution (Wägele and Mayer, 2007).

2.3.3. Phylogenetic reconstructions

Phylogentic reconstructions of this thesis base on the largest phylogenomic arthropod data set currently available. Selection of optimal subsets (SOS) was successful for all data sets with a power-law distribution (arthropod and endopterygote data sets), as it improved tree robustness and clade resolution in general. Results are focused on the monophyly of hexapods, entognaths and ectognaths and on relationships between primary wingless orders. Further, a possible sister group of hexapods and the position of myriapods with respect to Mandibulata *versus* Myriochelata or Atelocerata are addressed. Considering endopterygote data sets, it is focused on the position of Hymenoptera.

Arthropod data sets

Original data set AP_1 and its selected optimal subset (SOS). Maximum likelihood and Bayesian tree reconstructions of the SOS of AP_1 resolve arthropod relationships with many strongly supported nodes, especially for Hexapoda and internal hexapod clades. In contrast, the tree based on the original supermatrix is in many respects unresolved or shows low support values (Fig. 2.19). Moreover, this tree provides some suspicious results, for example polyphyletic lepidopterans or polyphyletic millipedes. The comparison of both trees, unreduced *versus* SOS, suggests that the strategy of MARE reduction heuristics has been successful. Tree resolution is increased; suggested clades (e.g. Hexapoda, Ectognatha, Endopterygota, Coleoptera, Lepidoptera etc.) are widely accepted. Suspicious clades, like polyphyletic lepidopterans disappeared (Figs. 2.20-2.21).

Original data set AP_1: Maximum Likelihood analyses
The tree inferred from data set AP_1 (original supermatrix) shows some clades with high support, for example Euarthropoda + Onychophora or Euchelicerata. Most clades of commonly recognized groups, however, are moderately (pancrustaceans, BS 89%) or weakly supported (chelicerates or Endopterygota, BS $< 80\%$). To single out suspicious results, the millipede *Julida* and the butterfly *Euclidia* cluster with two apterygotes, *Acerentomon* (Protura) and *Campodea* (Diplura), within a polytomy (Fig. 2.18). Many well founded groups are para- or polyphyletic (e.g. myriapods, hexapods, dicondylian insects, coleopterans, trichopterans, lepidopterans) due to misplaced single representatives. Especially, within endopterygote groups, for example coleopterans or lepidopterans, there is no reso-

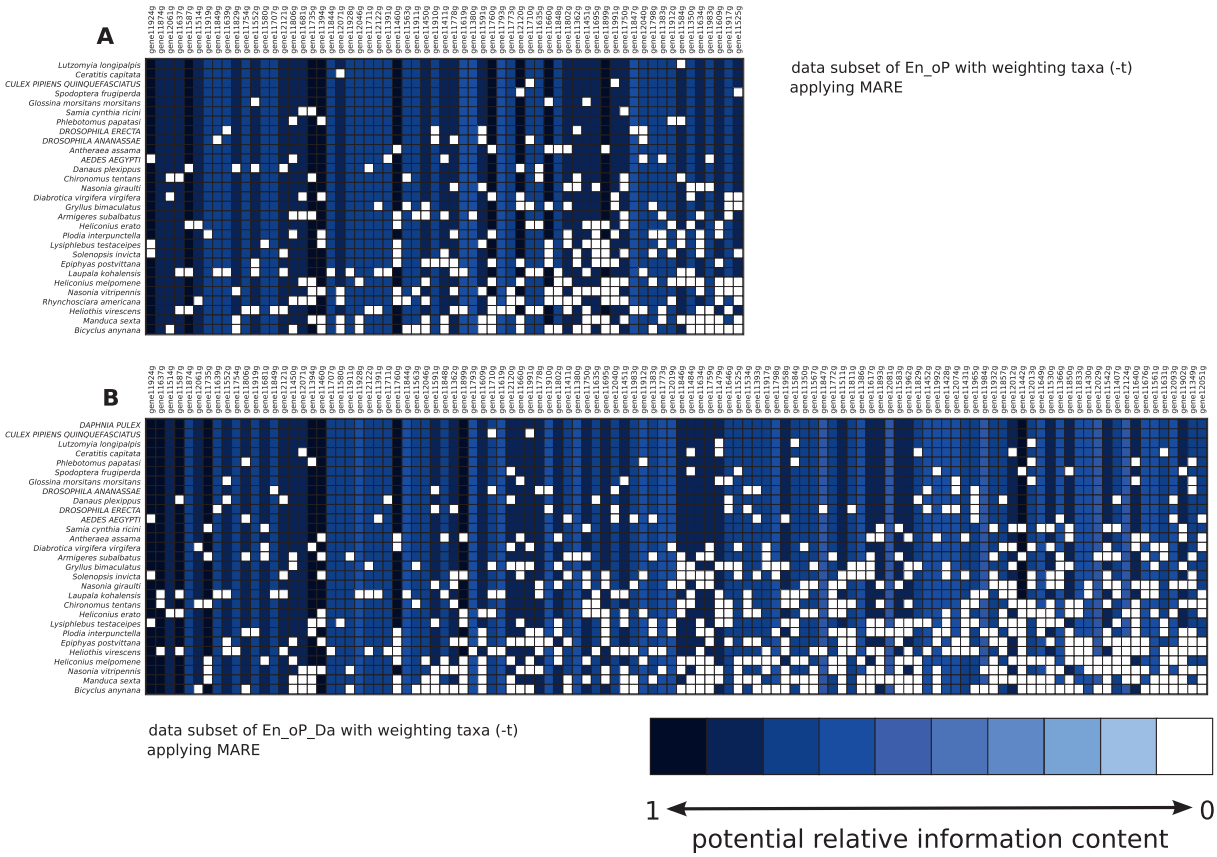

Figure 2.14.: Selected data subsets of `En_oP` and `En_oP_Da` with `-t` option. Labeling and color code are specified in Fig. 2.7. Proteome species are capitalized. **A** SOS of `En_oP`. 29 taxa and 63 genes were selected (total average information content: 0.67; matrix saturation of 82.9%). **B** SOS of `En_oP_Da`. 29 taxa (incl. *Daphnia*) and 112 genes were selected (total average information content: 0.6; matrix saturation: 78%).

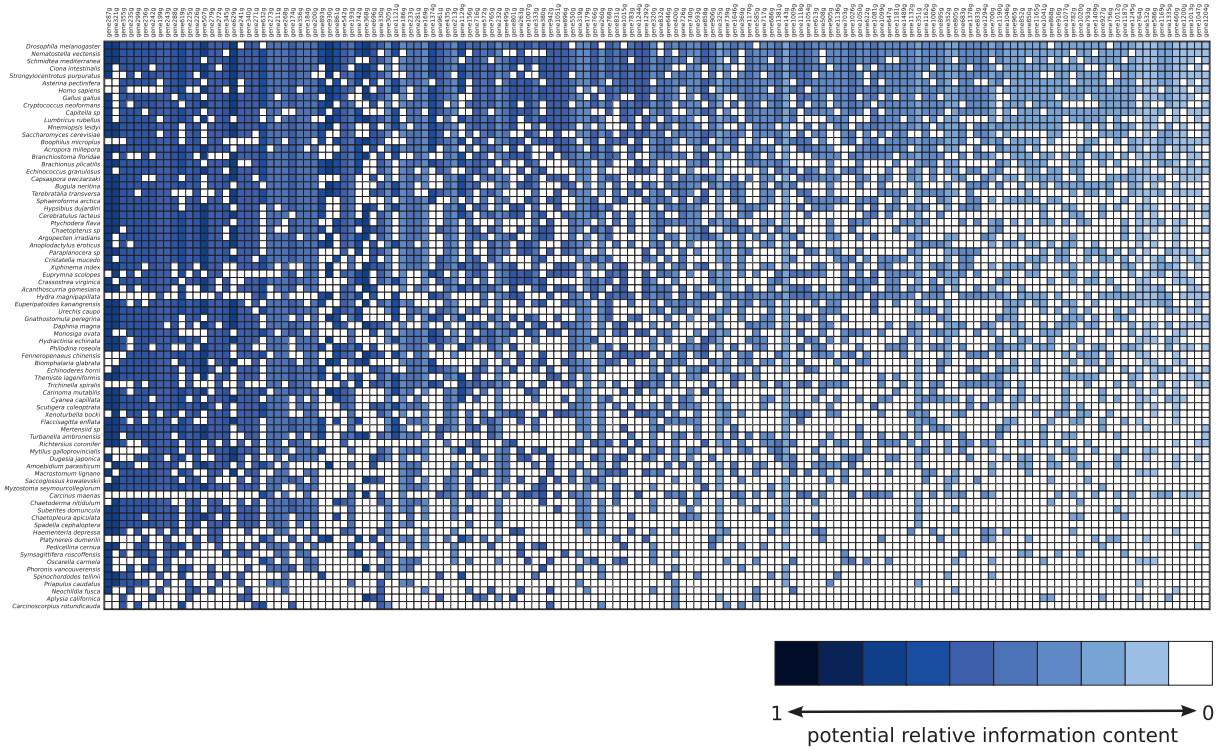

Figure 2.15.: Original data matrix of Dunn et al. (2008). Labeling and color code are specified in Fig. 2.7. The data set includes 77 taxa and 150 genes. Total average information content: 0.175, matrix saturation: 50.19%.

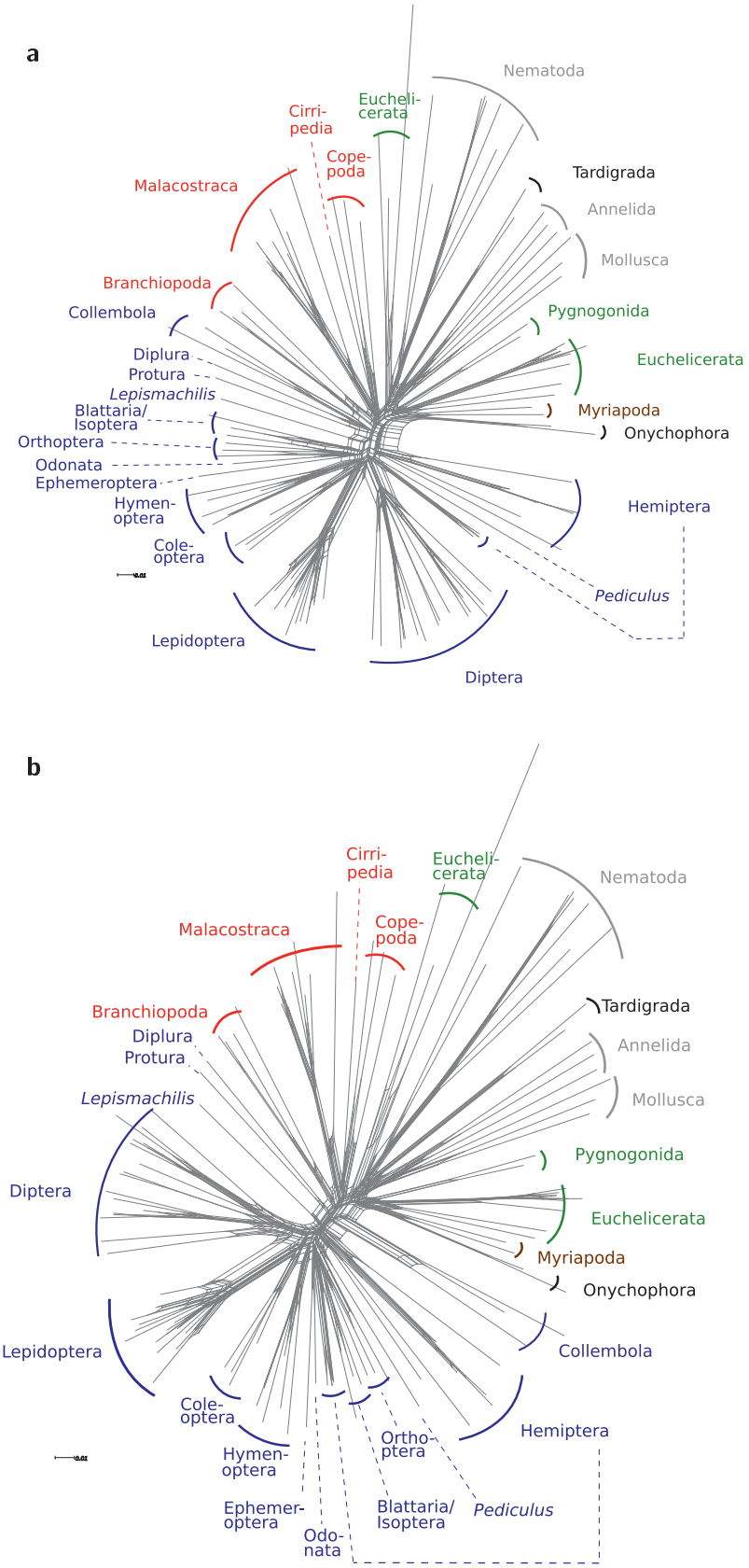

Figure 2.16.: Neighbornet graphs of the unmasked and masked SOS (117 taxa, 129 genes) of AP_1. Neighbornet graphs were calculated with uncorrected p-distances in SplitsTree 4.8. **a** Network of the unmasked SOS alignment (81,713 characters). **b** Network of the masked SOS alignment (37,476 characters) which was chosen for phylogenetic reconstruction. Color code: mollusks, annelids and nematodes: gray; tardigrades, onychophorans: black; myriapods: brown; chelicerates: green; crustaceans: red; apterygote hexapods: blue; pterygote insects: dark blue.

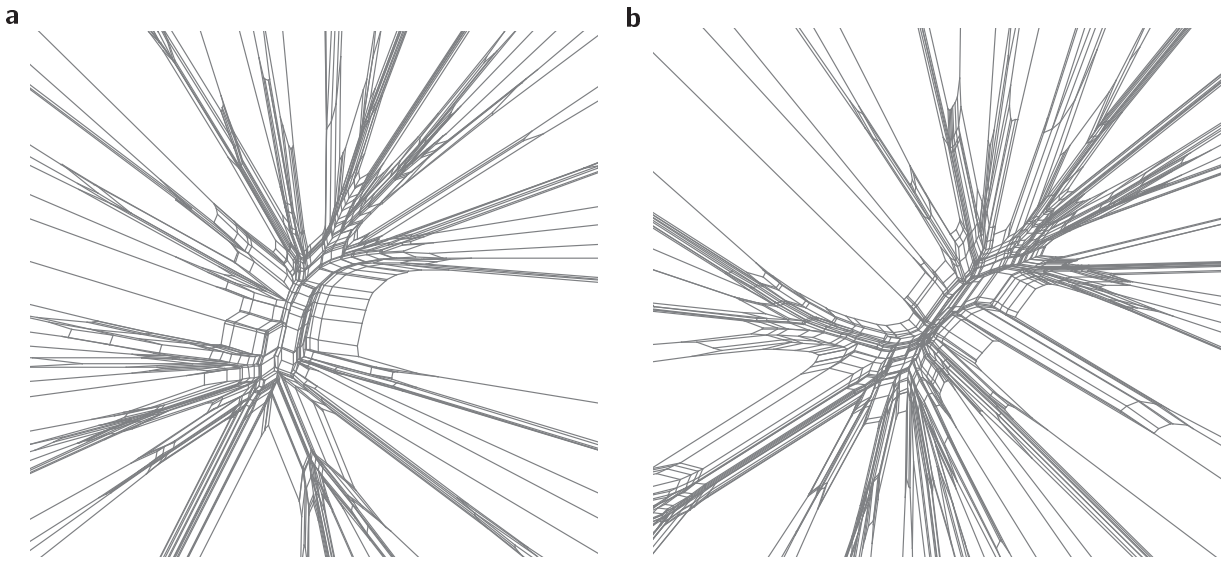

Figure 2.17.: Sections of the inner neighbornet graphs in Fig. 2.16. **a** Section of the unmasked network prior to alignment masking. **b** Section of the masked network after applying Aliscore and ALICUT. Especially within deep splits, the patterns are more distinct compared with the unmasked neighbornet.

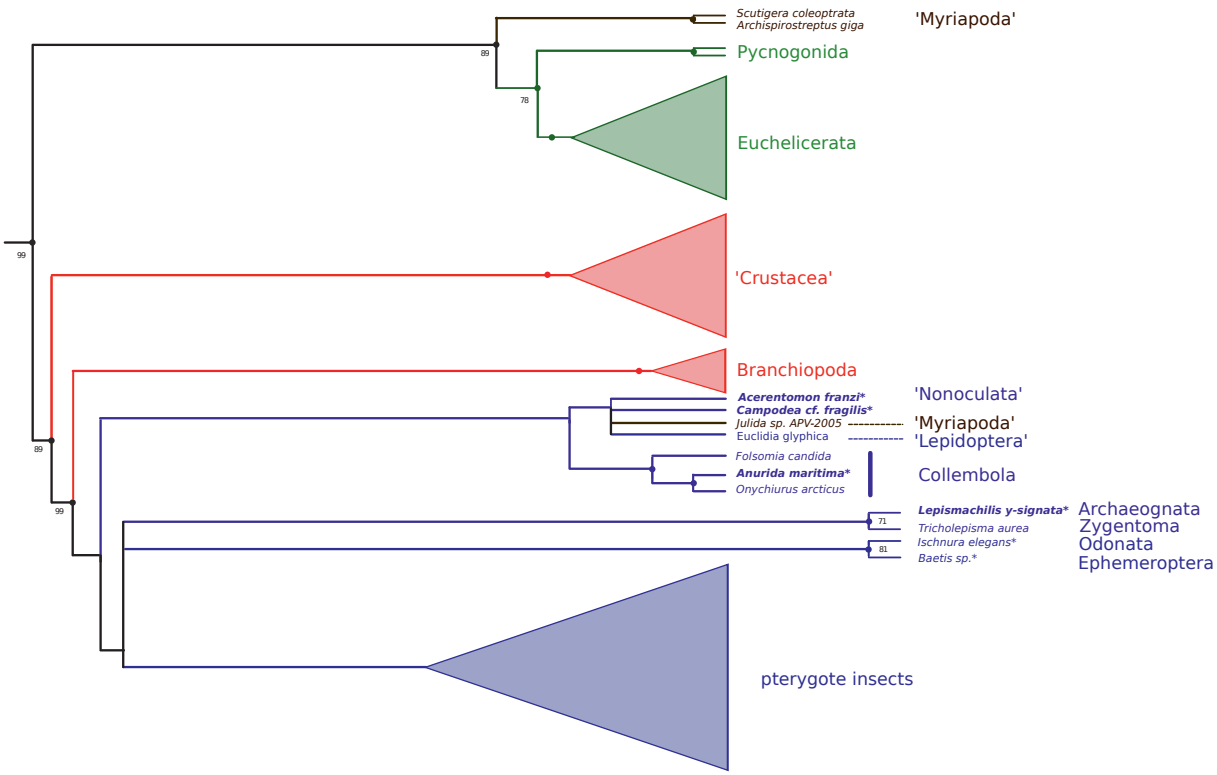

Figure 2.18.: Schematic cladogram (ML) inferred from AP_1 (233 taxa, 775 genes, see Fig. 2.19). Entognatha cluster with *Julida* (Diplopoda) and with *Euclidia* (Lepidoptera) within in a polytomyous clade. Consequently entognaths, hexapods, myriapods, lepidopterans are polyphyletic. Color code and labeling as in Fig. 2.19.

lution (see Fig. 2.19). In contrast, the Maximum likelihood and the Bayesian SOS tree show high support for euarthropod splits and most orders. Euchelicerate orders, Ixodida (ticks) and Astigmata (mites), crustaceans groups like Decapoda or Copepoda, and nearly all hexapod orders (Collembola, Hymenoptera, Coleoptera, etc.) show strong support (Figs. 2.20-2.21).

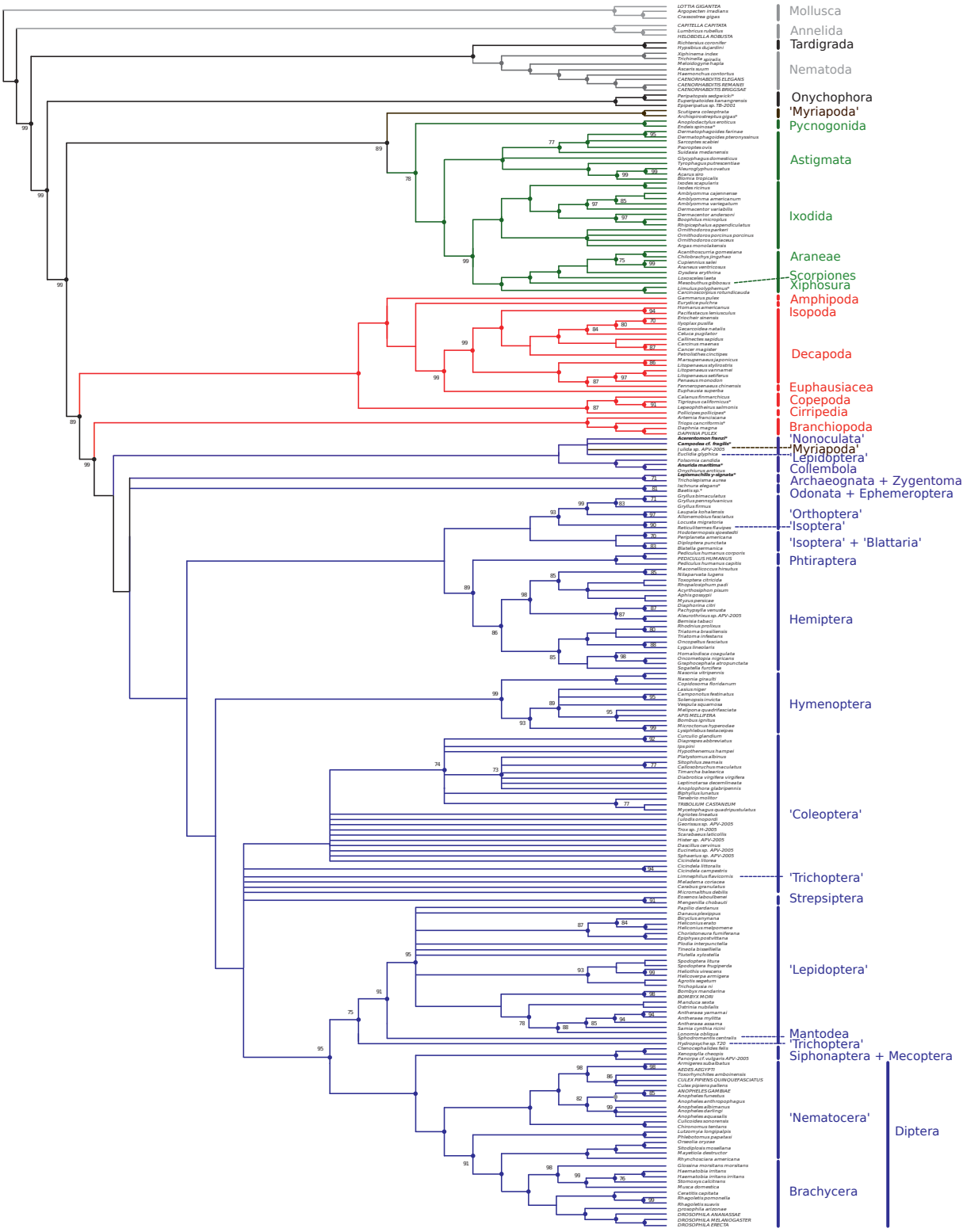

Figure 2.19.: Majority rule cladogram (ML) inferred from the masked data set AP_1 (233 taxa, 775 genes). ML tree search and bootstrapping were conducted with RAxML (see 2.2.7). Support values are derived from 100 bootstrap replicates. Support values < 70: not displayed, support values = 100: represented by a dot only. Quotation marks indicate para- or polyphyletic clades. Color code: mollusks, annelids and nematodes: lighter gray; tardigrades, onychophorans: black; myriapods: brown; chelicerates: green; crustaceans: red; basal hexapods: blue; pterygote insects: dark blue; * indicate EST taxa contributed by members of the 'Arthropod DMP Network'; own EST-taxa are bold printed.

Selected optimal subset (SOS) of AP_1: Maximum Likelihood and Bayesian analyses

In contrast to data set AP_1, trees inferred from the SOS show increased tree robustness with high support for most euarthropod splits (Figs. 2.20-2.21). Tardigrades (*Hypsibius* and *Richtersius*) emerge as a sister group of nematodes (BS 100%, pP 1.0) in both, ML and Bayesian, reconstructions. In both trees, the position of onychophorans is resolved with strong support for a clade (Onychophora,Euarthropoda). Monophyly of euarthropods is strongly supported in both trees (Tab. 2.6), but relationships between myriapods, sea spiders, chelicerates, crustaceans and hexapods remain unresolvedr. While the position of sea spiders is not resolved in the ML tree (Fig. 2.20), the Bayesian tree (Fig. 2.21) shows monophyletic chelicerates including sea spiders with high support (pP 0.99). Pycgnogonids are sister group to Eucheliceratа.Within mandibulates, two alternative clades, either Atelocerata (= Tracheata) or Pancrustacea (= Tetraconata) have been suggested. In the present analyses, the position of myriapods is not resolved. In the Bayesian tree, myriapods are a sister group to chelicerates (including sea spiders) with low support (< pP 0.7). In the ML tree, relationships between myriapods, sea spiders, euchelicerates and pancrustaceans are not resolved.

Both SOS trees suggest a clade Pancrustacea with maximal support while crustaceans are paraphyletic. Within crustaceans, relationships are still far from being resolved. In both trees, Branchiopoda are sister group to Hexapoda, with maximal support in the Bayesian tree and with moderate support (BS 92%) in the ML tree (Tab. 2.6). This clade contradicts results of rRNA based studies Mallatt and Giribet (2006); von Reumont et al. (2009) where copepods have been inferred as a sister group to Hexapoda, albeit weakly supported (chapter 3).

The ML and the Bayesian tree strongly support Hexapoda. Monophyly of Entognatha (Protura, Diplura and Collembola) is generally ambiguous (see review of Grimaldi, 2010). In present analyses, monophyly of Entognatha remains uncertain: in the ML and Bayesian tree, Entognatha are recovered, but albeit moderate or weakly supported (Tab. 2.6). Within Entognatha, there is strong support in both trees for a sister group relationship of Protura and Diplura (Nonoculata, Luan et al., 2005). Monophyly of ectognathous hexapods (Archaeognatha, Zygentoma + pterygote insects, see Hennig (1981) and Kristensen (1991)) "has likewise never been seriously challenged" (Grimaldi, 2010). In the present analyses, Ectognatha (Archaeognatha + pterygote insects) are corroborated with maximal support in both trees (Figs. 2.20-2.21). We should be aware, however, that sufficient EST data of Zygentoma are currently missing.

Addressing relationships of early winged insects, three possible scenarios have been suggested from morphological or molecular studies: Paleoptera ((Odonata,Ephemeroptera),Neoptera), see Hennig (1981) and Kukalová-Peck (1983), Chiastomyaria (Odonata (Ephemeroptera + Neoptera)) (Boudreaux, 1979; Kjer, 2004) or Metapterygota (Ephemeroptera (Odonata, Neoptera)), see Börner (1909) and Zhang et al. (2008). Most molecular analyses support either a Chiastomyaria or a Paleoptera clade. The present phylogenomic data are inconclusive in ML tree reconstructions. Paleoptera, however, are strongly supported in the Bayesian tree. This result contrasts for example with the study of Simon et al. (2009). Their analyses support Chiastomyaria, but this data set has a much smaller taxon sampling on pterygote insects. Within Neoptera, both SOS trees show strong support for the monophyly of all endopterygote orders. Hymenopterans are unambiguously resolved as a sister group to all other endopterygote insects (Figs. 2.20, 2.21).

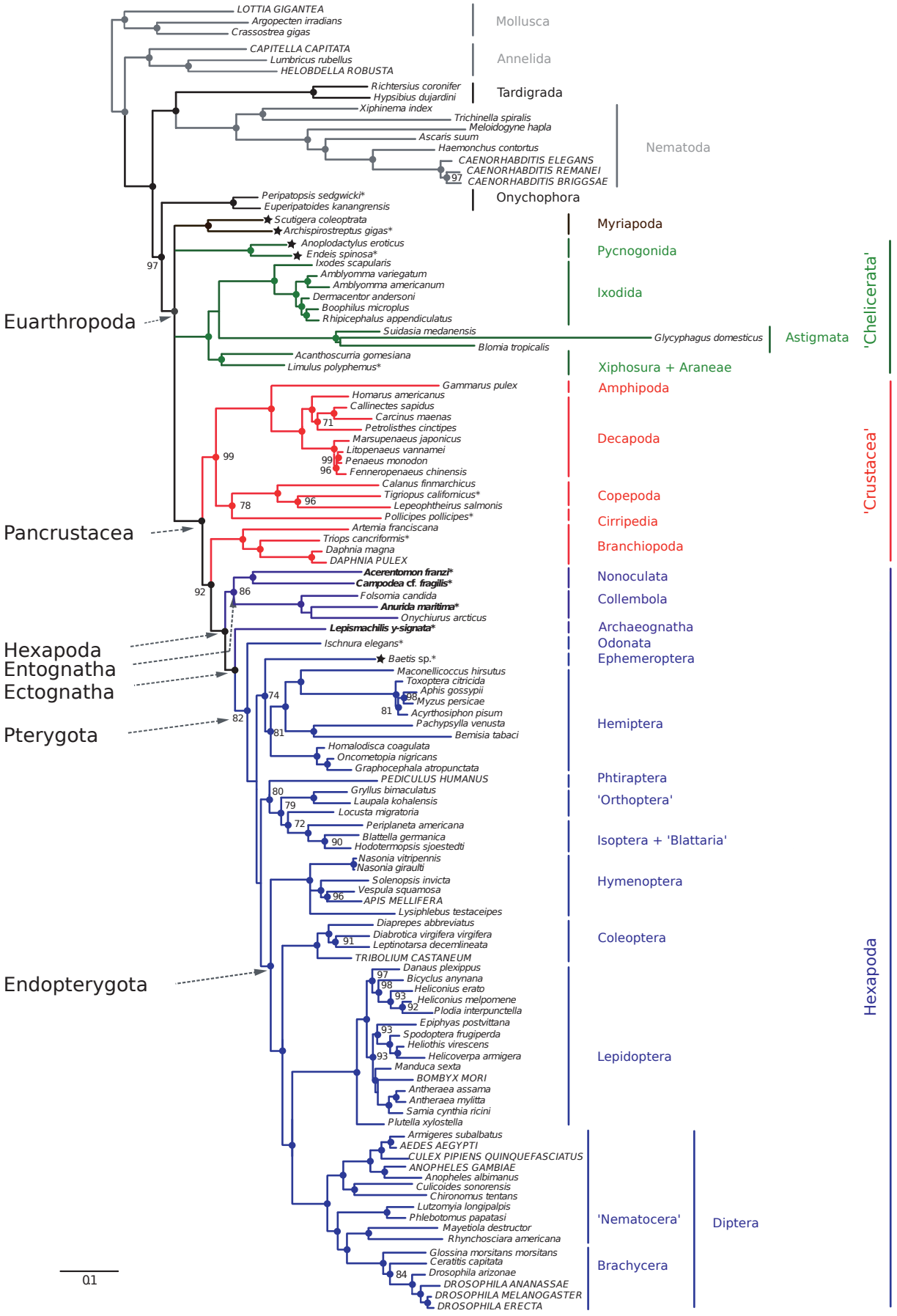

Figure 2.20.: Phylogram (majority rule) inferred from the 117-taxon ML analysis based on the SOS of AP_1. Support values were derived from 1,000 bootstrap replicates. Support values, color code and labeling are specified in Fig. 2.19. 'Unstable' taxa (leaf stability index < 0.95, see 2.2.7) are marked by a star in front of the taxon name.

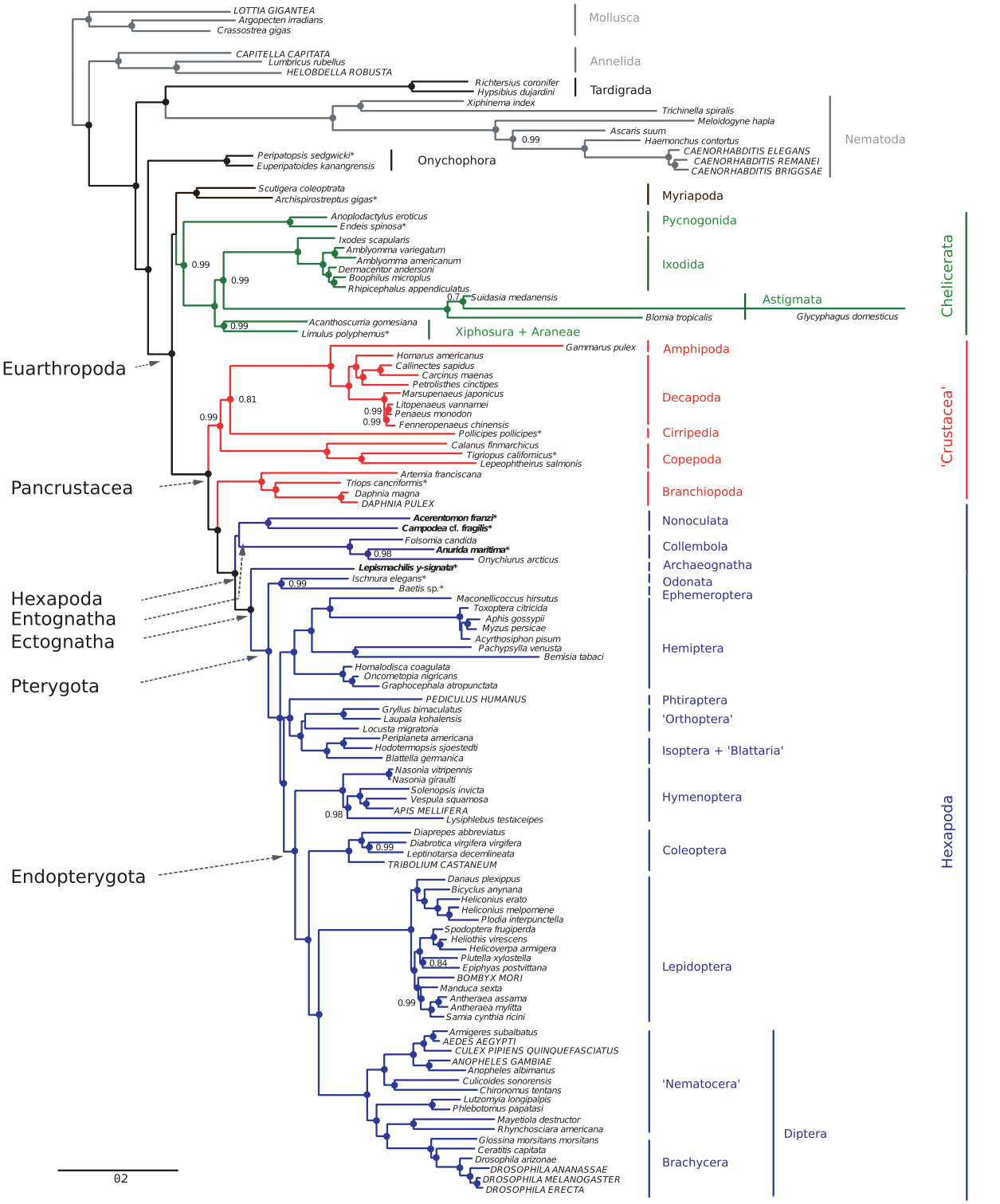

Figure 2.21.: Phylogram (majority rule consensus) inferred from the 117-taxon Bayesian analysis based on the SOS of AP_1. The mrc tree was inferred from a 'triple' chain (3 out of 25 chains) with the lowest *maxdiff* value (0.186). Concurrently, these chains showed best log-likelihood values (harmonic means, burn-in excluded) of all possible 'triple' chains. For each chain, 20,000 cycles were sampled, the burn-in was excluded (5,000 cycles). Posterior probabilities (pP) were estimated under the CAT mixture model (see 2.2.7. Values < 0.7: not shown, values = 1.0: represented by a dot. Color code and labeling are similar to Fig. 2.19.

Table 2.6.: Support of selected clades of ML and Bayesian SOS tree. Bootstrap support (BS in [%]) derived from 1,000 bootstrap replicates; posterior probabilities (pP) derived from a selected triple chain of Bayesian analyses (20,000 cycles / chain, burn-in excluded).

Selected clades	**Bootstrap Support**	**posterior Probability**
(Tardigrada,Nematoda)	100	1
(Onychophora,Euarthropoda)	97	1
((Tardigrada,Nematoda),(Onychophora,Euarthropoda))	100	1
Euarthropoda	100	1
Mandibulata	-	-
Myriochelata	-	0.57
Chelicerata	-	0.99
Eucheliceratа	100	1
Pancrustacea	100	1
(Amphipoda,Decapoda)	100	1
(Copepoda,Cirripedia)	78	-
((Amphipoda,Decapoda),(Copepoda,Cirripedia))	99	-
((Amphipoda,Decapoda),Cirripedia)	-	0.81
(((Amphipoda,Decapoda),Cirripedia),Copepoda)	-	0.99
(Branchiopoda,Hexapoda)	92	1
Hexapoda	100	1
Entognatha	86	0.5
(Collembola,(Protura,Diplura))	86	0.5
Nonoculata: (Protura,Diplura)	100	1
Ectognatha: (Archaeognatha,Pterygota)	100	1
Pterygota	82	1
Chiastomyaria: (Odonata,(Ephemeroptera,Neoptera))	-	-
Paleoptera: (Odonata,Ephemeroptera)	-	0.99
Neoptera	-	1
(Ephemeroptera,Hemiptera)	74	-
Endopterygota	100	1
(Hymenoptera,remaining endopterygote clades)	100	1
(Coleoptera,(Lepidoptera,Diptera))	100	1
(Lepidoptera,Diptera)	100	1

Bayesian consensus tree and incongruencies among Bayesian analyses

Discrepancy across all bipartitions (*maxdiff*) was checked by pairwise comparison and comparing triple chains with *bpcomp*, implemented in PhyloBayes (Lartillot et al., 2008). Triple chain c04−c18−c20 (*maxdiff* value = 0.186) was selected to infer a majority rule consensus (mrc) tree. These three chains show the best log likelihoods (harmonic means) of all chains used for triple chains (Tab. 2.7). The 25 Bayesian runs did not converge on a single topology (Tab. 2.8). Incongruent consensus trees might reflect different local optima of single chains. Topological differences among single runs cannot be displayed in a consensus tree (see Fig. 2.21). Instead, a consensus network (Holland and Moulton, 2003, see Fig. 2.22) is appropriate to visualize incongruencies. Some clades, e.g. (Onychophora,Euarthropoda), Pancrustacea, Branchiopoda as a sister group to Hexapoda and Nonoculata are robustly resolved in all trees inferred from single chains with maximal support. Incongruencies addressing other clades are caused by unstable positions of few taxa between single consensus trees (Fig. 2.21): (1) Mandibulata (Myriapoda,Pancrustacea) are maximally supported in consensus trees of two chains (c09, c15). Both chains show comparatively low harmonic means of log-likelihoods

(Tab. 2.8). In both chains, a clade (Mandibulata,Chelicerata) is negligibly supported (pP 0.52 and 0.55). In remaining chains, myriapods cluster with chelicerates with negligible or moderate support (pP 0.52–0.89). (2) The barnacle *Pollicipes* (Cirripedia) is suggested as a sister group of copepods in one chain, albeit weakly supported (pP 0.51). An alternative clade (Pollicipes,Malacostraca) gets various support (0.56–0.96 pP, Tab. 2.8). (3) in some single chain trees, the bristletail *Lepismachilis* (Archaeognatha) is suggested as a sister group of a clade joining isopterans and blattarian insects with moderate or low support (pP 0.52–0.82, Tab. 2.8). *Pediculus* (Phthiraptera) emerge as a sister group to this clade with maximal support in trees from respective chains. Log-likelihoods (harmonic means) of these chains are remarkably lower than log-likelihoods of chains used for the mrc tree (see Fig. 2.21 and Tab. 2.7). These chains have been rejected in a Bayes factor test (Kaas and Raftery, 1995; Nylander et al., 2004). (4) Among butterflies (Lepidoptera), five different topologies with distinctive clades are suggested. Differences occur among Yponomeutoidea, Papilionoidea, Pyraloidea, Tortricoidea and Noctuidea.

Leaf stability indices for ML analysis

Calculation of leaf stability indices (LSIs) was conducted for the ML analysis of the AP_1-SOS using all bootstrap trees (see 2.2.7). Values indicate the frequency a taxon occurs in the same position among investigated trees. A value of 1 means that a taxon has a unique position among investigated trees. The threshold for instability of a taxon is arbitrarily defined: here, taxa with an LSI < 0.95 have been defined as 'unstable'. Five taxa of the AP_1-SOS are unstable: pycnogonids (*Anoplodactylus* and *Endeis*), myriapods (*Archispirostreptus* and *Scutigera*) and the mayfly *Baetis* (Tab. 2.9). LSIs of orthopterans, isopterans, blattarian insects, the louse *Pediculus* and the bristletail *Lepismachilis* show values around 0.952–0.955. However, the validity of this instability index are calculated from triplet likelihood mapping (Thorley and Wilkinson, 1999; Thorley and Page, 2000) and has to be critically considered.

Pruning of unstable taxa from the SOS of AP_1 provides neither topological differences nor differences in support for remaining taxa. To test the impact of 'unstable' taxa, ML analysis (1,000 bootstrap replicates) was repeated excluding 'unstable' taxa. This procedure has been also suggested by Dunn et al. (2008): they propose the identification of a 'robust backbone tree' based on sufficient signal among included taxa. In contrast to Dunn's study, the exclusion of 'unstable' taxa in present analysis causes few, but important topological differences. The position of velvet worms and the position of the horseshoe crab *Limulus polyphemus* is aberrant (Fig. A.4). *Limulus* is placed in a basal position within euchelicerates. Onychophorans slip down and emerge as a sister group to a clade ((Nematoda,Tardigrada),Euarthropoda) with maximal support. (Nematoda,Tardigrada) are suggested as a sister group to Euarthropoda with negligible support (BS 56%). A clade (Xiphosura,Araneae), previously inferred as a sister group to all other eucheliceratates with high support (Figs. 2.20, 2.21), clusters as a sister group of Ixodida with moderate support (BS 85%). Astigmata branch off first within eucheliceratates. Bootstrap values change for several clades throughout the entire topology. A general decrease or increase of bootstrap values is not perceptible. A recalculation of LSIs results in 'instability' of both onychophorans (LSI: 0.8638886158886). Summarized, the removal of 'unstable' taxa does not increase tree robustness and reliability for the present data set. Instead, highly suspicious clades occur, e.g. (Onychophora(Tardigrada,Nematoda)).

Table 2.7.: Log-likelihood values and triples of PhyloBayes chains from the SOS of `AP_1`. Chain ID: Identifier. triple chain: a triple chain consists out of 3 chains, *maxdiff* < 0.3. *maxdiff*: discrepancy value across all bipartitions for given triples. Selected chains and triples for the mrc tree: blue-printed.

chain ID	log-Likelihood (harmonic mean)	triple chain	maxdiff (< 0.3)
c18	948174.861012454	c05 - c14 - c16	0.162100
c04	948217.993492174	c04 - c18 - c20	0.186000
c20	948376.710282837	c22 - c05 - c16	0.186530
c16	948469.642382507	c21 - c23 - c01	0.188330
c06	948491,752015474	c23 - c01 - c06	0.202933
c05	948525.74067471	c21 - c23 - c08	0.207870
c22	948678.821621205	c01 - c23 - c08	0.207870
c23	948708.71215524	c21 - c08 - c01	0.207870
c21	948752.989770425	c22 - c05 - c14	0.236470
c08	948757.764925626	c22 - c14 - c16	0.236470
c14	948779.209757328		
c01	948865.845517544	all 25 chains	1

Table 2.8.: Selected clades and support values of single PhyloBayes chains inferred from the SOS of `AP_1`. Chains with a *maxdiff* value < 0.3 within a triple: bold printed; selected chains for the mrc tree: blue, bold-printed; log LH: harmonic mean of log likelihood values. MyCh: Myriochelata; Ma: Mandibulata; Che: Chelicerata; Pa: Pancrustacea; Cirr: Cirripedia, Mala: Malacostraca; H: Hexapoda, Ento: Entognatha; No: Nonoculata; Lepis: *Lepismachilis*; Pt: Pterygota; Ped: *Pediculus*; I/B: mixed clade of isopterans and blattarians; Paleo: Paleoptera.

chain ID	log LH	support value (posterior probability) of selected clades									
		MyCh	Ma	Che	(Cirr + Mala)	H	Ento	(Lepis + Pt)	(Ped(Lepis + I/B))	(Lepis + I/B)	Paleo
c17	948096.9	0.89	-	0.99	0.56	1	0.54	1	-	-	1
c18	948174.9	0.64	-	0.99	0.77	1	< 0.5	1	-	-	1
c04	948218	0.52	-	1	0.74	1	0.55	1	-	-	1
c11	948247.5	0.60	-	0.99	0.91	1	0.59	1	-	-	1
c20	948376.7	0.54	-	0.99	0.92	1	0.5	1	-	-	0.99
c13	948378.6	0.66	-	0.99	-	1	0.55	1	-	-	1
c16	948469.6	0.78	-	0.98	0.59	1	< 0.5	-	1	0.74	0.99
c03	948487.7	0.63	-	0.99	0.82	1	< 0.5	1	-	-	0.99
c06	948491.8	0.52	-	0.99	0.90	1	< 0.5	-	1	0.76	1
c05	948525.7	0.65	-	0.99	0.70	1	< 0.5	-	1	0.73	1
c12	948557.6	0.66	-	0.99	0.61	1	< 0.5	1	-	-	1
c07	948566.4	0.56	-	0.99	0.56	1	0.5	1	-	-	1
c09	948587	-	1	0.99	0.92	1	< 0.5	1	-	-	0.99
c19	948644.6	0.73	-	0.99	0.87	1	< 0.5	-	1	0.69	1
c22	948678.8	0.71	-	0.99	0.76	1	< 0.5	-	1	0.82	0.99
c15	948696.1	-	1	0.99	0.84	1	< 0.5	-	1	0.52	1
c23	948708.7	0.68	-	0.98	0.94	1	< 0.5	-	1	0.69	1
c02	948735.6	0.75	-	0.89	0.74	1	< 0.5	-	1	0.52	1
c21	948753	0.86	-	0.99	0.82	1	< 0.5	-	1	0.76	1
c08	948757.8	0.85	-	0.99	0.96	1	< 0.5	-	1	0.71	1
c14	948779.2	0.79	-	0.99	0.73	1	< 0.5	-	1	0.59	1
c24	948799.5	0.86	-	0.98	0.86	1	< 0.5	-	1	0.58	1
c01	948865.8	0.73	-	1	0.89	0.5	-	-	1	0.79	1
c25	948925.8	0.79	-	0.98	0.57	1	< 0.5	-	1	0.75	1
c10	949085.1	0.61	-	0.99	0.57	1	< 0.5	-	1	0.69	0.99

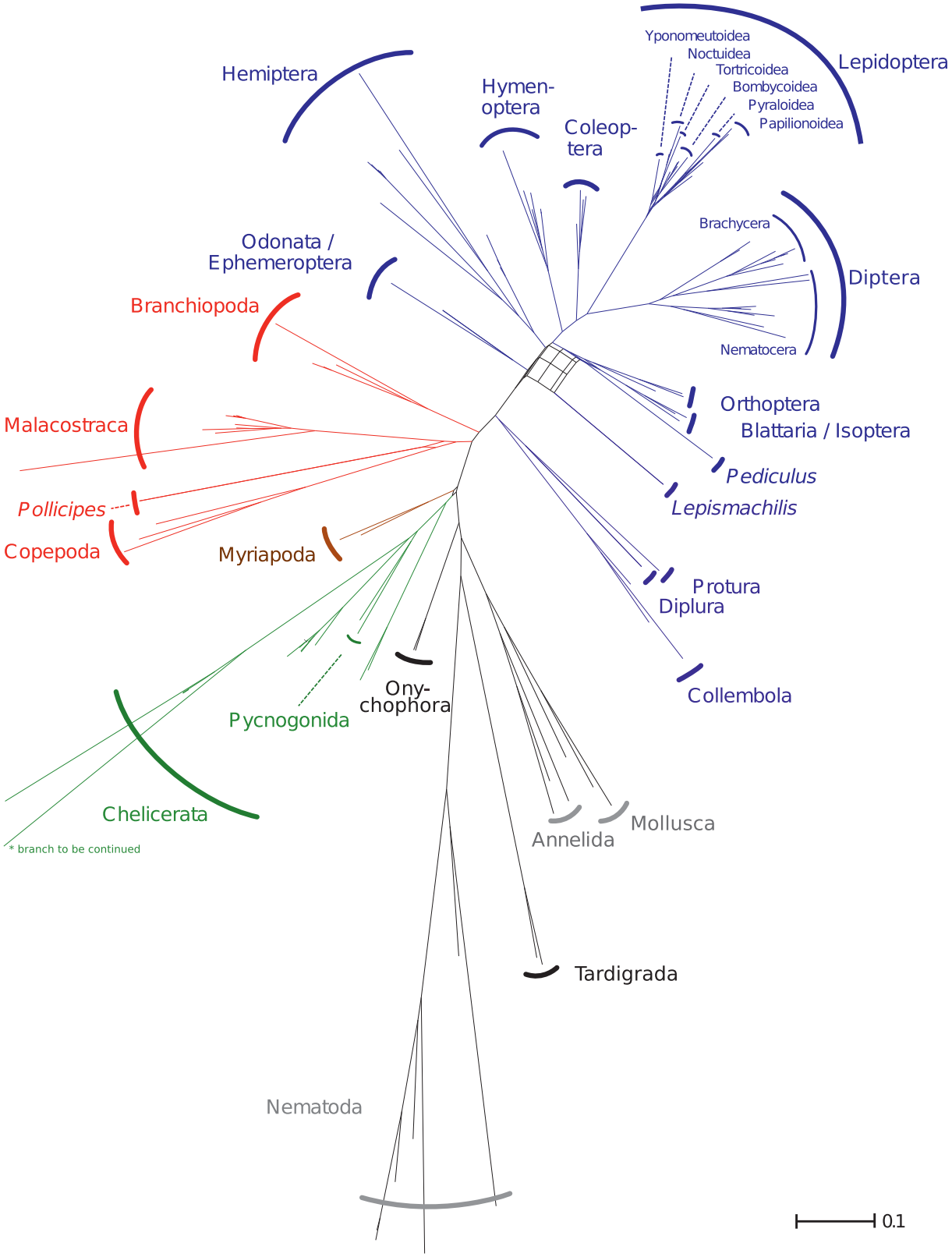

Figure 2.22.: Consensus network calculated from all 25 single Phylobayes consensus trees based on the AP_1-SOS. The network was calculated with SplitsTree 4.8 (threshold = 0.01 with averaged edge weights) and displays inconsistencies between single topologies. Color code: mollusks, annelids and nematodes: gray; tardigrades, onychophorans: black; myriapods: brown; chelicerates: green; crustaceans: red; apterygote hexapods: blue; pterygote insects: dark blue, see also Fig. 2.16.

Table 2.9.: Leaf stability indices (LSIs) of selected taxa (ML analysis, SOS, AP_1). LSIs were calculated with Phyutility (Smith and Dunn, 2008) based on 1,000 collected bootstrap trees. The threshold for 'instability' was arbitrarily set (< 0.95). Taxa with LSIs $<$ 0.96 are listed. Group specification follows the NCBI taxonomy.

species	group*	class / order	Leaf Stability Index (LSI)
Archispirostreptus gigas	Myriapoda	Diplopoda	0.8601010494753
Scutigera coleoptrata	Myriapoda	Chilopoda	0.8601010494753
Anoplodactylus eroticus	Chelicerata	Pycnogonida	0.9081712143928
Endeis spinosa	Chelicerata	Pycnogonida	0.9081712143928
Baetis sp.	Hexapoda	Ephemeroptera	0.9373833583209
Lepismachilis y-signata	Hexapoda	Archaeognatha	0.9522524737631
Pediculus humanus	Hexapoda	Phthiraptera	0.9523611694153
Locusta migratoria	Hexapoda	Orthoptera	0.9544979010495
Gryllus bimaculatus	Hexapoda	Orthoptera	0.9547434782609
Laupala kohalensis	Hexapoda	Orthoptera	0.9547434782609
Periplaneta americana	Hexapoda	Blattaria	0.9547899550225
Blattella germanica	Hexapoda	Blattaria	0.9547989505247
Hodotermopsis sjoestedti	Hexapoda	Isoptera	0.9547992503748

Ribosomal and non-ribosomal data subsets. Both data subsets, AP_1_ri and AP_1_nri, are based on the SOS of AP_1. Maximum likelihood analyses of AP_1_ri (32 ribosomal genes, 0.522 total average information content, 82.9% matrix saturation) and AP_1_nri (97 non-ribosomal genes, 0.4 total average information content, 55.5% matrix saturation) provide different topologies. The consensus network (Fig. 2.23) from both trees, displays contradicting signal for several clades.

Both topologies are similar for most clades. For example, Nonoculata is highly supported in the ribosomal and in the non-ribosomal tree. This is also true for Endopterygota and Hymenoptera, branching off first within Endopterygota. Contradictory signal that produce 'cobwebs' in a consensus network, especially concerns myriapods with *Scutigera* (Chilopoda, Myriapoda) and *Archispirostreptus* (Diplopoda, Myriapoda), entognathous hexapods and the position of *Lepismachilis* (bristletail). The placement of *Lepismachilis* seems correlated with the position of Odonata, Ephemeroptera, the louse *Pediculus*, Cicadellidae (Hemiptera) and a mixed clade of blattarians, orthopterans and Isoptera. Clades occurring in only one of both trees (Figs. A.5, A.6) cause following major topological inconguencies: (1) *Archispirostreptus* is sister group to Pancrustacea (weakly supported) in the non-ribosomal tree. *Scutigera* groups within a clade with Pycnogonida and Euchelicerata, but relationships are not resolved. Both myriapods share only two non-ribosomal genes with moderate information content (gene 11841: 0.46; gene 11383: 0.53. In contrast, monophyletic Myriapoda are suggested from the ribosomal tree, maximally supported. Both myriapods have 23 genes in common. Myriapoda are positioned within Pycnogonida and Euchelicerata, but relationships between Myriapoda, Pycnogonida and Euchelicerata are not resolved. (2) *Pollicipes* (Cirripedia) shows as a sister group relationship with copepods in the non-ribosomal tree (as well as before in the ML tree, Fig. 2.20). In the ribosomal tree, *Pollicipes* is sister group to Malacostraca with low support. This clade has been previously recovered in the Bayesian tree (Fig. 2.21). (3) Entognatha (Collembola,Nonoculata), is

only present in the non-ribosomal topology and moderate supported. Entognathous orders share 8 non-ribosomal genes (information content ranges from 0.57 to 0.92). Within the ribosomal tree, the position of springtails is not resolved. Here, Protura, Diplura and Collembola share 14 ribosomal genes (information content varies between 0.42 and 0.77). In the non-ribosomal tree, *Folsomia* (Entomobryoidea) is the most basal split within Collembola. In contrast, *Anurida* (Poduroidea) is the most basal split in the ribosomal topology. Thus, Poduroidea would be paraphyletic in the latter scenario. (4) *Lepismachilis* is suggested as a sister group to pterygote insects (maximally supported) only in the ribosomal tree. In contrast, it clusters with *Periplaneta* and the louse *Pediculus* in the non-ribosomal tree. The resolution is poor among deep pterygote splits (Odonata, Ephemeroptera, hemipterans, orthopterans, *Pediculus*, etc.).

In general, resolution and support is lower for both separate ML trees compared with the complete SOS ML tree of `AP_1`. Bootstrap values of some clades present in both trees, partially strongly vary. For example, Euarthropoda or Hexapoda have support values that differ more than 40% comparing both trees (see Tab. 2.10). For other clades, support is similar (Pancrustacea, Endopterygota), or only marginally differs (see Nonoculata) if both topologies are compared.

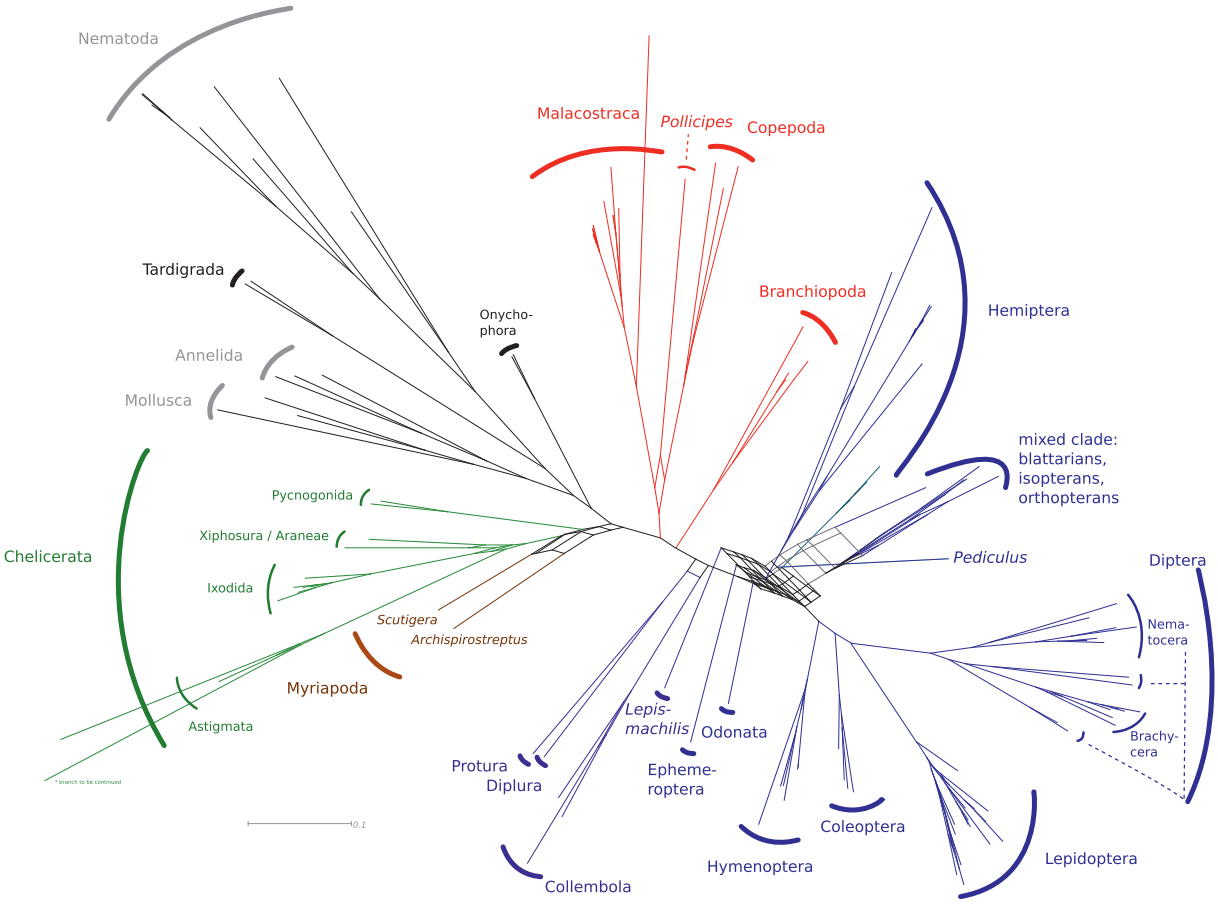

Figure 2.23.: Consensus network of the ribosomal and non-ribosomal ML tree inferred from data subsets `AP_1_ri` and `AP_1_nri`. The network was calculated with SplitsTree 4.8 (threshold = 0.01, averaged edge weights). It displays inconsistencies between both trees. Color code: mollusks, annelids and nematodes: black; tardigrades: lighter gray, onychophorans: gray; myriapods: brown; chelicerates: green; crustaceans: red; apterygote hexapods: blue; pterygote insects: dark blue; *Pediculus*: sea-blue; Clypeorrhyncha, Cicadellidae within Hemiptera: turquoise. Clades do not reflect necessarily monophyletic clades.

Table 2.10.: Selected clades and BS support of ML trees inferred from data subsets AP_1_ri (32 ribosomal genes, total average information content: 0.522, matrix saturation: 82.9%) and AP_1_nri (97 non-ribosomal genes, total average information content: 0.4, matrix saturation: 55.5%), compared with the 'full' SOS of AP_1 (129 genes). BS: Bootstrap support.

Selected clades recovered in both data subsets AP_1_ri and AP_1_nri	**BS [%]** AP_1_ri	**BS [%]** AP_1_nri	**BS [%]** AP_1, SOS
(Tardigrada,Nematoda)	79	100	100
(Onychophora,Euarthropoda)	99	55	97
((Tardigrada,Nematoda),(Onychophora,Euarthropoda))	100	100	100
Euarthropoda	51	98	100
Pancrustacea	100	100	100
(Branchiopoda,Hexapoda)	54	92	92
Hexapoda	60	100	100
Nonoculata	93	97	100
Endopterygota	100	100	100
(Hymenoptera,remaining endopterygote clades)	99	100	100

Selected optimal subset of AP_2 (SOS). In data set AP_2 (Tab. 2.3, 2.4), the velvet worm *Epiperipatus* had been additionally defined as constraint and included in tree reconstruction. In few instances, the topology differs with respect to that of SOS-AP_1 : (1) Pycnogonida emerges as a sister group to Eucheliceratа: chelicerates are monophyletic, albeit weakly supported (BS 64%). The relationships between (Xiphosura,Araneae), Astigmata and Ixodida are not resolved, in contrast to the SOS-AP_1 topology. (2) With *Epiperipatus* included, Mandibulata are inferred. Myriapoda (Diplopoda,Chilopoda) are proposed as a sister group to Pancrustacea. However, Mandibulata only shows negligible support (BS 50%, Fig. 2.24). (3) The hymenopteran species *Lysiphlebus* (Ichneumonoidea) shows a sister group relationship to paraphyletic Vespoidea + Apoidea, but with negligible support (BS 53%, Fig. A.7).

In general, about one third of all bootstrap values marginally differ with respect to the SOS-AP_1 tree. Mostly, bootstrap supports are lower, but not more than 5% with one exception. For some clades bootstrap support is increased, e.g. for Entognatha (BS 90% instead of 86%, cf. Fig. 2.20 and Fig. A.7). For *Epiperipatus*, only a low number of genes are present (33 of 127 genes, information content: 0.92–0.44, averaged 0.65). Since this SOS has only two genes less, the impact of inclusion of this single taxon, (*Epiperipatus*) on the topology is astonishing. It might reflect the instability of certain splits, e.g. Mandibulata or Myriochelata.

Selected optimal subset of AP_3_oP (SOS). Exclusion of all proteome species in data set AP_3_oP results in an SOS which slightly differs in taxa- and gene selection compared with SOS-AP_1 (section 2.3.1, Fig. 2.12). In tree reconstruction, Pancrustacea, Hexapoda, Ectognatha, Endopterygota or Hymenoptera are robustly resolved. Hymenoptera branch off first within endopterygote insects. Topology and support values differ in following instances: (1) Euarthropoda show a remarkable weak

support (BS 67%). (2) Myriochelata are monophyletic, but negligibly supported. Chelicerata, including Pycnogonida, are monophyletic. Pycnogonida are sister group to Eucheliceratа with maximal support (Tab. 2.11, Fig. 2.25).

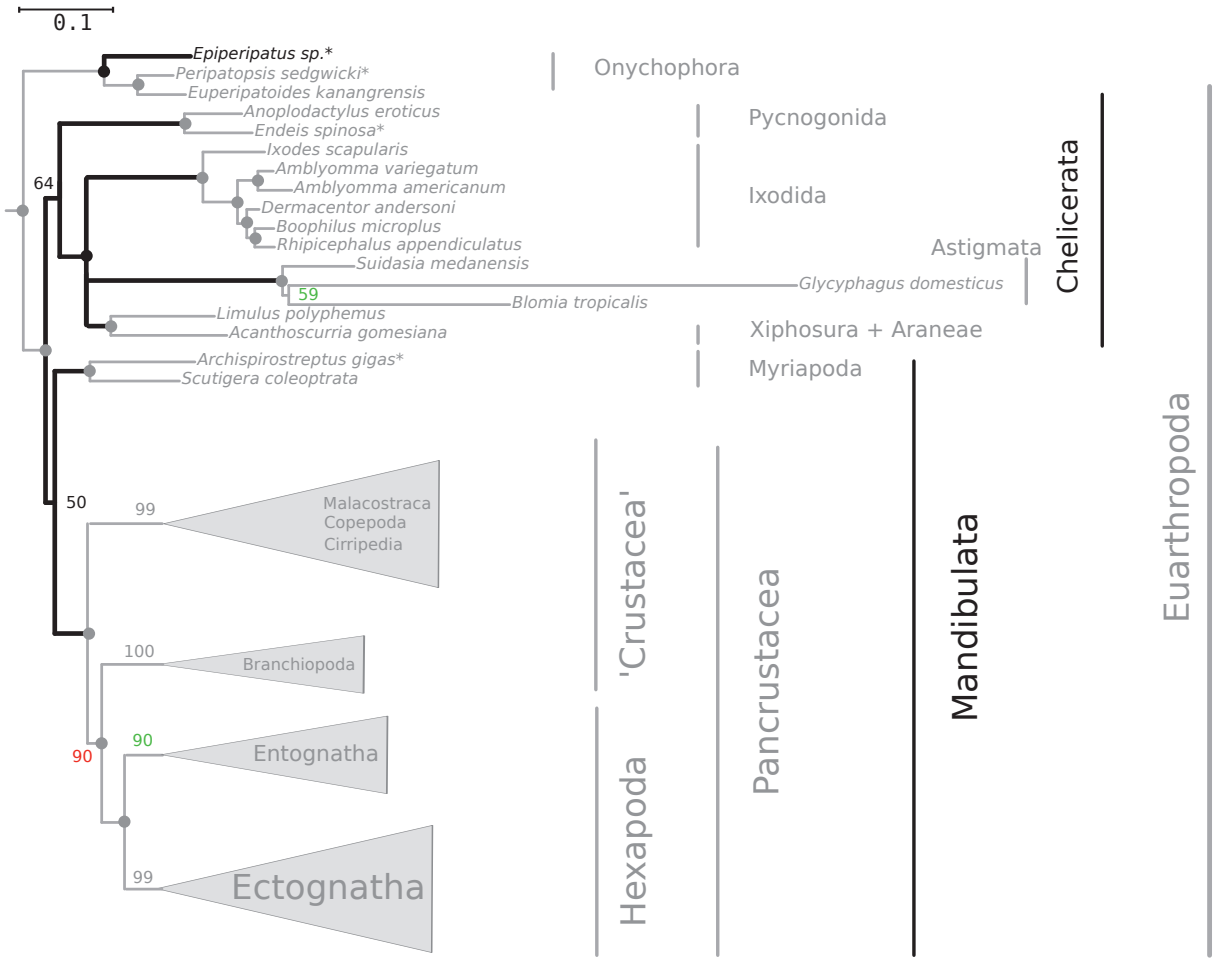

Figure 2.24.: Schematic section of the phylogram (majority rule, 1,000 bootstrap replicates) of 118-taxon ML analysis (127 genes) based on the SOS of AP_2. Topological changes compared with Fig. 2.20 are in black; increased BS values: green; decreased BS values: red. Asterisks * indicate EST taxa contributed members of the 'Arthropod DMP Network'; own EST-taxa are in bold-print. Support values and labeling are specified in Fig. 2.19. For the complete topology see Fig. A.7.

The support for Eucheliceratа is remarkably lower than in the SOS tree of AP_1. (3) Branchiopoda are suggested as a sister group to Hexapoda with negligible support. (4) Within primary wingless hexapods, support for Entognatha increases while support for Nonoculata slightly decreases. (5) Within pterygote insects, the damselfly *Ischnura*, the mayfly *Baetis* and the louse *Pediculus* branch off first, but relationships are unresolved. Pterygota are maximally supported. Within hemimetabolans, most relationships are moderately supported. The flea *Ctenocephalides* is proposed as a sister group to (Lepidoptera,Diptera). The support for Lepidoptera and a clade (Lepidoptera,Diptera) is considerably lower than in the AP_1-SOS tree. Leaf stability indices (Smith and Dunn, 2008) for this data set identify following taxa as 'unstable' (LSI $< 95\%$): myriapods, pantopods, the barnacle *Pollicipes*, all included branchiopods and, within pterygote insects, the mayfly *Baetis*, the louse *Pediculus*, Isoptera, blattarians as well as all orthopteran insects.

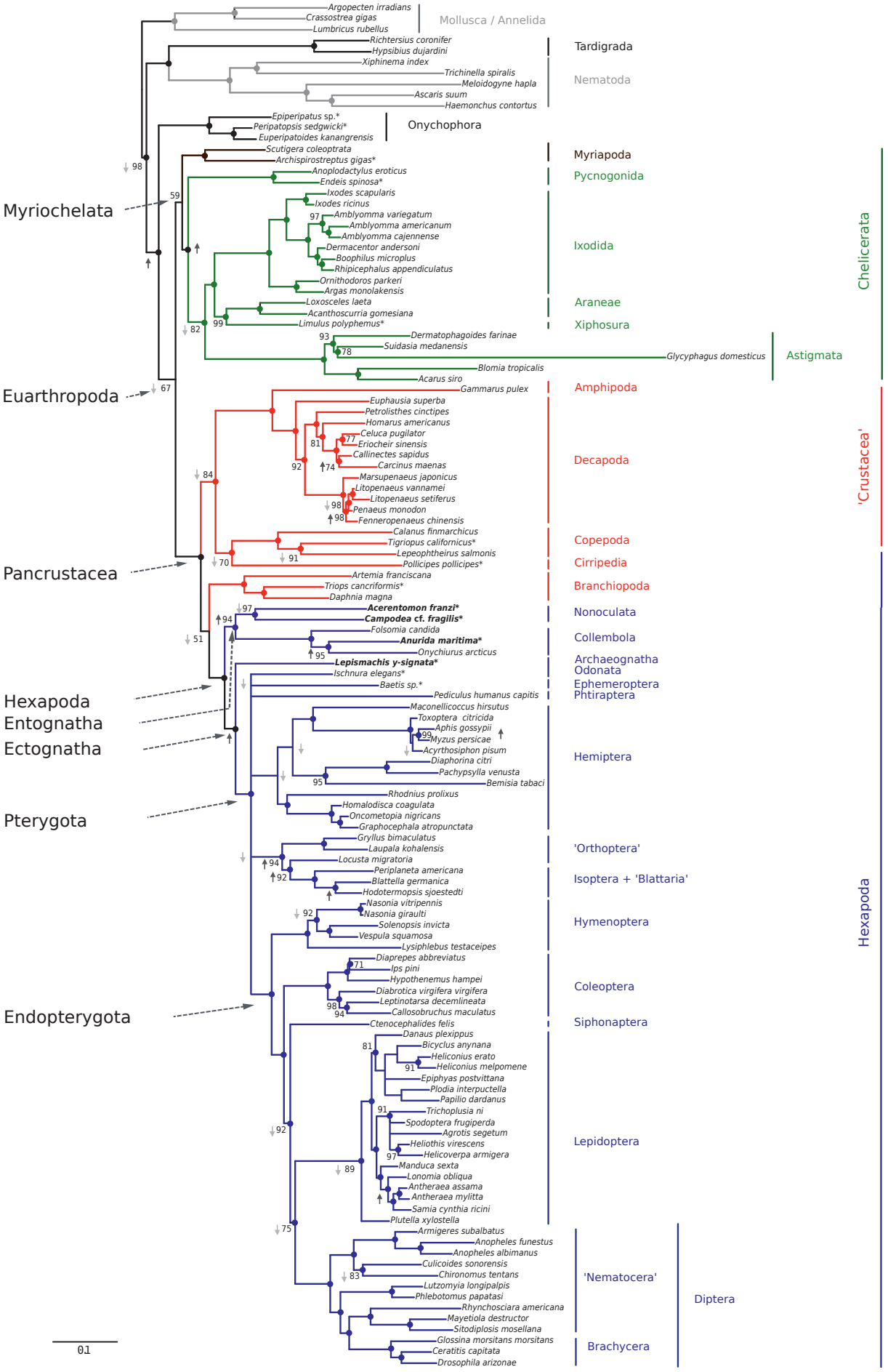

Figure 2.25.: Phylogram (majority rule) of 125-taxon ML analysis (100 genes, 1,000 bootstrap replicates) based on the SOS of AP_3_oP. Black arrows: increased BS values, gray arrows: decreased BS values compared to SOS of AP_1 (Fig. 2.20). For color code and labeling see Fig. 2.19.

Table 2.11.: Selected clades and bootstrap values from ML analysis of the AP_3_oP SOS (proteome species excluded, 125 taxa, 100 genes) compared with the AP_1 SOS.

Selected clades of SOS data subsets	**Bootstrap support** AP_3_oP, SOS	**Bootstrap support** AP_1, SOS
Euarthropoda	67	100
Myriochelata	59	-
Chelicerata	100	-
Euchelicerata	82	100
crustaceans excluding Branchiopoda	84	99
(Cirripedia,Copepoda)	70	78
(Branchiopoda,Hexapoda)	51	92
Entognatha: (Collembola,Nonoculata)	94	86
Nonoculata	97	100
Ectognatha	100	100
Pterygota	100	82
Lepidoptera	89	100
(Lepidoptera,Diptera)	75	100

Endopterygote data sets

The topological differences are marginal between trees including and excluding *Daphnia pulex*. This is true for both, the original data sets and respective SOS. Differences between the original data sets occur mainly for bootstrap support values. Three topological differences are perceptible, one each within hymenopterans, lepidopterans and dipterans. Exemplarily, only differences between the original data set En_oP_Da and its SOS including *Daphnia* are outlined. For topologies inferred from data set En_oP and its SOS (Figs. A.8, A.9). In many instances, the topology inferred from En_oP_Da is unresolved (Fig. 2.26). The butterfly *Euclidia* (Noctuidea, covering only 16 of 775 genes) is placed as a sister group to all ensiferans (Orthoptera). Thus, endopterygotans are not monophyletic, lepidopterans are polyphyletic. A suspicious placement of *Euclidia* was already obtained in a previous tree (Figs. 2.18, 2.19). Relationships within hymenopterans and coleopterans and between hymenopterans, coleopterans and strepsipterans are unresolved. Lepidopterans (excluding *Euclidia*) are placed as a sister group to paraphyletic mecopterans, but with weak support (BS 51%). The mecopteran *Panorpa* is a sister group of Siphonaptera with high support. (Mecoptera,Siphonaptera) is inferred as a sister group to monophyletic Diptera with high support (BS 99%). Within dipterans, Brachycera are monophyletic with maximal support; Nematocerans are not monophyletic caused by the placement of *Phlebotomus* and *Lutzomyia*.

In contrast, the topology of the SOS of En_oP_Da shows mostly well resolved splits (Fig. 2.27). All orders are monophyletic with maximal support. Within Endopterygota, Hymenoptera branch off first and are suggested as a sister group to remaining insect orders. Lepidoptera are sister group of Diptera. Within Diptera, Nematocera are paraphyletic as previously inferred. Considering the resolution of clades, the SOS provide much more robust subtrees compared with the original data set (Fig. 2.26). Thus, selecting an optimal data subset (SOS) pays of its effort for present endopterygote data sets.

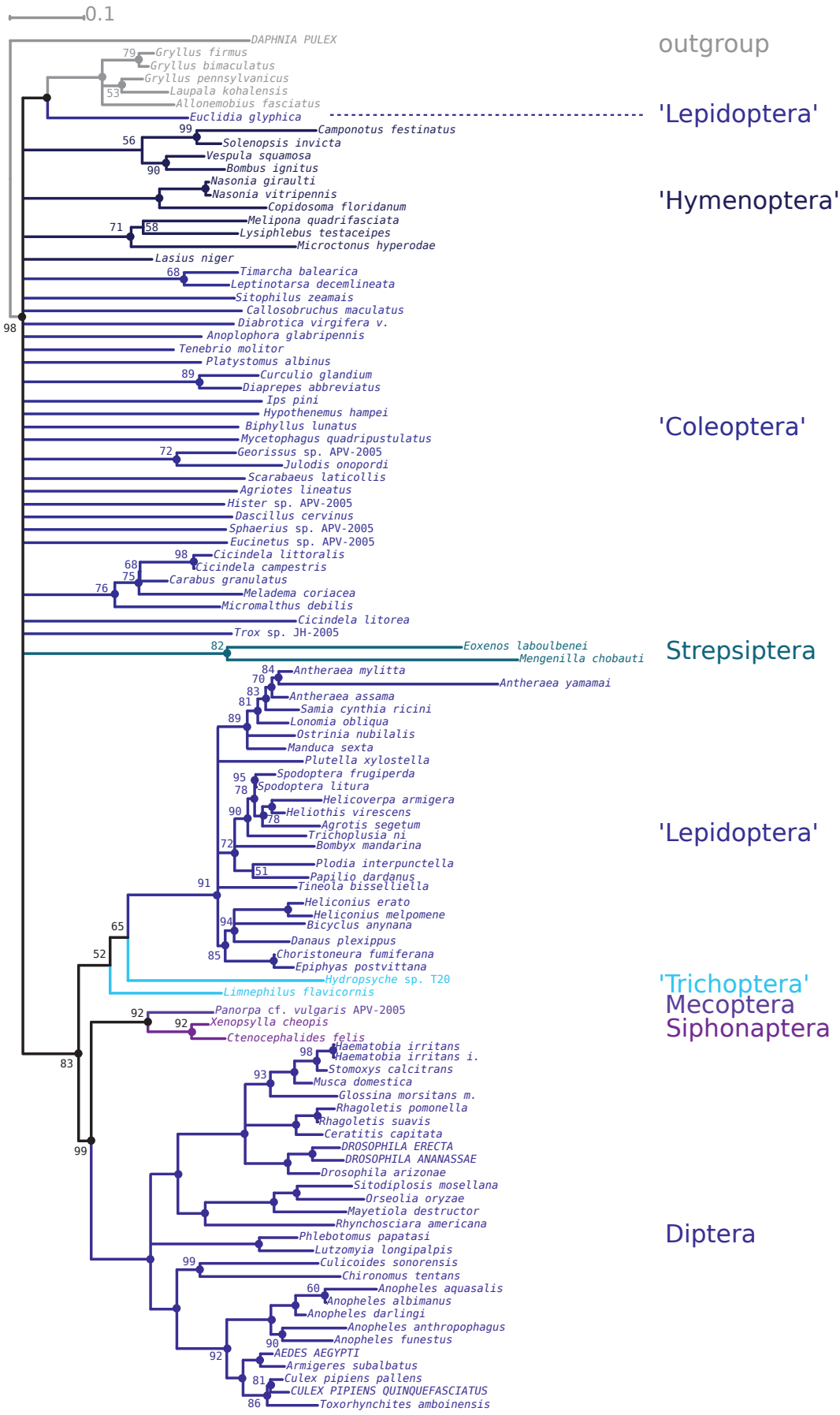

Figure 2.26.: Phylogram (majority rule) of 107-taxon ML analysis (775 genes, 100 bootstrap replicates) based on data set `En_oP_Da`. Outgroup species: *Daphnia* and ensiferans (Orthoptera), the tree was rooted with *Daphnia*. Support values are derived from 100 bootstrap replicates. Support values = 100: a dot only; support values < 70 labeled without dot; Quotation marks: clades are not monophyletic. Color code: outgroups: gray; hymenopterans: dark blue; coleopterans: navy blue; strepsipterans: sea blue; lepidopterans: royal blue; trichopterans: turquoise; mecopterans: lilac; siphonapterans: violet, dipterans: mid-blue.

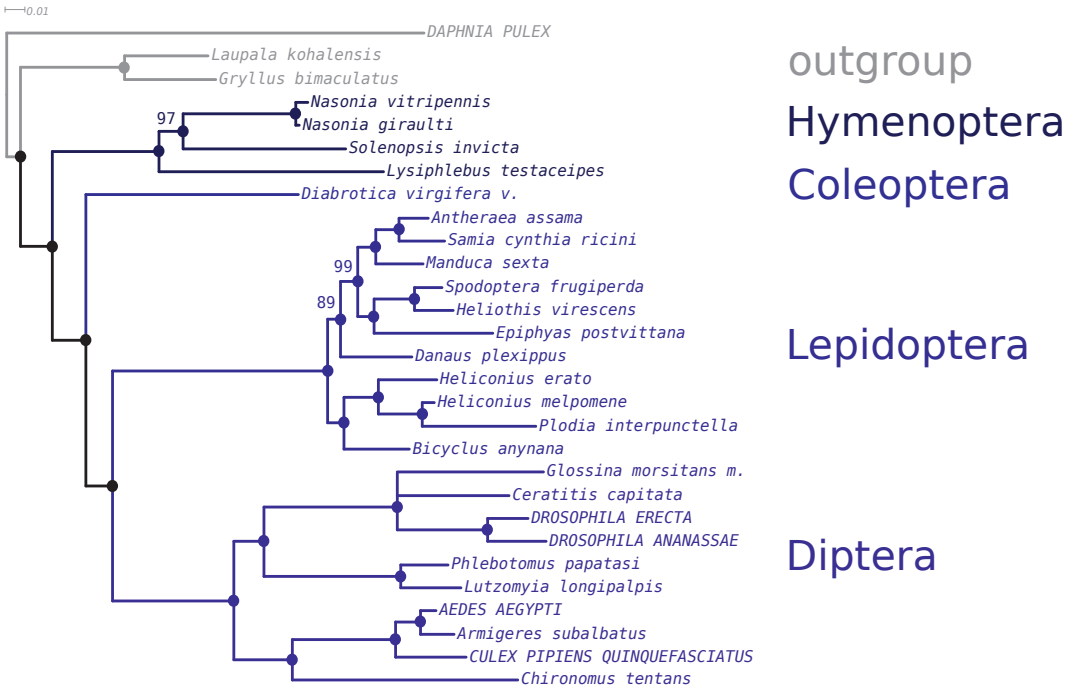

Figure 2.27.: Phylogram (majority rule) of 29-taxon ML analysis (112 genes, 1,000 bootstrap replicates) based on the SOS of `En_oP_Da`, including proteome species *Daphnia* where the proteome is available (Tab. 2.5). Outgroup species: *Daphnia* and ensiferans (Orthoptera). The tree was rooted with *Daphnia*. Color code: outgroups: gray; hymenopterans: dark blue; coleopterans: navy blue; lepidopterans: royal blue; dipterans: mid-blue.

Reanalysis of Dunn et al. (2008)

All ML trees inferred from selected optimal subsets based on Dunn's original data set (whatever option with MARE was applied), show a decreased resolution and decreased support values (cf. Figs. 2.28, 2.29 and A.10). However, total average information content and matrix saturation are increased when MARE is applied (see section 2.3.1 and Fig. 2.15). Especially in deep metazoan splits, there is no resolution at all. Several arthropod clades, e.g. Euarthropoda, Pancrustacea and a clade (Branchiopoda,Hexapoda) show high support in the tree inferred from the in original data set (Fig. 2.28). In contrast, both SOS topologies propose Onychophora to be positioned within euarthropods, but with weak support (BS $< 70\%$). Pancrustacea show only low support (BS $< 80\%$) as well. Pycnogonida are inferred as a sister group to a clade (Onychophora,Myriapoda,Euchelicerata) while internal relationships are unresolved. Branchiopoda show as a sister group relationship with hexapods, moderately supported (BS $< 90\%$).

For Dunn's data set, selecting an optimal subset (SOS) with increased total average information content and increased matrix saturation did not improve tree robustness. For possible reasons, the very low information content of genes and a similar Gaussian distribution of present|absent data should be considered.

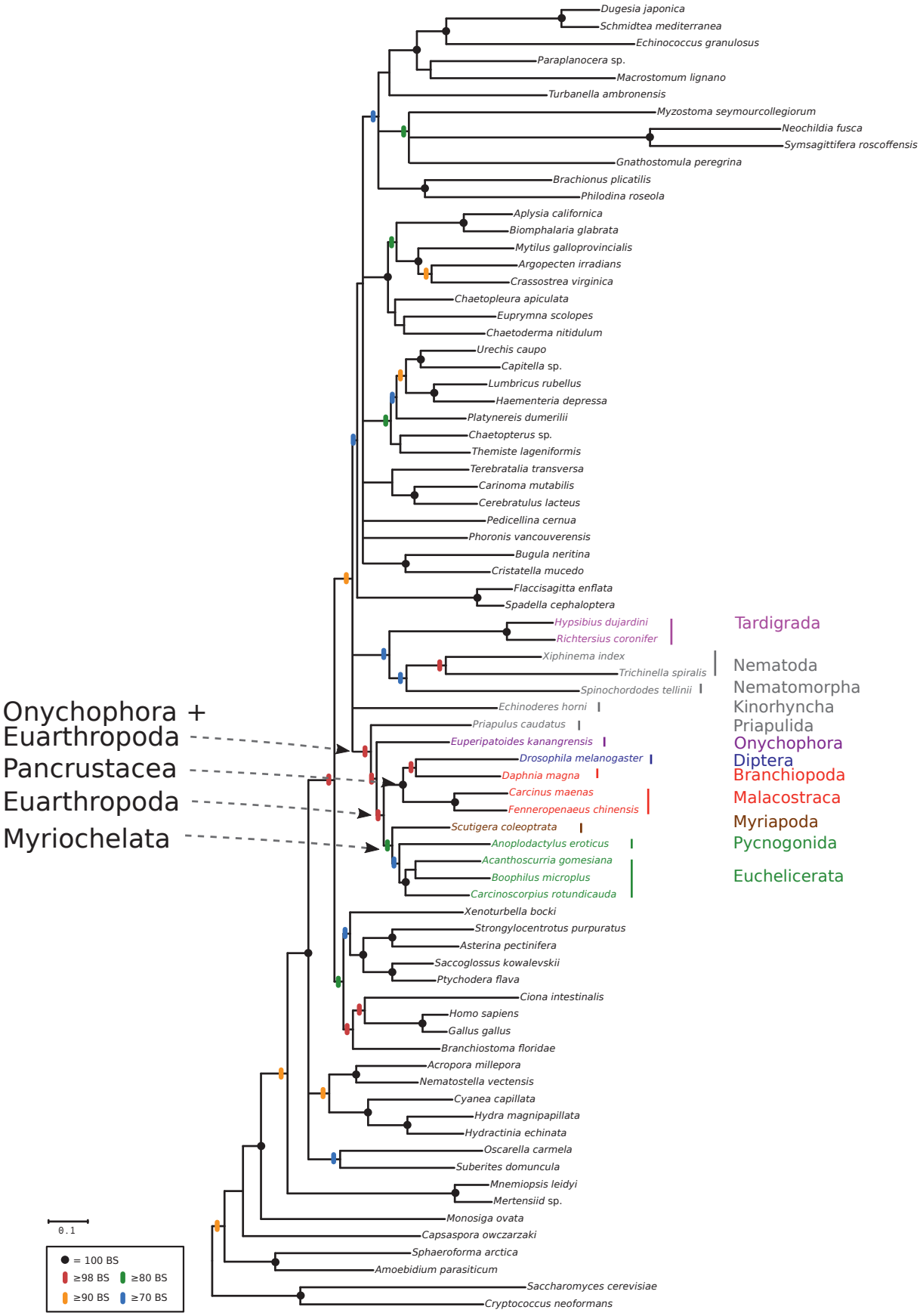

Figure 2.28.: Phylogram (ML, majority rule) based on the original data set of Dunn et al. (2008) with 77 taxa, 150 genes, see section 2.3.1 and Fig. 2.15. Bootstrap values were derived from 1,000 bootstrap replicates. The was tree rooted with *Saccharomyces* and *Cryptococcus*. Color code: Priapulida, Kinorrhyncha, Nematoda, Nematomorpha: gray; Tardigrada: pink; Onychophora: violet; chelicerates: green; Myriapoda: brown, crustaceans: red; Hexapoda: dark blue.

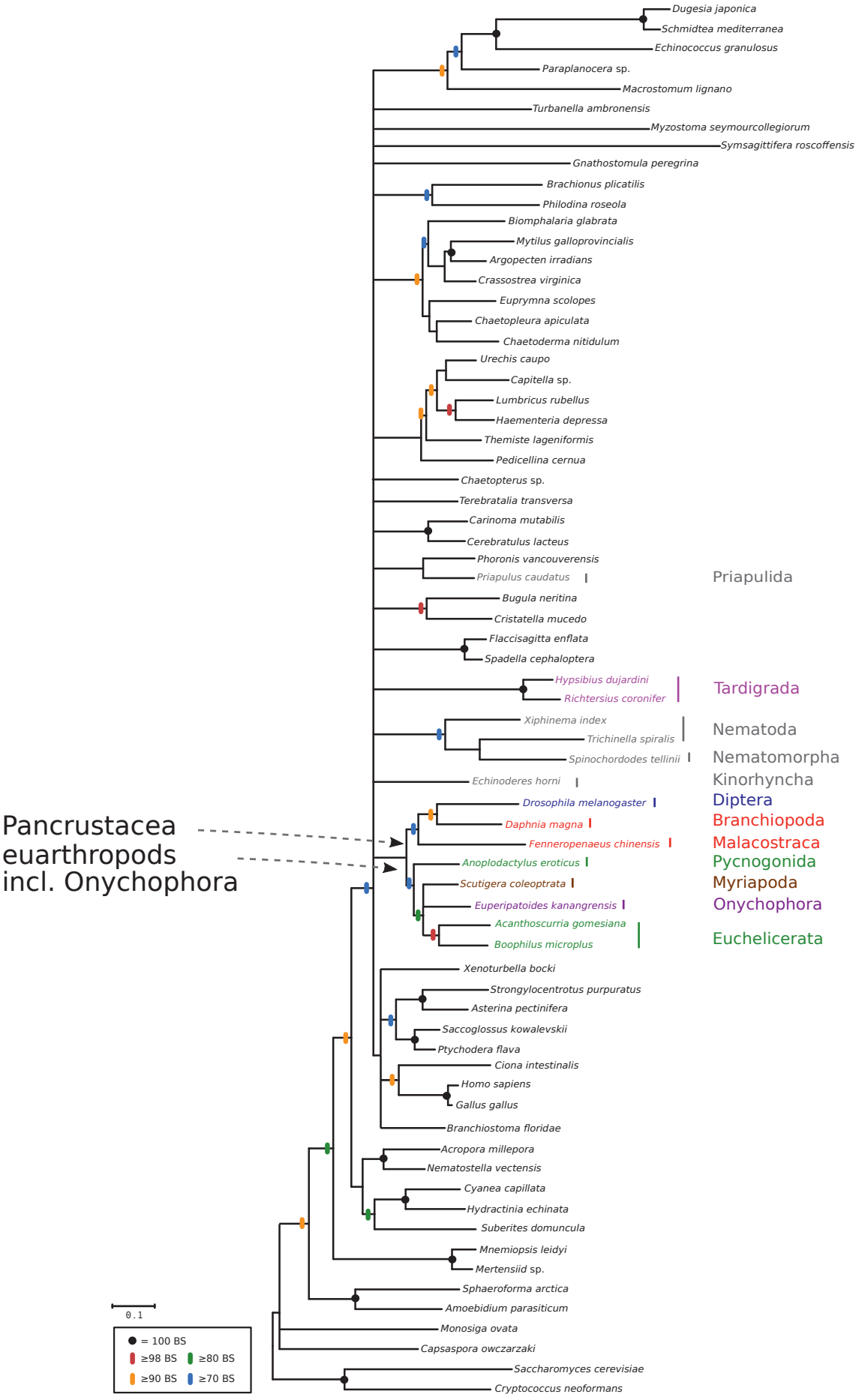

Figure 2.29.: Phylogram (majority rule) of a 70-taxon ML analysis (29 genes, 1,000 bootstrap replicates) based on the SOS (-t) of the provided data of Dunn et al. (2008). For labeling and color code, see Fig. 2.28.

2.4. Discussion

This is the first study with an extensive arthropod taxon sampling and a large-scale gene coverage. Gappiness in data matrices, cautious taxa and gene selection and gene overlap between taxa have extensively addressed. Phylogenomic data sets with an large euarthropod taxon sampling are mostly focused on particular genes (Timmermans et al., 2008; Aleshin et al., 2009). Studies including a broad gene coverage comprise only small euarthropod taxon samplings and are focused on Ecdysozoa or Metazoa (Roeding et al., 2007; Dunn et al., 2008; Philippe et al., 2009).

Present analyses illustrate upcoming demands: the interaction of parameters, for example taxa/gene overlap, matrix saturation, data distribution, information content of genes, heterogeneity of signal among genes and its distribution, and their effects on tree reconstruction is not clearly understood. Following questions are introduced and should be examined in future research: a) Can genes with low or no signal lead to resolved trees with high support and could this be caused by randomness? b) Is contradicting signal present for particular splits and how does this affect node resolution? c) Which method can reliably estimate influences of single genes/gene groups or single taxa/taxon groups prior to tree reconstruction? d) Are mixtures of fast evolving genes to resolve terminal nodes and slow evolving genes to resolve deep splits appropriate within a supermatrix approach? e) How can long branch taxa be identified prior to tree reconstruction and how can they be properly modeled?

The application of MARE outlines possible problems within large-scale data sets. Observed effects on taxa and gene selection and tree reconstruction due to different settings and data sets allow statements about what we have understood yet and what we still do not understand. This provides hints in which direction refinements or development of new methods could lead (e.g. influences of single taxa, interaction between gene heterogeneity, distribution of data and matrix saturation).

A drawback of present analyses are computational limitations: large scale data demand new reconstructions tools that require less computational power and can be preferably executed on desktop PCs in a reasonable time. We should be aware that the amount of data will increase 10–20 times with 454 sequencing. New bioinformatical tools are necessary to properly handle these data in processing, quality assignment and tree reconstruction.

2.4.1. Methodological aspects

Sequence processing and orthology assignment

Due to the complex nature of genome evolution (gene loss, duplications, functional diversification), a cautious assignment of gene orthology is indispensable when using ESTs for phylogenetic analyses (Hughes et al., 2006). HaMStR (Ebersberger et al., 2009) implements the database InParanoid (Remm et al., 2001; Berglund et al., 2008), similar to the procedure introduced by Chen et al. (2007). The use of trained profile Hidden Markov Models with a reciprocal BLAST for orthology assignment seems more conservative than other, more simple reciprocal BLAST procedures. Therefore, the number of common genes compared to studies that used different orthology procedures was evaluated. Data of the present study share most orthologous genes (`AP_1`: 51 genes, `AP_1`-SOS: 46 genes) with Philippe et al. (2009). Marginally smaller is the number of shared genes with Baurain et al. (2007) and Delsuc et al. (2008). The number of genes used in the selected optimal subset (SOS) of data set `AP_1` is similar to all three studies. This is remarkable, because the present orthology assignment is quite different. In contrast, the number of shared genes is much smaller in comparison to the study of

Dunn et al. (2008) (AP_1: 37 genes, its SOS: 19 genes). Dunn et al. (2008) developed a new strategy for orthology prediction. Selected genes by Dunn et al. (2008) have been evaluated with MARE as less informative (see Fig. 2.15). An orthology assignment benchmark test would be worthwhile to evaluate strengths and problems of current popular orthology prediction methods (e.g. Ebersberger et al., 2009; Schreiber et al., 2009).

Alignment masking

Alignment masking improves the signal-to-noise ratio within data sets and leads to more plausible trees. Recently, this has been demonstrated for simulated and real data Misof and Misof (2009) and by Kück et al. (2010). In the present analyses, alignment masking removed noise based on random similarity in all arthropod and endopterygote data sets. This is visible in split decomposition networks comparing the unmasked and masked SOS alignment of data set AP_1 (Figs. 2.16, 2.17). The masked data set shows a denser pattern that separates, for example, pterygote insects from remaining euarthropods. By the process of alignment masking, clear patterns are emphasized and conflicts in the data set are highlighted which are not caused by random alignment similarities. The positive effect of alignment masking is also obvious concerning information content (Fig. 2.7 and 2.10). The masked SOS shows twice as much total average information content than the unmasked SOS.

MARE: Reduction heuristics for taxa gene selection

MARE can be used to evaluate the information content of data matrices, of single genes and taxa within those matrices as well as the matrix saturation of a given data set. On this basis, MARE can be used to select an optimal data subset due to high total average information content using new reduction heuristics (section 2.1.4).

In this thesis, applying MARE was highly appropriate for arthropod and endopterygote data sets with power-law distributed data. Trees derived from unreduced data sets show misplaced taxa and many unresolved relationships, especially within hexapods (Figs. 2.18, 2.19 and 2.26). In contrast, trees from selected optimal subsets (SOS) are much more resolved, and suspicious phenomenons like polyphyletic Lepidoptera or Coleoptera disappear. The selection of an optimal data subset with MARE is a promising strategy for unreduced, power-law distributed data sets with low matrix saturation and low total average information content. Comparing reduction heuristics of MARE with currently used 'threshold selection' of genes and taxa from a raw matrix (see e.g. Dunn et al., 2008) would be useful. Another idea is a comparison of reducing data matrices based on presence|absence information only with subsequent tree reconstruction to evaluate a possible influence of information content. A reduction based on presence|absence information should lead to much larger data matrices with a lower total average information content. Meyer (2009) suggests from simulation studies that reduction relying on information content of power-law distributed data matrices provides equal or more accurate trees than unreduced data matrices with sparse data availability. It is unclear, in which way the distribution of data availability and information content affect each other during tree reconstruction. Wiens and Moen (2008) argue that missing data will not be a problem if there is enough information for available data. It remains unclear what "enough information" means. This might be highly dependent on the data set and the taxonomic group of interest. Thus, there is a risk that chosen criteria are biased by preferred topological hypotheses. The strategy to select taxa and genes with MARE is convenient, comprehensible and has been successful for present data sets. MARE implies many possibilities to exploit provided information. Using MARE may impede systematic errors caused by accumulated data

that provide no information. A dominance of uninformativeness and noise in a data set entails the risk of a split resolution by chance with erroneously high node support values (see also Lemmon et al., 2009).

Extended, weighted geometry quartet mapping
Compared to likelihood quartet mapping (LM), extended geometry quartet mapping (eGM) is a conservative approach. LM can overestimate signal, and it was significantly shown that apparently resolved quartet trees actually are star-like (Nieselt-Struwe, 1997; Nieselt-Struwe and von Haeseler, 2001). On the other hand, eGM might be too conservative to detect low signal (Nieselt-Struwe and von Haeseler, 2001). Other strategies for measuring information content are woth to take into account, for example invariant techniques (Allman and Rhodes, 2007; Casanellas and Fernández-Sánchez, 2007).

Information content of genes and data matrices
High scores of information content. In MARE, a high score of information content in a single gene (partition) is deried from the number of (partially) resolved quartets obtained via eGM. Although the score is high, single quartets can support different topologies. This may be problematic in tree reconstruction. Furthermore, genes that have a high score information content, but for example code for proteins involved in different pathways, do not necessarily support an identical or similar topology. Thus, a high score of total average information content of the matrix does not necessarily lead to completely resolved trees. Nevertheless, as first crucial step MARE currently separates signal from non-signal (see Fig. 2.2). MARE could be easily extended in a way that single quartets in genes and in the data matrix can be analyzed. This could be a first step to understand possible interaction of genes. Another possibility is the refinement to restrict the choice of taxa for every quartet to predefined groups of interest (hypothesis-driven quartet mapping).
Detected signal and genuine phylogenetic signal. High relative information content is not necessarily congruent to phylogenetic signal. Since data availability and signal might not be correlated, if many genes are uninformative, focusing on data presence will not fully use the potential of supermatrix preprocessing. Information content of genes in a phylogenomic scale has not been considered before with respect to taxa and gene selection.

Reduction heuristics algorithm
The optimality function currently implemented in MARE uses a hill climbing algorithm (section 2.1.4). In the pre-alpha-version, the overlap of genes between taxa is guaranteed for two genes. According to Steel and Sanderson (2010), gene overlap should be raised to three genes. An minimum overlap of nucleotide positions may be worth to consider. Another idea is to define a core set of taxa of interest and to search for maximal (quasi-)bicliques (Yan et al., 2005; Li et al., 2008b). Afterwards, the taxon set could be expanded in the same way. Thereby, information content of single genes should be taken into account towards a "weighted" quasi-biclique approach. This might generate a more balanced submatrix, if the data set is extremely power-law distributed. This is the case, if e.g. data matrices include many proteome species (see e.g. Driskell et al., 2004).

Distribution of data and of heterogeneous information content of genes
The data set of Dunn et al. (2008) was evaluated with MARE. SOS were selected with different settings and subsequently tree reconstruction was conducted to examine performance of MARE with respect to non power-law distributed data. The data set of Dunn shows low signal for most genes and

low heterogeneity among them (Fig. 2.15). The distribution of available data is Gaussian-like with ca. 50% matrix saturation. We have to keep in mind that provided data were already preselected by the authors. In present reanalysis, this preselected data matrix has been treated as unreduced because the original data matrix was not available and a request to the authors remained unanswered. The unreduced data matrix apparently is sufficient to reconstruct a resolved tree. Nevertheless, the resolution and node support is astonishing in view of to the low total average information content and low information content of single genes. In both SOS, total average information content was increased twice (section 2.3.1), but a selection of subsets was not 'successful' with respect to tree reconstruction. Any removal of taxa and genes highly affected the data set towards lower tree resolution. Anyway, total average information content of both SOS (~ 0.4) should be sufficient for reconstructing a robust tree. Reasons for unresolved topologies could be: i) Reduction of Dunns data set is not useful because data are nearly Gaussian distributed and the total average information content and heterogeneity of signal among genes is low (Fig. 2.15). Reduction has an negative effect on the tree resolution. ii) There is not much signal in Dunns data set and nodes show particularly erroneously high support caused by noise. Dunns data set might be highly unstable according to different tree reconstruction methods. A possible way to evaluate reasons is an optional implementation of hypothesis-based quartet mapping in MARE and studying single quartets for particular clades.

In summary, MARE is appropriate to evaluate signal of genes and heterogeneity of signal among genes for respective taxa. MARE exposes problems within data sets, and the implemented reduction heuristics is appropriate for data matrices with power-law distributed data and (high) signal heterogeneity among genes which corresponds to most real world data sets. Some issues of MARE should be expanded and refined.

Unbalanced taxon sampling and possible impact on deep splits

One aim of the present study was to include all available euarthropod EST data to avoid long branch attraction (Philippe et al., 2005c; Baurain et al., 2007; Brinkmann and Philippe, 2008). Li et al. (2008a) pointed out that more taxa are not necessarily better for reconstruction of ancestral character states (see also Brinkmann and Philippe, 2008; Heath et al., 2008a,b). Therefore, a cautious evaluation of putative signal is crucial. A further step is the examination of provided signal addressing e.g. randomness or homoplasies.

The arthropod data sets of this study predominantly include hexapods and crustaceans. Therefore, myriapods are under-sampled, but unfortunately, no additional myriapod EST data are available. This might influence the selection of an optimal subset, if the selection is shifted towards genes that resolve mainly pancrustacean relationships. If more pancrustacean data are available, the number of possible (partially) resolved quartets is higher than e.g. for myriapods. Then, the probability to detect more signal for a pancrustacean clade might be higher. Whether a taxon sampling with emphasis on pancrustacean data indeed influences tree reconstruction could be easily tested by an hypothesized extended geometry quartet mapping (eGM).

Bayesian analyses of the selected optimal arthropod subset

Bayesian analyses of the SOS of `AP_1` showed topological differences and support for selected clades in trees derived from single PhyloBayes chains (section 2.3.3, Fig. 2.22 and Tab. 2.8). A check

of convergence between chains with the *maxdiff* tool is recommended by the authors of PhyloBayes (Lartillot et al., 2008). A chain which does not converge with other chains may be trapped within a local optimum. In this case, it remains unclear in which way log-likelihood values shall be assessed. The program lacks explanations for handling chain convergence with respect to log-likelihood values. The present data set shows no convergence when considering all chains (maxdiff = 1, Tab. 2.8). The authors of PhyloBeayes recommend at least a *maxdiff*-value of <0.3 (Lartillot et al., 2008). Chains with the best log-likelihood values which concurrently converged (*maxdiff* < 0.3) were chosen for the majority rule consensus (mrc) tree. For the present study, this is proposed as reasonable compromise.

Taxa with various placements according to different chains (e.g. the bristletail *Lepismachilis* of the current data set) should be additionally analyzed with other Bayesian reconstruction methods. Kolaczkowski and Thornton (2009) showed that Bayesian inference (MrBayes) is biased "in favor of topologies that group long branches together, even when the true model and prior distributions of evolutionary parameters over a group of phylogenies are known". This should also be tested for PhyloBayes considering the CAT model which was proposed to handle taxa with long branches properly (Lartillot and Philippe, 2004; Lartillot et al., 2007). Marshall (2010) and Brown et al. (2010) addressed possible incorrect branch length estimates in Bayesian analyses which should be also tested for PhyloBayes.

Impact of single taxa and proteome taxa on subset selection and tree reconstruction

Effects of including a single taxon

Analyses of this study clearly show that single taxa can have a massive effect on tree reconstruction. The inclusion of the velvet worm *Epiperipatus* led to very little differences in taxa and gene selection (cf. SOS of `AP_1` without *Epiperipatus* and SOS of `AP_2`, including *Epiperipatus*, section 2.3.1). The tree inferred from the SOS of `AP_2` including *Epiperipatus* shows few, but very remarkable differences. For example, Mandibulata are inferred with myriapods as a sister group to Pancrustacea, although marginally supported (Fig. 2.24). The placement of myriapods seems highly sensitive and may be easily influenced by single taxa, single genes and/or reconstruction methods. The inclusion of *Epiperipatus* does not 'improve' the tree towards tree resolution, but it shows effects on apparently sensitive nodes that are worth to investigate.

Exclusion of proteome species

Proteome taxa have a strong influence on taxa/gene selection (see section 2.3.1). An exclusion of proteome taxa leads to more taxa in the SOS for arthropod (`AP_3_oP`) and endopterygote data sets (Figs. 2.12,2.13,2.14). This is favorable, if data matrices include species of interest, e.g. apterygote hexapods, which only partially cover genes of proteome species. Therefore, proteome species conceivably shift selection of a data subset with MARE. Optionally weighting taxa in MARE can compensate such shifts. Currently, all available proteome species serve as model organisms and show highly derived morphological characters. Artifacts caused by these species are hard to predict and identify. Several studies and unpublished data state that some proteome species, e.g. the louse *Pediculus*, show various placements depending on the reconstruction method (e.g. Odronitz et al., 2009, and Ebersberger & Strauss, pers. comm). In the present study, this was also observed for *Pediculus*. Because of huge data availability, it is thinkable that proteome taxa intensely shift tree reconstruction. Considering support values, only few splits show enormous differences between trees with and without proteome species. In the SOS on `AP_3_oP`, the support for Euarthropoda is astonishingly low (BS 67%, see

Fig. 2.25, Tab. 2.11). Also, the decrease of support for a sister group relationship of Branchiopoda and Hexapoda is considerable. The bootstrap support falls down from 92% to 51% (Fig. 2.25). It is conspicuous that remaining branchiopods in this SOS show instability after tree reconstruction (see section 2.3.3, SOS `AP_3_oP`). An analysis of selected 125 genes should be conducted with previously excluded proteome species. A much higher support for a clade (Branchiopoda,Hexapoda) would imply a major influence of proteome taxa. Otherwise, the difference in support of (Branchiopoda,Hexapoda) might have arisen from particular different gene selection between the SOS of `AP_3_oP` and `AP_1`.

Unstable taxa: appropriateness of leaf stability indices

Leaf stability indices (LSIs) should be considered critically. The post-processing technique is implemented in Phyutility (Smith and Dunn, 2008) based on Thorley and Wilkinson (1999) and Thorley and Page (2000), who used LSIs among a set of collected trees to identify unstable taxa.

Unstable taxa are not necessarily misplaced. Taxa that are putatively wrongly placed in all collected trees are not identified and may support artificial clades. Dunn et al. (2008) suggest to prune unstable taxa from the alignment and repeat analyses. The idea is to obtain a stable and more accurate "backbone" tree. In the present analyses, pruning of unstable taxa (namely Pycnogonida, Myriapoda and *Baetis*) with subsequent reanalyses resulted in a tree with Onychophora as sister group maximally supported to ((Nematoda,Tardigrada)Euarthropoda) (section 2.3.3, and Fig. A.4). However, a well resolved clade Onychophora + ((Nematoda,Tardigrada)Euarthropoda) is suspicious. No molecular, morphological or developmental study has ever proposed this clade (Roeding et al., 2007; Dunn et al., 2008; Zantke et al., 2008; Edgecombe, 2009; Budd and Telford, 2009). By discarding Pycnogonida and Myriapoda, possible synapomorphies with onychophorans are dropped. Consequently, onychophorans slip down the tree and result in a biased topology with a falsely maximal support (see Wägele and Mayer, 2007).

Instead of pruning unstable taxa, it is necessary to model these taxa correctly and avoid artificial clades. An idea is to examine possible signal per gene for these taxa. Therefore, quartet mapping from predefined groups might be appropriate.

2.4.2. Phylogenetic relationships

Inferring the sister group of hexapods: still, a puzzling problem

The knowledge of the sister group of hexapods is crucial for the polarization of characters and impotant if we want to resolve the hexapod (and arthropod) tree of life. Previous molecular and morphological studies have suggested various taxa as sister group of hexapods, e.g. Copepoda, Branchiopoda, Remipedia, Malacostraca (crustaceans) or Myriapoda (see Szucsich and Pass, 2008; von Reumont, 2010). In all trees inferred in this study, Branchiopoda are sister group to hexapods, but with different support. In both ML and Bayesian SOS trees of data set `AP_1`, the sister group relationship of Branchiopoda and Hexapoda is strongly supported (Tab. 2.6). This corroborates single- and multi-gene analyses (Regier and Shultz, 1997; Shultz and Regier, 2000; Regier et al., 2005; Roeding et al., 2007; Dunn et al., 2008; Philippe et al., 2009). However, the support derived from exclusively ribosomal protein coding genes is negligible (BS 54%, Tab. 2.10). Non-ribosomal protein coding genes cover ca. 3/4 of the SOS matrix (97 of 129 genes) and might dominate tree reconstruction of the complete SOS: support values are similar (BS 92% for the complete SOS, BS 94% for the non-ribosomal SOS

tree, cf. Tab. 2.10). Studies based on nuclear rRNA markers suggest Copepoda as sister group of Hexapoda, albeit weakly supported (Mallatt and Giribet, 2006; von Reumont et al., 2009; Mallatt et al., 2010, see chapter 3). Other studies based on non-ribosomal genes provide contradicting trees. While results of Regier et al. (2008) are inconclusive, Regier et al. (2010) infer a clade "Xenocarida" (= Remipedia + Cephalocarida) as a sister group to Hexapoda: support is strong at a nucleotide level, but insignificant at an amino acid level. Ertas et al. (2009) suggest a close relationship of Remipedia and Hexapoda based on hemocyanin.

By excluding proteome species (AP_3_oP), the low support for a clade (Branchiopoda,Hexapoda) is unexpected. From branchiopods only *Daphnia pulex* was excluded (see SOS of AP_3_oP, Fig. 2.25). Gene coverage of remaining branchiopods resembles the SOS of AP_1: *Artemia* covers 93, *Daphnia magna* 66 and *Triops* 42 genes. Thus, an insufficient gene coverage of branchiopods seems unlikely. It should be examined why *Daphnia pulex* apparently provides a more substantial support for Branchiopoda + Hexapoda than remaining branchiopods.

Non-molecular studies provide ambiguous results. A clade (Branchiopoda,Hexapoda) is suggested by Schram and Koenemann (2004) based on a "preliminary" cladistic analysis. Glenner et al. (2006) suggest a scenario for a common ancestor of hexapods and fresh water crustaceans like branchiopods, but they have mainly derived characters and are highly adapted to seasonal freshwater ponds. In contrast, this sister group relationship is contradicted by paleontological data (Waloszek, 2003). Recent studies also reject a sister group relationship of Branchiopoda and Hexapoda, but propose Malacostraca or Remipedia as possible sister group to hexapods (Fanenbruck et al., 2004; Fanenbruck and Harzsch, 2005; Harzsch, 2006; Strausfeld, 2009; Strausfeld et al., 2009; von Reumont, 2010). The incongruence within and between molecular and morphological studies addressing the hexapods sister group is not resolved yet. Careful analyses of signal quality in molecular and morphological data are still required, along with more molecular data from Leptostraca, Remipedia and Cephalocarida.

Hexapoda are unambiguously supported from EST data

All selected optimal subset (SOS) trees of this thesis unambiguously resolve monophyletic Hexapoda. Thereby, primary wingless entognathous orders are proposed as most basal splits within hexapods. Moreover, advanced methods applied here, payed off their effort (section 2.4.1). Present results strongly reject studies based on mitochondrial (mt) data (Nardi et al., 2003a,b; Carapelli et al., 2005, 2007). Mt data suggest a polyphyly of hexapods, which implies that features of the hexapod bauplan must have evolved at least twice. Likewise, Aleshin et al. (2009) outlines that mitochondrial data might be not appropriate for deep arthropod and hexapod phylogeny. Studies based on mt markers or on nuclear multi-gene data did not include proturans (Carapelli et al., 2007; Timmermans et al., 2008; Aleshin et al., 2009; Regier et al., 2010, see chapter 1, section 1.2). EST-analyses, presented in this thesis based on an extensive phylogenomic data set, including all orders of monocondyl, primary wingless hexapods yield strong support for monophyletic Hexapoda. Moreover, the resolution and support of Hexapoda seems independent from model selection, tree reconstruction method and various taxa/gene selections (see section 2.3.3 and 2.4.1). Supporting the traditional view, it can be concluded that hexapods are monophyletic and that the distinctive hexapod bauplan evolved only once, even though non-molecular characters may provide ambiguous support (Dohle, 2001; Bitsch and Bitsch, 2004; Harzsch et al., 2005; Harzsch, 2006; Ungerer and Scholtz, 2008; Szucsich and Pass, 2008, see chapter 1).

Entognatha and Nonoculata

Analyses of the present study provide ambiguous support for a monophyly of entognathous hexapods taking ML and Bayesian SOS-analyses of the full data set (AP_1) into account. Entognatha are recovered, albeit moderate or weakly supported (Tab. 2.6). Considering Bayesian analyses, one chain suggests a sister group relationship of Collembola and Ectognatha. Nonoculata as sister group to (Collembola,Ectognatha) showed low support (pP 0.5). Chains showing convergence with other chains, having much better log-likelihoods (Tab. 2.8) strongly support Hexapoda. In these chains Entognatha are recovered with a support at least > pP 0.5. Separate analyses of ribosomal and non-ribosomal protein coding genes also suggest Entognatha (non-ribosomal topology) or a sister group relationship of springtails and Ectognatha (ribosomal topology, see 2.3.3). The second scenario is negligibly supported (BS < 50). The position of Collembola is not resolved and the average information content (IC) of genes is remarkably lower than the IC of non-ribosomal protein coding genes. A paraphyly of collembolan Poduromorpha, suggested by the ribosomal tree, gains no support from morphological and molecular studies (Deharveng, 2004; Gao et al., 2008; Xiong et al., 2008). Nevertheless, present ML analyses with more than 100 genes with an additional included onychophoran (SOS of AP_2) or excluding proteome taxa (SOS of AP_3_oP) strongly corroborate Entognatha with increased support. In summary, a support for Entognatha seems sensitive to different tree reconstruction methods (ML *versus* Bayesian). Its inference may also be dependent on particular signal coming from different genes. This confirms previous molecular and morphological studies that also point out these ambiguities (Szucsich and Pass, 2008; Grimaldi, 2010). Within morphological studies, the interpretation of character states remains difficult (Szucsich and Pass, 2008), due to e.g. extreme adaptations to subterranean or cryptic habitats. At least, from a molecular perspective, there is currently no alternative hypothesis which gains more support than a clade Entognatha.

Within Entognatha, there is a strong support for Nonoculata in all analyses of the present study without exception. This corroborates results of several single, mostly rRNA gene analyses (Luan et al., 2003, 2004; Kjer, 2004; Mallatt and Giribet, 2006; Dell'Ampio et al., 2009; von Reumont et al., 2009; Mallatt et al., 2010). Ribosomal genes had been characterized as potentially biased (e.g. Luan et al., 2005; Dell'Ampio et al., 2009), based on findings of nuclear rRNA genes. von Reumont et al. (2009) obliterate, at least for rRNA genes, previously proposed arguments by including methods that account for non-stationary processes (see chapter 3). Furthermore, not only separate analyses of ribosomal protein coding genes, but also analyses of non-ribosomal protein coding genes strongly support Nonoculata (Tab. 2.10). Morphological evidence for this clade is still ambiguous (Szucsich and Pass, 2008). The amount of shared genes for entognathous hexapod orders of all data sets in this study is the highest currently published. Since this number is still moderate (around 30 genes are shared), 454-EST approaches of apterygotes, especially proturans and diplurans, and missing potential sister taxa of hexapods (e.g. Remipedia) may be a promising approach for future studies, to cover a much higher number of genes shared by apterygote hexapods.

Ectognatha

The monophyly of Ectognatha seems well founded by single- gene analyses (e.g. Kjer, 2004; Kjer et al., 2006; Misof et al., 2007; von Reumont et al., 2009). Recently, this has been corroborated by nuclear protein coding genes (Regier et al., 2010). Again, Ectognatha is strongly corroborated by all phylogenomic analyses of this study (see Figs. 2.20-2.21). The placement of *Lepismachilis* (Archaeognatha) as sister group to remaining pterygote insects appears robust (cf. ML majority rule tree and Bayesian mrc tree, Figs. 2.20, 2.21). However, the placement of *Lepismachilis* varies between single Bayesian trees and therefore might be affected by unknown parameters within single chains. An alternative scenario, where *Lepismachilis* is closely related to the louse *Pediculus* (suggested by Bayesian trees from single chains with less well log-likelihood values) is likewise proposed by the non-ribosomal ML tree. Although in theory ML and Bayesian approaches should accomplish imbalances of a different number of shared genes if gene overlap is present, it is remarkable that *Lepismachilis* shares much more genes with *Pediculus* than with Odonata or Ephemeroptera. Tests for a potential influence on a taxon placement with respect to different amounts of shared genes are pending. However, in both trees the poor resolution does not allow any comprehensible conclusion. If we consider that an analysis of both ribosomal and non-ribosomal protein coding genes provides more signal than a separate approach, and if inconsistencies between single Bayesian trees are due to local optima (Fig. 2.22), current analyses strongly corroborate Ectognatha with Archaeognatha as sister group of Dicondylia. Future analyses, however, should include Zygentoma which currently lack any qualitatively acceptable EST data.

Ambiguous placement of myriapods

Addressing mandibulate taxa, two alternative clades, either Atelocerata (= Tracheata) or Pancrustacea (= Tetraconata) have been suggested (chapter 1). All analyses of the current study that base on selected optimal subsets (SOS), unambiguously suggest a clade Pancrustacea with maximal support, no matter whether with or without proteome species and whether analyzing ribosomal protein and non-ribosomal protein coding genes together or separately. Crustaceans are always paraphyletic, and Atelocerata are never recovered. This corroborates previous molecular analyses, based on single and on multi-gene approaches inferred with various data sets and various methods of alignment, alignment masking, taxa/gene selection, and tree reconstructions (e.g. Dunn et al., 2008; von Reumont et al., 2009; Mallatt et al., 2010; Regier et al., 2010). Therefore, it is concluded that crustaceans are more closely related with hexapods than with myriapods, at least based on current molecular analyses. The current phylogenomic approach contradicts morphological character sets supporting Atelocerata, described e.g. by Bäcker et al. (2008).

There is little morphological support for a clade Myriochelata (Mayer and Whintington, 2009), in contrast to data supporting Mandibulata (e.g. Wägele, 1993; Harzsch, 2006; Bäcker et al., 2008). In present analyses, the position of myriapods remains ambiguous. Current ML analyses could either not resolve the position of myriapods, placed them as a sister group to chelicerates (Myriochelata), within chelicerates, or as sister group to pancrustaceans. All placements are negligibly supported (cf. section 2.3, Figs. 2.20, 2.21, 2.24, 2.25). Most unexpected is the polyphyly of myriapods due to the separation of chilopods and diplopods in the ML tree derived from non-ribosomal protein coding genes. Chilopoda are placed within chelicerates, but remains unresolved; Diplopoda are sister group to Pancrustacea with poor support (see section 2.3, Fig. 2.23, separate analyses of ribosomal and non-ribosomal protein

coding genes). Methodologically this seems suspicious due to the very low number of shared genes (section 2.4.1). In the Bayesian mrc tree, myriapods are placed as sister group of chelicerates, again weakly supported. Summarized, the phylogenetic inference of myriapods is highly dependent on the selected data set and taxa/gene selection. Due to advanced methods (e.g. MARE) applied here, the outcome, namely an ambiguous phylogenetic placement of myriapods, should be more realistic than results of some other studies. With current methods, their position cannot be clearly addressed in the arthropod tree of life. This ambiguity is evident by comparing studies (Roeding et al., 2007; Dunn et al., 2008; Regier et al., 2010) which suggest a well supported resolution with different multi-gene approaches, but support conflicting hypotheses, namely Mandibulata *versus* Myriapoda. Thereby, other studies have demonstrated a high sensitivity of reconstructing Myriochelata with respect to gene choice, taxon sampling and outgroup selection (Bourlat et al., 2008; Rota-Stabelli and Telford, 2008; Philippe et al., 2009). In molecular tree reconstruction, myriapods may be "rogue" taxa whose position in the tree remains ambiguous and fall into different branches across the tree in e.g. bootstrap replicates and thereby significantly reduce the resolution in the resulting consensus trees or the support on the best-scoring (ML) tree (Stamatakis, pers. comm., Thomson and Shaffer, 2010). The unstable position is probably related to a systematic phenomenon of myriapod molecular evolution (see also chapter 3). To resolve the myriapod position in the arthropod tree of life, there is a high demand to better understand heterogeneity of substitution processes among arthropods. Equally important is the inclusion of good quality data of all myriapod groups in phylogenomic analyses. Currently, data from the single chilopod representative *Scutigera* lack good quality and gene coverage. Further, there are no EST data from two more important ancient lineages, Symphyla and Pauropoda. Additionally, no EST data are available from remipedes (crustaceans) which are discussed as possible common ancestor of Atelocerata (Bäcker et al., 2008). With respect to these challenges, we may be able to resolve the myriapod position and thus the Mandibulata - Myriochelata and the Pancrustacea - Atelocerata conflict in further studies.

Placement of water bears and velvet worms

Tardigrades (water bears) and onychophorans (velvet worms) display a mosaic of plesiomorphic and autapomorphic features of segmental differentiation. The position of tardigrades and onychophorans is crucial to resolve the evolution of e.g. segmentation, appendages, and the central nervous system (Edgecombe, 2009; Budd and Telford, 2009).

Tardigrades, represented by two species in the present EST analyses, are monophyletic and consistently reconstructed as a sister group to Nematoda with maximal support in nearly all trees (section 2.3.3). This is in line with recent molecular studies based on various gene selections (Baurain et al., 2007; Roeding et al., 2007; Lartillot and Philippe, 2008; Bleidorn et al., 2009). This sister group relationship suggests that tardigrades belong rather to Cycloneuralia and allies (Aguinaldo et al., 1997) than to arthropods including onychophorans. In general, molecular, morphological and neuroanatomical studies consider the position of tardigrades as ambiguous. For further discussion refer to Giribet (2003); Mallatt et al. (2004); Dunn et al. (2008); Zantke et al. (2008); Budd and Telford (2009); Edgecombe (2009) and Meusemann et al. (2010)[1].

[1] Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, 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). Advanced Access, doi:10.1093/molbev/msq130.

The position of onychophorans is resolved in the majority of the current EST trees (Figs. 2.20, 2.21). They show strong support for a clade (Onychophora,Euarthropoda). This agrees with morphological and molecular studies (Hou and Bergström, 1995; Roeding et al., 2007; Dunn et al., 2008; Edgecombe, 2009). Morphologically, velvet worms resemble euarthropod-like animals with a reduction of locomotory cilia, a body cavity with a pericardial septum, a heart with ostia, etc. In contrast, they lack for example a complete disintegration of the muscular tube into segmentally arranged muscle systems, segmentally arranged sclerotized exoskeletal structures, see discussion in Meusemann et al. (2010). Earlier morphological and molecular analyses place onychophorans either as a sister group to (Tardigrada,Euarthropoda) (Budd and Telford, 2009) or to Euarthropoda (Roeding et al., 2007; Dunn et al., 2008; Edgecombe, 2009). A clade (Onychophora,Euarthropoda), corroborated by the present analyses, suggests that fully differentiated segmentation, including ganglionization of the central nervous system evolved in a common stem-lineage of onychophorans and euarthropods. This implies that onychophorans primarily lack many characters of the euarthropod body organization (Hou and Bergström, 1995; Edgecombe, 2009). The interpretation of the fossil record of "lobopodian"-grade organisms as possible stem group representatives of euarthropods is also compatible with this conclusion (Hou and Bergström, 1995). In the present thesis, a placement of velvet worms as sister group to ((Tardigrada,Nematoda)Euarthropoda) is considered as unlikely due to methodological bias (see section 2.4.1) which has been reconstructed in the present ML tree excluding pycnogonids and myriapods. Nevertheless, molecular phylogenomic data of tardigrades and velvet worms are still sparse. These taxa may highly influence early euarthropod (Mandibulata, Myriochelata) or deeper splits (e.g. possible Panarthropoda). Therefore, more data and careful analyses are crucial for a plausible reconstruction of the phylogenetic position of water bears and velvet worms.

Early Pterygota

We have no clear picture of the early evolution of winged insects (Whitfield and Kjer, 2008). Morphological and molecular analyses either support a clade (Odonata (Ephemeroptera + Neoptera)) coined Chiastomyaria (Boudreaux, 1979; Kjer, 2004), Metapterygota (Ephemeroptera (Odonata, Neoptera)), see Börner (1909) and Zhang et al. (2008), or Paleoptera ((Odonata,Ephemeroptera), Neoptera), see Hennig (1981) and Kukalová-Peck (1983). Most molecular analyses support either Chiastomyaria or Paleoptera (see discussion in Simon et al., 2009). The present phylogenomic data are inconclusive. The maximum likelihood SOS tree does not show one of three suggested scenarios but, Paleoptera are strongly supported in the Bayesian SOS tree (Figs. 2.20, 2.21). This contrasts with the study of Simon et al. (2009) which corroborates Chiastomyaria, but their data sets include a different gene selection and a much smaller taxon sampling on pterygote insects than analyses of this thesis. Metapterygota have been favored by Staniczek (2003) examining structures of the pterygote head capsule. This study substantially lacks data from Zygoptera and Anisozygoptera (Odonata): only one anisopteran species is included as representative of Odonata. Thorough morphological reanalyses of mandible muscle structures within dragon- and damselflies based on computer tomography (CT), again support Paleoptera (Blanke, unpublished data, pers. comm.).

Currently, EST data of Odonata and Ephemeroptera are only represented by one single species each (Simon et al., 2009). Still, it lacks data from dragonflies (Anisoptera) and Anisozygoptera, but will be crucial in future analyses. Data from both taxa may help to improve a robust tree reconstruction. Examination of early pterygote relationships should cautiously consider evolutionary rates and possible

artifacts of a rapid gene evolution (Simon et al., 2009). Analyses of fast and slow evolving genes, e.g. using SCaFoS (Roure et al., 2007) or SlowFaster (Kostka et al., 2008), might be worth its effort to reliably resolve early pterygote splits. Detailed analyses of genes responsible for the development of wings, wing-related muscles and physiological requirements that enables flight also may provide suitable information on an early evolution and relationships of Odonata and Ephemeroptera with respect to Neoptera. Therefore, data from larvae as well as adult specimens are important.

Endopterygote insects and the position of Hymenoptera

Within Neoptera, relationships among early endopterygote insects are a major focus of scientific activity. It is still unclear whether Coleoptera or Neuropteroidea branch off first, or whether hymenopterans are a sister group to all other endopterygote insects (Kristensen, 1999; Kukalová-Peck and Lawrence, 2004; Beutel and Pohl, 2006; Wiegmann et al., 2009). All arthropod and endopterygote SOS trees of the present thesis show strong support for the monophyly of included endopterygote orders. Hymenoptera are unambiguously resolved as a sister group to all other endopterygote insects (Figs. 2.20-2.25, and 2.27), with various taxa selection (SOS of `AP_2` and `AP_3_oP`), different gene groups (ribosomal proteins *versus* non-ribosomal proteins, Figs. A.5, A.6) or different gene selection (SOS of `AP_3_oP` and endopterygote SOS). This is in line with previous studies (Savard et al., 2006; Wiegmann et al., 2009; Simon et al., 2009), but contrasts with conclusions based on complete mitochondrial genomes (Castro and Dowton, 2005). Single nuclear rRNA analyses again corroborate this scenario (see chapter 3, mixed model trees and von Reumont et al., 2009). In nuclear rRNA studies, the placement of Hymenoptera seems dependent on selected model strategies (cf. Mallatt and Giribet, 2006; von Reumont et al., 2009; Letsch et al., 2010). In the study of von Reumont et al. (2009), mixed model trees show Hymenoptera as sister group to all remaining endopterygote insects. Applied standard DNA models on the same rRNA data set presented in chapter 3 cannot resolve the position of hymenopterans with respect to Coleoptera. In the study of Letsch et al. (2010, Figs. 7-11), the position of Hymenoptera is unstable according to different alignment strategies and / or modeling. In contrast, nuclear protein coding genes may provide a stronger signal in favor of Hymenoptera as sister group to remaining endopterygote insects. EST analyses of the present thesis corroborate Hymenoptera as sister group to all remaining endopterygotes (section 2.3). Resolution of the hymenopteran position is important in interpreting and understanding early extinct endopterygote insects and the evolution of this most species-rich group of arthropods. Still, the inclusion of Strepsiptera, Neuropterida and Mecoptera remains crucial to resolve the position of Hymenoptera as accurately as possible. Molecular data are still sparse and available EST data are of low quality (GenBank, access October 2010). These taxa are discussed as possibly deep positioned within (or strepsipterans even outside) Endopterygota. Yet, current studies suggest contradicting results and consider deep endopterygote relationships as not completely resolved (Beutel and Gorb, 2006; Wiegmann et al., 2009; Friedrich and Beutel, 2010; Beutel et al., 2010; McKenna and Farrell, 2010).

2.5. Conclusions

Applied methods in this study, for example, evaluation of data sets with MARE and applying reduction heuristics to select optimal subsets, demonstrate that data quality assessment considering information content and a cautious selection of data sets pay off within phylogenomics. These methods can be recommended for phylogenomic data in general. Still, several effects, for example an impact of proteome species and single taxa on the tree topology are not well understood. This is in particular obvious considering clades like Entognatha or Mandibulata / Myriochelata. They seem very sensitive to taxa/gene selection and to different reconstruction methods. While Euarthropoda, Pancrustacea and Hexapoda including primary wingless hexapods, Ectognatha, Endopterygota and hymenopterans as sister groupt to remaining endopterygotes are robustly resolved, inference of the hexapod sister group and Entognatha remains difficult. Current data sets, which comprise the largest phylogenomic study for arthropods, (not unequivocally) support Entognatha with a high support for a sister group relationship of proturans and diplurans (Nonoculata). Currently, there is no alternative hypothesis which gains more support from multi-gene approaches. The placement of some taxa is still problematic even in this phylogenomic approach, for example the position of myriapods which could not clearly be addressed. Admittedly, the data availability for myriapods is not that good, but applied methods point towards a systematic phenomenon of myriapod molecular evolution. Still, EST data are missing from important taxa, e.g. Symphyla, Pauropoda, Remipedia, Zygentoma, which are crucial for the arthropod tree of life. Applied methods in the current thesis are promising for high quality analyses, but are worth to refine. In tree reconstructions, there is a high need to understand and to handle properly 'rogue' taxa. Additionally, EST data from upcoming 454 techniques will require much more computational power and a careful data assembly and processing techniques. This phylogenomic study, although raising hope to reach a resolved arthropod tree, still faces challenges in interpreting the strength and quality of the phylogenetic signal. Unresolved incongruencies between morphological and molecular analyses should challenge systematists to present the strength, quality and deficiencies of their evidence, and work towards resolving outstanding issues.

3. Arthropod Phylogeny inferred from nuclear rRNAs genes

3.1. Background

Ribosomal RNA genes and its use for phylogenetic inference

The large (28S) and the small (18S) nuclear ribosomal RNA gene are popular markers for inferring deep phylogenies. Studies focusing on the reconstruction of ancient splits in metazoans, and for example in arthropods or arthropod subgroups have relied on these markers (e.g. Mallatt and Giribet, 2006; Mallatt et al., 2010; Kjer, 2004; Luan et al., 2005; Misof et al., 2007; Gao et al., 2008; Dell'Ampio et al., 2009; Letsch et al., 2009).

Ribosomal RNA genes are structural genes and exhibit loop and stem regions. Within loops, nucleotide sites are unpaired and evolve independently. Within stem regions, a nucleotide site has a corresponding pairing site. Stems contribute to a secondary structure of rRNA molecules that form subunits of ribosomes. Structure features (secondary and tertiary structures) maintain functions within a ribosome and are targets of natural selection. Primary sequences may vary as long as functional domains are structurally retained. Stem regions underly compensatory mutations: mutation of a single site within a stem covaries with its corresponding paired position to retain the secondary structure within the molecule. Ignoring correlated variance of paired positions may mislead tree reconstructions (e.g. Jow et al., 2002; Galtier, 2004; Misof et al., 2007).

Recent molecular rRNA studies show particular nodes, for example Nonoculata, that were discussed as biased result due to non-stationary evolutionary processes (Luan et al., 2004, 2005; Dell'Ampio et al., 2009). The reconstruction of ancient splits is dependent on taxon sampling and character choice, as signal might be low in single lineages. All the more important is a proper handling of data and its quality assessment prior to tree reconstructions (e.g. secondary structure guided alignments, alignment masking and evaluation of compositional heterogeneity within data sets) to obtain reliable results. The resolution of arthropod phylogeny remains controversial (see chapter 1).

The idea behind the present approach is to improve phylogenetic reconstruction from rRNA genes. Following issues were considered: i) a cautious taxon sampling, covering all euarthropod orders with special focus on apterygote hexapods and myriapods ii) consideration of paired sites with structure guided alignments iii) alignment masking iv) the employment of realistic models of sequence evolution which take secondary structure information and non-stationary substitution processes into account[1].

[1] A part of present analyses has been published in von Reumont et al. (2009): von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits R, Luan YX, Wägele JW, Pass G, Hadrys H, Misof B (2009): Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. BMC Evolutionary Biology (2009) 9:119.

Available and new data for arthropod rRNA single gene analyses

Molecular analyses based on nuclear rRNA markers usually comprise large data sets, but mostly lack important representatives (e.g. pauropods are not included in Mallatt and Giribet, 2006). Other studies are exclusively focused on arthropod subgroups, for example on myriapods (Gai et al., 2006), apterygote hexapods (Luan et al., 2005; Gao et al., 2008; Dell'Ampio et al., 2009) or insects (Kjer, 2004; Misof et al., 2007). Mostly, only one nuclear rRNA marker or gene fragments have been included in analyses (Luan et al., 2003, 2005; Gao et al., 2008; Dell'Ampio et al., 2009). The present data set includes both (nearly) complete nuclear rRNA genes. It represents a well-balanced taxon sampling across euarthropod groups plus velvet worms and water bears (148 taxa). Whenever possible, taxa were included with putatively primitive morphological characters (as recommended in Philippe et al., 2005a) and (Lartillot and Philippe, 2008). Complete 18S genes have been sequenced from 26 species of apterygote hexapods and myriapods. Complementary, complete or nearly complete 28S sequences of identical or similar species were kindly provided by E. Dell'Ampio and D. Bartel (Group Pass, University of Vienna).

Alignment strategies and modeling data with respect to paired sites

Correct modeling is of particular importance for reliable tree reconstructions (Philippe et al., 2000; Rodríguez-Ezpeleta et al., 2007; Lartillot and Philippe, 2008). The extent to which biological processes can or should be modeled is unclear. Problems might occur, if models are misspecified or over-parametrized (Kelchner and Thomas, 2007; Grievink et al., 2010). Therefore, analyses of rRNA sequences can still deliver new insights in this direction, since a comprehensive background knowledge allows a distinction between different aspects. Structure guided alignments should help to reduce erroneous homology assignment within rRNA alignments. Mainly, structure guided alignments and hypothetical consensus structures (see below) are manually constructed from a structure constraint and comparative sequence analyses (Kjer, 2004; Kjer and Honeycutt, 2007; Dohrmann et al., 2008; Dell'Ampio et al., 2009; Mallatt et al., 2010). In order to model covariation, secondary structure interactions were estimated by the alignment software RNAsalsa (Stocsits et al., 2008, 2009). The software takes secondary structure into account. Evolutionary constraints resulting from secondary structure interactions are well known (e.g. Gillespie et al., 2005a). The accuracy of rRNA comparative structure models (Fox and Woese, 1975; Wuyts et al., 2000; Gutell and Cannone, 2002) was confirmed by crystallographic analyses (Ban et al., 2000; Noller, 2005). By incorporation of a given secondary structure string, the software corroborates underlying hypotheses of positional homology as accurately as possible. Essentially, this approach combines prior knowledge of conserved site interactions in a canonical eukaryote secondary structure consensus model. Additionally, alternative and/or additional site interactions are estimated, supported by the specific data. Site covariation patterns were used to guide the application of mixed substitution models (RNA/DNA models) in subsequent phylogenetic analyses. This may help to impede inadequate modeling of rRNA substitution processes in deep phylogenetic inference (Brown and Lemmon, 2007; Misof et al., 2007). A consensus secondary structure out of the alignment approach is required applying mixed models. It can be understood as model parameter: it defines site interactions and thus character interdependence due to compensatory mutations (Hancock et al., 1988; Stephan, 1996; Misof et al., 2007). The software *PHASE-2.0* includes specialized substitution models for RNA genes with conserved secondary structure; concurrently, it allows modeling non-stationarity (Gowri-Shankar and Jow, 2006).

Modeling non-stationary processes

If present, non-stationary processes clearly violate assumptions of stationarity regularly assumed in phylogenetic analyses (Blanquart and Lartillot, 2006; Gowri-Shankar and Rattray, 2006, 2007). Inhomogeneous base composition across taxa is a frequently observed phenomenon indicating non-stationary substitution processes (Galtier and Gouy, 1995; Tarrío et al., 2001; Gowri-Shankar and Rattray, 2007). In several studies where non-stationarity has been recognized, LogDet analyses were applied, or respective taxa which caused non-stationarity were excluded (Luan et al., 2005; Dell'Ampio et al., 2009). Alternatively, non-stationarity was ignored in modeling (Gao et al., 2008; Mallatt et al., 2010), but tools for modeling non-stationarity are available, for example the software P4 (Foster, 2004, maximum likelihood inference) or *PHASE-2.0* (Gowri-Shankar and Jow, 2006, Bayesian inference). In the present study, observed non-stationarity was modeled in combination with mixed RNA/DNA substitution models in a Bayesian framework. The idea was to obtain a better fit than with standard substitution models (Telford et al., 2005; Gowri-Shankar and Rattray, 2007). Additionally, modelling non-stationarity allowed for lineage specific variation of the model of evolution. In *PHASE-2.0*, a non-homogeneous substitution model is implemented "[...] by introducing a reversible jump Markov chain Monte Carlo method for efficient Bayesian inference of the model order along with other phylogenetic parameters of interest" (Gowri-Shankar and Rattray, 2006; Gowri-Shankar and Jow, 2006). Application of a new hierarchical prior leads to more reasonable results when only a small number of lineages share a particular substitution process. As far known, *PHASE-2.0* is the only software where concurrently it is possible to apply mixed models and non-stationary processes.

3.2. Materials and Methods

3.2.1. Taxon sampling

Specimens of apterygote, hexapod orders and myriapods were sampled with a self-made mini-exhauster from different sampling locations (Tab. A.7). Specimens were placed into a box and finally preserved in 99% ethanol. Alternatively, DNA extractions or preserved specimens were provided by our cooperation partners (European Basal Hexapod Work group, Siena, Vienna; Group Machida, Japan and Group Luan, China). 18S and 28S new sequences of 35 species were included (Tab. 3.1) which cover all primary wingless hexapods orders and Pauropoda, Symphyla, Diplopoda and Chilopoda (Myriapoda). Sequences has been deposited at GenBank, NCBI.

A total of 148 arthropod taxa were sampled (Tab. A.8) representing all arthropod clades including onychophorans and tardigrades. New sequences were provided from other cooperation partners (crustaceans: Reumont/Wägele, ZFMK Bonn; odonates and mayflies: Letsch/Misof, ZFMK Bonn; Simon/Hadrys, ITZ Hannover).

3.2.2. Laboratory work

Gene amplification, purification and sequence editing for ribosomal RNA genes took place at the molecular unit of the ZFMK. Collected material and provided DNA extractions were preserved in 99% ethanol and stored at $-20\,^{\circ}C$.

DNA extraction, gene amplification and sequencing

DNA extraction of complete specimens or tissue was conducted with the DNeasy Blood & Tissue Kit (Qiagen®). Single specimens were previously macerated. The manufacturers protocol was slightly modified (overnight incubation and adding 8 µl RNAse [10 mg/ml] after lysis). Extracted genomic DNA was amplified with the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare®) for tiny

Table 3.1.: Species list of newly sequenced nuclear rRNA genes (12/2008). 18S rRNA sequences (blue printed) were sequenced at the ZFMK, 28S rRNA sequences for apterygote hexapods and myriapods were kindly provided by work group Pass (University of Vienna). All sequences has been deposited in GenBank, NCBI with given Accession numbers.

Taxon	Group	Accession numbers	Gene
Craterostigmus tasmanianus	Chilopoda	EU376009, EU368617	28S, 18S
Lithobius forficatus	Chilopoda	EU368618	18S
Polyxenus lagurus	Diplopoda	EU376011, EU368619	28S, 18S
Monographis sp.	Diplopoda	EF192437	28S
Polydesmus complanatus	Diplopoda	EU376010, EU368620	28S, 18S
Cylindroiulus caeruleocinctus	Diplopoda	EF199985, EU368621	28S, 18S
Pauropodidae	Pauropoda	EU376012, EU368622	28S, 18S
Acerentomon franzi	Protura	EF199976, EU368597	28S, 18S
Baculentulus densus	Protura	EU376049	28S
Eosentomon sp.	Protura	EU376047, EU368598	28S, 18S
Eosentomon sakura	Protura	EF192434	28S
Sinentomon erythranum	Protura	EF192442	28S
Campodea augens	Diplura	EF199977, EU368599	28S, 18S
Lepidocampa weberi	Diplura	EU376050	28S
Catajapyx aquilonaris	Diplura	EF199978, EU368600	28S, 18S
Parajapyx emeryanus	Diplura	EF192440	28S
Octostigma sinensis	Diplura	EF192439	28S
Tetrodontophora bielanensis	Collembola	EU376051	28S
Gomphiocephalus hodgsoni	Collembola	EF199969, EU368601	28S, 18S
Bilobella aurantiaca	Collembola	AJ251729, EU368602	28S, 18S
Anurida maritima	Collembola	AJ251738, EU368603	28S, 18S
Podura aquatica	Collembola	EF199970, EU368604	28S, 18S
Cryptopygus antarcticus	Collembola	EF199971, EU368605	28S, 18S
Isotoma viridis	Collembola	EU376052	28S
Orchesella villosa	Collembola	EF199972, EU368606	28S, 18S
Pogonognathellus flavescens	Collembola	EU376053, EU368607	28S, 18S
Megalothorax minimus	Collembola	EF199975, EU368608	28S, 18S
Sminthurus viridis	Collembola	EF199973, EU368609	28S, 18S
Allacma fusca	Collembola	EU376054, EU368610	28S, 18S
Dicyrtomina saundersi	Collembola	EF199974, EU368611	28S, 18S
Machilis hrabei	Archaeognatha	EF199981, EU368612	28S, 18S
Lepismachilis y-signata	Archaeognatha	EF199980, EU368613	28S, 18S
Pedetontus okajimae	Archaeognatha	EU376055, EU368614	28S, 18S
Lepisma saccharina	Zygentoma	EU376048, EU368615	28S, 18S
Ctenolepisma longicaudata	Zygentoma	EU368616	18S

and rare specimens or those which are hard to collect (proturans or pauropods). The 18S rRNA was split into 3 or 4 fragments for amplification with PCR (primers see Fig. 3.1, Table 3.2). Following primer combinations were used:

A) 1F/5R, 3F/18Sbi, 5F/9R

B) 18S L0001/18S R0532, 18S L0466/18S R1100, 18S L0922/18S R1524, 18S L1362/18S R2090

C) 18S 1L/18S 1R, 18S L500/ 18S R1470, 18S L1210/18S R1790, 18S 3L/18S 3R

D) 18SV0000/18Sbi5.0, 18Sai/18Sbi or alternative 18Sai/18SR1900, 18Sbirev/18SR1900

If necessary, primer pairs of A, B, C and D were combined to amplify the complete 18S sequence.

Table 3.2.: Primer list for amplification and sequencing of the 18S rRNA gene used for apterygote hexapods and myriapods

Primer	direction	sequence 5' - 3'	Reference
18SL0001	forward	TACCTGGTTGATCCTGCCAGT	Luan et al. (2003)
1F	forward	TACCTGGTTGATCCTGCCAGTAG	Giribet and Ribera (2000)
18S1L	forward	TACCTGGTTGATCCTGCCAGT	Luan et al. (2005)
18SV0000	forward	TACCTGGTGGATCCTGCCAGTA	Chalwatzis et al. (1995)
18SL0466	forward	GTTCGATTCCGGAGAGGGAG	Luan et al. (2003)
3F	forward	GTTCGATTCCGGAGAGGGA	Giribet et al. (1996)
18SL500	forward	GTTCGATTCCGGAGAGGGAG	Luan et al. (2005)
18Sai	forward	CCTGAGAAACGGCTACCACATC	Maddison et al. (1999)
18SL0922	forward	AATTGGAGTGCTCAAAGCAGGC	Luan et al. (2003)
5F	forward	GCGAAAGCATTTGCCAAGAA	Giribet et al. (1996)
18SL1210	forward	CCTTGAGAAAATTGGAGTGCT	Luan et al. (2005)
18Sbi rev	forward	TCCGATAACGAACGAGACTC	De Salle et al. (1992)
18SL1362	forward	CTTAATTTGACTCAACACGGG	Luan et al. (2003)
18S3L	forward	AGGAATTGACGGAAGGGCAC	Luan et al. (2005)
18SR0532	reverse	TTGCGCGCCTGCTGCCTTCC	Luan et al. (2003)
5R	reverse	CTTGGCAAATGCTTTCGC	Giribet et al. (1996)
18S1R	reverse	TAATATACGCTATTGGAGCTGG	Luan et al. (2005)
18Sbi5.0	reverse	TAACCGCAACAACTTTAAT	De Salle et al. (1992)
18SR1100	reverse	CGACGATCCAAGAATTTCAC	Luan et al. (2003)
18Sbi	reverse	GAGTCTCGTTCGTTATCGGA	Maddison et al. (1999)
18SR1470	reverse	TTAGAACTAGGGCGGTATCTG	Luan et al. (2005)
18SR1524	reverse	AGTCTCGTTCGTTATCGGAAT	Luan et al. (2003)
9R	reverse	GATCCTTCCGCAGGTTCACCTAC	Giribet et al. (1996)
18SR1790	reverse	CGTTACCGGAATGAACCAGAC	Luan et al. (2005)
18SR1900	reverse	TAATGATCCTTCTGCAGGTTCACCTACG	Chalwatzis et al. (1995)
18SR2090 or 18S3R	reverse	CCTACGGAAACCTTGTTACG	Luan et al. (2003, 2005)

Primers were ordered from Metabion™. For amplification of some fragments, the PCR-Multiplex Kit (Qiagen®) was more effective, because pooling of weak PCR products was avoided. In general, Hotstart PCRs (Chou et al., 1992) were conducted. For PCR conditions and profiles, see Tab. A.9. Products were purified with the NucleoSpin Extract II Kit (Macherey-Nagel®). Alternatively, purification was conducted with enzymes ExoI/SAP: 0.12 μl ExoI (20 u/μl,Biolabs), 0.45 μl SAP (Shrimp Alkaline Phosphatase, 1 u/μ, Promega) and 2.43 μl RNAse-free sterile water was mixed on ice. 10 μl of PCR product was added. This mix was incubated for 15 minutes at 37 °C, following 20 minutes 75 °C incubation time and finally cooled down to 12 °C. Purified products were checked on agarose gels. Weak PCR products were pooled for purification. In case of multiple bands, fragments with expected size were cut from 1–1.5% agarose gel and purified according to manufacturers protocol (NucleoSpin Extract II, Macherey-Nagel). DNA concentration was measured with a mass marker (BioRad) and with a Nanodrop Spectrophotometer ND-1000 (peqLab). Cycle Sequencing (CS) reactions were performed using DNA Quick Start Mastermix (Beckman Coulter). CS products were ethanol-precipitated or

purified with a CleanSeq magnetic bead system (Agencourt). Double stranded Sanger-sequencing was conducted on Beckman Coulter capillary sequencers CEQTM 8000 and CEQTM 8800. Few amplified and purified PCR products were sequenced by Macrogen (Inc.), Korea.

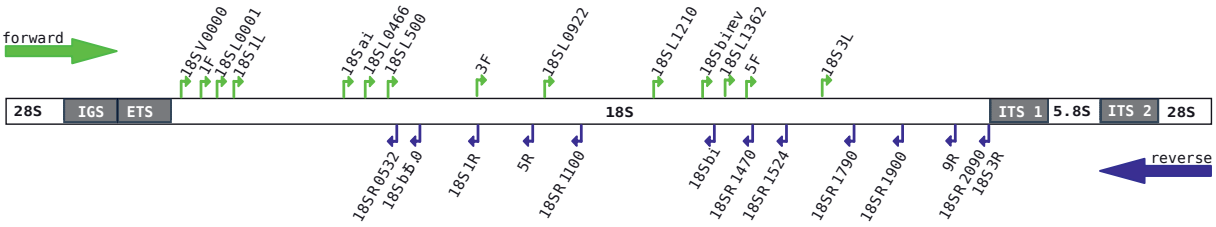

Figure 3.1.: Primer card for the amplification of the 18S used for hexapods and myriapods. Positions of forward primers: green arrows; positions of revers primers: blue arrows. Different primers that dock on similar positions are given at a single arrow.

3.2.3. Sequence editing, quality check and data setup

Sequence electropherograms were analyzed and assembled into consensus sequences with SeqMan (DNAStar Lasergene) or BioEdit 7.0 (Hall, 1999). Consensus sequences were checked in NCBI using BLASTN, mega BLAST or "BLAST 2 SEQUENCES" (Tatusova and Madden, 1999) to exclude contamination. Additionally, sequences were provided from cooperation partners and downloaded from GenBank, NCBI. In total, 148 concatenated 18S and 28S rRNA sequences were included in analyses. Only sequences were considered, which span at least 1,500 bp for the 18S and 3,000 bp for the 28S (Tab. A.8). For 29 taxa, 18S and 28S rRNA sequences were concatenated from different species (Tab. A.10). Species were chosen as closely related as possible. *Milnesium* sp. (Tardigrada) was included as outgroup taxon.

3.2.4. Alignment and alignment masking

Secondary structure guided alignments

Secondary structures of rRNA genes were considered to improve sequence alignment, as advocated in Kjer (1995); Hickson et al. (2000); Buckley et al. (2001) and Misof et al. (2006). Both genes were separately prealigned with MUSCLE v3.6 (Edgar, 2004b). Sequences of 24 pterygotes were added applying a profile to profile alignment (Edgar, 2004a). 28S sequences of *Hutchinsoniella macracantha* (Cephalocarida), *Speleonectes tulumensis* (Remipedia), *Raillietiella* sp. (Pentastomida), *Eosentomon* sp. (Protura) and *Lepisma saccharina* (Zygentoma) were incomplete. Apart from *L. saccharina*, prealignments of these taxa were manually corrected; correct positions of each sequence fragment were identified by "BLAST 2 SEQUENCES" (NCBI Blast).

The 28S + 5.8S (U53879) and 18S (V01335) from *Saccharomyces cerevisae* and corresponding secondary structures were extracted from the European Ribosomal Database (Van de Peer et al., 2000; Wuyts et al., 2002, 2004). The 5.8S was included in the constraint to avoid artificial stems due to folding interactions between 28S and 5.8S (Michot et al., 1983; Gillespie et al., 2005b,a). Structure strings were converted into dot-bracket-format using Perl scripts. Alignment sections presumably involved in formation of pseudo-knots were locked from folding. Pseudo-knots in *Saccharomyces cerevisae* are known for the 18S (stem 1 and stem 20, V4-region: stem E23_9, E23_10, E23_11 and E23_13). Prealignments and structure constraints served as input for structure guided alignments

with RNAsalsa (Stocsits et al., 2009). RNAsalsa alignments ran with default settings.

Due to missing parts of the 18S of *Speleonectes tulumensis* and the 28S of *Raillietiella* sp., chimeran sequences were generated manually (*S. tulumensis*: EU370431, L81936: 18S; *Raillietiella* sp.: EU370448, AY744894: 28S). For *Speleonectes*, position 1–1644 (L81936) and position 1645–3436 (EU370043) were concatenated. For *Raillietiella*, position 1–3331 (AY744894) and position 3332–7838 (EU370448) were concatenated into one sequence (numbers refer to aligned positions).

Alignment masking

Structure guided alignments were checked for randomly similar aligned positions with Aliscore (Misof and Misof, 2009). The sliding window size was set to default ($w = 6$) and gaps were treated as ambiguities (-N option). The maximum number of possible random pairwise comparisons (-r: 10,878) was used for consensus profiles. Alignments were masked with ALICUT (Kück, 2009). Since Aliscore currently ignores base pairings, ambiguously aligned positions within stems were discarded and corresponding positions were handled as unpaired in tree reconstructions. Finally, masked 18S and 28S alignments were concatenated, including respective consensus structure strings for tree reconstructions with mixed RNA/DNA models.

3.2.5. Split analyses

Phylogenetic networks (Huson and Bryant, 2006) were calculated from unmasked and masked alignments using SplitsTree 4.10 to analyze the information content of present data. Neighbornet graphs based on the neighbornet algorithm (Bryant and Moulton, 2004) with uncorrected p-distances. Additionally, LogDet transformation was applied (e.g. Steel, 1994; Penny et al., 1994; Steel et al., 2000). The proportion of invariable (constant) sites displayed in SplitsTree 4.10 was considered in calculation of LogDet networks. LogDet is a distance transformation (Steel, 1994) that corrects for biases in base composition. Neighbornet graph gives first indications of signal-like patterns and conflicts in alignments.

3.2.6. Compositional base heterogeneity

Alignments were checked for compositional base heterogeneity applying a χ^2-test implemented in *PAUP* 4.0b10 (Swofford, 2003). Both alignments (18S, 28) were separately tested before concatenation. Parsimony uninformative positions and randomly similar alignment blocks were excluded. Additionally, paired and unpaired sites were separately tested for each gene due to determine the minimal number of groups (submodels) for tree reconstruction in *PHASE* modeling non-stationarity. The test was repeated for both partitions as used in tree reconstruction: if stems were disrupted by discarding a paired site due to alignment masking, retained and formerly paired positions were treated as unpaired.

3.2.7. Phylogenetic reconstructions

Phylogenetic reconstructions were conducted within a Bayesian framework from the concatenated, masked gene alignment. To compare a possible impact of mixed models and modeling *versus* ignoring non-stationarity analyses were conducted a) with mixed models for both, a stationary (time-

homogeneous) *versus* a non-stationary (time-heterogeneous) setup and b) with 'standard' DNA models modeling stationarity *versus* modeling non-stationarity. All trees were rooted with *Milnesium* (Tardigrada).

Mixed RNA/DNA models and modeling stationarity versus non-stationarity

Within a mixed model approach, loop partitions were modeled with DNA models. Stem partitions were modeled with RNA models considering covariation patterns. Among site rate variation (ASRV, Yang, 1996) was implemented in both types of substitution models. Inhomogeneity of base composition suggested non-stationary processes of sequence evolution across the data set (see section 3.3). Analyses were performed with *PHASE-2.0* (Gowri-Shankar and Jow, 2006) to consider non-stationarity and concurrently mixed models. This may accommodate compositional heterogeneity and minimize bias in tree reconstructions according to ideas developed by Foster (2004).

The number of candidate models was restricted to REV + Γ, TN93 + Γ and the HKY85 + Γ models for loops and corresponding RNA models for stems (RNA16I + Γ, RNA16J + Γ and RNA16K + Γ). Site heterogeneity was modeled by a discrete gamma distribution (Yang, 1994) with six categories. The extent of invariable characters was not estimated. Considering invariable characters was recommended by (Shoemaker and Fitch, 1989). It has been shown that an estimation of invariable sites strongly correlates with an estimation of gamma shape parameter (Yang, 1996; Waddell et al., 1997; Sullivan and Swofford, 2001; Kelchner and Thomas, 2007). The concatenated data set was divided into four partitions (loop / stem, 18S and loop / stem 28S) with respect to different evolutionary rates for nuclear rRNA genes (Hillis and Dixon, 1991). RNA and DNA substitution model parameters were independently estimated per partition. Substitution models were selected based on results of a stationary pre-run (Fig. 3.2). Three different RNA (16 state) substitution models with corresponding DNA models were preselected and tested: REV + Γ & RNA16I + Γ, TN93 + Γ & RNA16J + Γ and HKY85 + Γ & RNA16K + Γ. For priors, a Dirichlet distribution and proposals for a set of exchangeable parameters were used[2].

Appropriate sampling of the parameter space according to the posterior density function (Zwickl and Holder, 2004) was checked by plotting values of each parameter. Parameter convergence was monitored with the statistical software package R (R Development Core Team, 2008, v2.9) for all model combinations (500,000 generations, sampling period: 150). Model combinations were excluded where several parameter values did not converge. MCMC processes were repeated with 3 million generations (sampling period: 150 generations) for selected model combinations (burn-in: 299,999 generations). The mixed model with the best fitness was selected by a Bayes Factor Test (BFT) (Kaas and Raftery, 1995; Nylander et al., 2004) after a second check on convergence by eye. The favored model ($2lnB_{10} > 10$) was used for final phylogenetic reconstructions.

For each approach, 14 independent chains of 7 million generations plus two chains of 10 million generations (sampling period: 1000) ran on Linux clusters (AMD Opteron Dual Core, 64bit systems, 32 GB RAM). For each chain, a burn-in of 2 million generations was excluded. The stationary (time-homogeneous) setup was similar to the pre-run, except for the number of generations, sampling period and burn-in. The non-stationary (time-heterogeneous) setup differed (Fig. 3.2): following

[2]Model = MIXED; Tree, proposal priority = 1; Model, proposal priority = 5; Topology changes, proposal priority = 10; Branch lengths, proposal priority = 40; Model 1, 3, proposal priority = 7; Model 2, 4 proposal priority = 8; Average rates, proposal priority = 1; Frequencies, proposal priority = 2; Rate ratios, proposal priority = 1; Gamma parameter, proposal priority = 1; random seed: individual seed set per run, (see Gowri-Shankar and Rattray, 2007)

Foster (2004) and Gowri-Shankar and Rattray (2007), only a limited number of composition vectors can be shared by different branches in the tree. Exchangeability parameters (average substitution rate ratio values, rate ratios and alpha shape parameter) served as input values and remained fixed during tree reconstruction. These parameter values had been calculated from results of preliminary stationary pre-runs in two steps: a) a consensus tree was inferred in *PHASE mcmcsummarize* from the output of the pre-runs. The topology of the consensus tree and the respective model file served as input for a ML estimation of parameters in *PHASE optimizer*. b) Estimated values of exchangeability parameters from the *optimizer* output file and estimated start values for base frequencies were loaded into *mcmcphase*. The number of allowed base frequency categories (submodels) along the tree was fixed. Accounting for submodels to individual branches in a tree by the MCMC process allow for a maximum of flexibility without losing the proper mix of parameters (Gowri-Shankar, pers. comm.). The number of submodels was set to 3, reflecting compositional base heterogeneity.

Harmonic means of *lnL* values of all chains per approach were compared with a BFT. Thereby, possible local optima were identified in which a single chain might have been trapped. Only sample data of chains with a $2lnB_{10}$–value < 10 (Kaas and Raftery, 1995) were merged into a "metachain" with Perl scripting (Beiko et al., 2006). The non-stationary consensus tree was inferred from ten non-stationary chains, the stationary consensus tree from three chains. Assembled metachains included 56 million generations for the non-stationary approach and 18 million generations for the stationary approach (see 3.3); burn-ins were discarded. Consensus trees and posterior probability values were inferred with *mcmcsummarize*. Branch lengths of the stationary and non-stationary consensus tree were estimated from different initial states of three *mcmcphase* chains (4 million generations, sampling period 500, topology changes turned off, starting tree = consensus tree, burn-in: 1 million generations). For this purpose, *PHASE* was especially modified by V. Gowri-Shankar. Data were combined with *mcmcsummarize* to infer mean branch lengths. Mean branch lengths were used to redraw the consensus tree.

DNA models and modeling stationarity versus non-stationarity

All analyses were repeated as described before, but instead applying mixed models, only the previously selected DNA model (TN93 + Γ) was used. Thus, interdependence of paired positions (stems) was ignored. Aim was to examine a possible impact of mixed models compared to DNA models (independent-sites models). The setup followed the description starting from pre-run II (first row, Fig. 3.2). Only sample data of chains with a $2lnB_{10}$–value < 10 (Kaas and Raftery, 1995) were merged into a metachain (see before). Inferring the non-stationary consensus tree, only three chains were included. For the stationary consensus tree, only one chain was included. All other chains were rejected by a BFT (see section 3.3). Assembled metachains included 15 million generations for the non-stationary approach and 5 million generations for the stationary approach (burn-ins discarded). Inferring consensus trees, estimation of posterior probability values and branch lengths were conducted likewise to the mixed model approach.

3.2.8. Consensus networks: Differences between 'mixed model' and 'DNA' trees

Topological differences between mixed model and DNA trees were visualized by consensus networks (Holland and Moulton, 2003). Consensus networks rely only on tree topology, node support values are not considered. Consensus networks were computed with SplitsTree 4.10 (Huson and Bryant, 2006).

The threshold of 0.01 visualizing differences was set to 0.01, edge weights were not considered. This was conducted for both, the stationary approach ('mixed model' *versus* 'DNA' tree) and the non-stationary approach ('mixed model' *versus* 'DNA' tree).

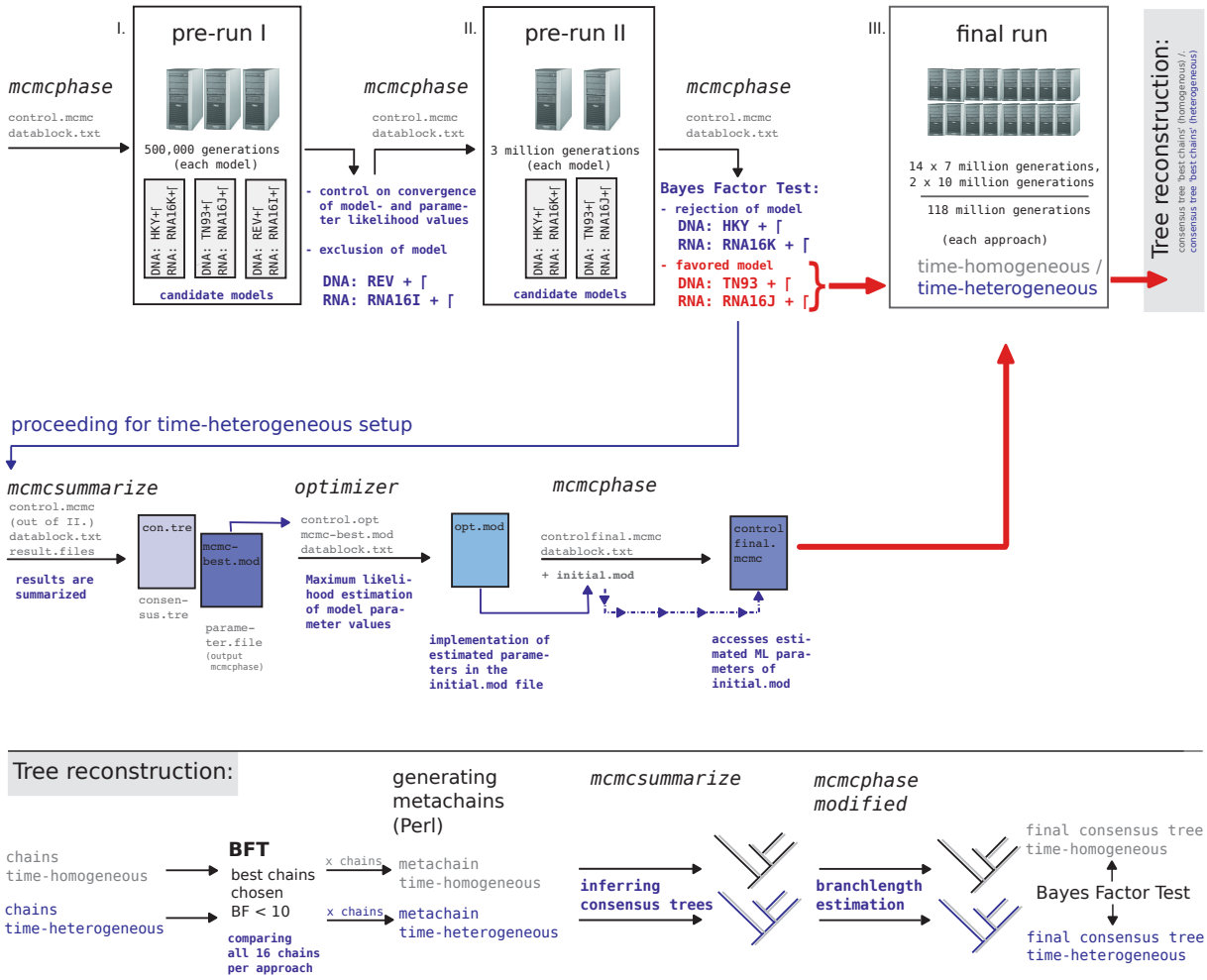

Figure 3.2.: Analyses setup for the time-homogeneous (stationary) and time-heterogeneous (non-stationary) approach with *PHASE-2.0* applying mixed RNA/DNA models. *PHASE* sub-packages are written in italics above black arrows, input files are written in Currier. Black arrows: analyses steps, blue arrows: results or parameter values were accessed by the following process. Red arrows: final run of the time-heterogeneous / time-homogeneous approach with 16 chains each. **First row:** I. 3 control files (control.mcmc) were prepared for *mcmcphase* using 3 different mixed models. Pre-run I was used for a first model selection (500,000 generations / model). Model REV + Γ & RNA16I + Γ was excluded because parameter values did not converge. II. Step one (I.) was repeated with 3 million generations using similar control files, but different number of generations and random seeds for remaining models. In a Bayes Factor Test (BFT) ln likelihoods values (harmonic means) of both chains were compared: mixed model HKY85 + Γ & RNA16K + Γ was rejected. Parameter values of the selected model TN93 + Γ & RNA16J + Γ were implemented in the time-heterogeneous setup. III.) 16 chains for both, the time-homogeneous and time-heterogeneous approach, were used in final analyses (final run). Stationary control files were similar to step II, except number of generations and random seeds. **Second row:** For the final time-heterogeneous setup, additional steps were necessary. The selected model *mcmcsummarize* was used to calculate a consensus tree. With *PHASE optimizer* a ML estimation was conducted for each parameter value (opt.mod) based on the inferred consensus tree and optimized parameter values (mcmc-best.mod, provided by *mcmcphase*). Estimated values were implemented in an initial.mod file. The initial.mod file and its parameter values were accessed by the control file of each final time-heterogeneous chain. Only topology and base frequencies were estimated. **Third row:** Trees were inferred separately for both approaches. For each approach, all chains were tested in a Bayes Factor Test against the chain with the best ln L. Chains with a $2lnB_{10}$–value < 10 were merged into a metachain with Perl scripting. A consensus topology for each approach was inferred with *mcmcsummarize*. To estimate branch lengths properly *mcmcphase* (modified) was used with a fixed topology. Resulting branch lengths were redrawn in consensus topologies.

3.3. Results

For the present data set, modeling non-stationarity clearly statistically outperforms stationary analyses. This is true for both, a) using mixed RNA/DNA models and b) using only standard DNA models. Several clades accepted by morphological, neuroanatomical and molecular studies, are only resolved with modeling non-stationarity, at least for mixed model analyses (Hexapoda, Ectognatha, Dicondylia etc.). Resolution of these clades corroborates the statistical result, but only serve as adequate and not as mandatory argument. Topological differences between the non-stationary, DNA tree and the non-stationary, RNA/DNA tree are marginal. The computational effort of a structure guided alignment procedure and alignment masking pays off: potential phylogenetic signal has been increased, noise has been reduced. Split-decomposition networks still display noise or conflicts which may give explanations for some inconsistencies.

3.3.1. Alignment and alignment masking

The 18S rRNA alignment span 3,503 and the 28S rRNA 8,184 positions. Secondary consensus structures found by RNAsalsa include 794 paired positions in the 18S and 1,326 paired positions in the 28S. Consensus structures include paired sites that were present in 60% or more sequences (default $s3 = 0.6$ in RNAsalsa). Aliscore scored 1,873 positions of the 18S and 5,712 positions of the 28S alignment as randomly similar. Masked alignments show 1,630 nucleotide positions (18S) and 2,472 positions (28S). The concatenated masked alignment spans 4,102 positions and has been used for all tree reconstructions.

3.3.2. Split decomposition patterns

Both neighbornet graphs, with uncorrected p-distances (Fig. 3.3a) and LogDet correction (Fig. 3.3b) show little tree–likeness. This might indicate some problems typical in studies of deep phylogenetic questions: i) Some taxa like Diptera (which do not cluster with ectognathous insects), Diplura, Protura and Collembola each appear in a different part of the network. Diplura and Protura are separated from other hexapods. *Lepisma* is separated from the second zygentoman *Ctenolepisma* which is nested within Ectognatha. Symphyla, Pauropoda (Myriapoda), Remipedia and Cephalocarida (crustaceans) have very long branches. They might be misplaced due to signal erosion or occurrence of homoplasies: their placement in trees should be considered critically (Wägele and Mayer, 2007). The usage of LogDet distances adjusts the length of some branches but does not decrease the amount of conflicts in deep divergence splits. ii) The inner part of both networks (based on the masked alignment) show little treeness. Some taxa have long stem-lineages: these species might share distinct nucleotide patterns that are absent in other taxa. Such well separated groups are Copepoda, Branchiopoda, Cirripedia, Symphyla, Collembola, Diplura, Protura and Diptera. In contrast, e.g. Myriapoda "partim", Chelicerata and Ectognatha excluding Diptera share weaker patterns.

The neighbornet graph (LogDet) based on the concatenated unmasked alignment has a tendency to a fuzzier pattern than the neighbornet graph which was calculated from the masked alignment. Although differences were not that obvious, less conflict is observed for several clades, for example chelicerates, branchiopod crustaceans or ectognathous insects (excluding Diptera and *Lepisma*). By zooming the inner parts of both LogDet networks (Fig. 3.4), differences are obvious. In the masked graph, the pattern is slightly more distinct for several clades, e.g. pterygote insects (excluding Diptera),

copepods or branchiopods (Fig. 3.4b). Thus, alignment masking increased signal and reduced noise within the data set.

3.3.3. Compositional base heterogeneity

Separate tests for compositional base heterogeneity of both gene alignments (18S and 28S) in *PAUP* 4.0b10, reject the null hypothesis (H_0) which assumes a homogeneous base composition across taxa (18S: $\chi^2 = 1168.94, df = 441, P = 0.00$; 28S: $\chi^2 = 1279.98, df = 441, P = 0.00$).

Separate testing of all four partitions confirm base compositional heterogeneity across taxa ($P = 0.00$ in all four partitions; 18S: loops: 477 positions, paired sites: 424 positions; 28S: loops: 515 positions, paired sites: 637 positions). Repetition of tests for single partitions as used in tree reconstruction mainly show similar results. Stems disrupted by discarding a corresponding paired site due to alignment masking are treated as unpaired (see 3.2.6). Hence, 1848 sites of the concatenated alignment (18S: 706; 28S: 1,142) are treated as paired in all analyses. Heterogeneity is corroborated for unpaired sites of both genes ($P = 0.00$ for both genes; 18S: 506 characters; 28S: 567 characters). For paired sites, again the null hypothesis H_0 is rejected (18S, 395 characters included: $P < 0.0003$, 28S, 585 characters included: $P = 0.00$). Since non-stationary processes in all tests is strongly supported, modeling non-stationarity has been implemented to account for lineage-specific substitution patterns. To fix the number of "free base frequency sub-models" in non-stationary analyses, the minimal exclusive set of groups was evaluated. Based on χ^2-tests, the data set was divided into three groups for both rRNA genes. In both genes, Diptera show a high A/T content and Diplura show a low A/T content. Exclusion of only one of the groups was not sufficient to obtain homogeneity (18S: excluding Diptera: $\chi^2 = 972.91, df = 423, P = 0.00$, excluding Diplura: $\chi^2 = 532.13, df = 423, P < 0.0003$; 28S: excluding Diptera: $\chi^2 = 986.72, df = 423, P = 0.00$, excluding Diplura: $\chi^2 = 813.8, df = 423, P = 0.00$). Simultaneous exclusion of both groups led to acceptance of H_0 for the 18S alignment ($\chi^2 = 342.22, df = 405, P = 0.99$). For the 28S, exclusion of both, Diptera and Diplura, H_0 was still rejected ($\chi^2 = 524.98, df = 405, P < 0.0001$). After sorting taxa according to base frequencies in ascending order, only an additional exclusion of *Peripatus* sp. (Onychophora) and *Sinentomon erythranum* (Protura) resulted in base compositional homogeneity for the 28S (H_0: $\chi^2 = 434.99, df = 399, P = 0.1$). Thus, three sub-models are sufficient for this data set to model non-stationary. χ^2-tests were subsequently repeated for stem and loop regions of each gene. The exclusion of Diplura was sufficient to obtain homogeneity in loop regions for both genes (18S: 474 characters, $P = 0.9757$; 28S: 541 characters, $P = 0.0684$). For 18S stem regions, it was likewise sufficient to exclude either Diptera (378 characters, $P = 0.6635$) or Diplura (385 characters, $P = 0.99$). Thus, two sub-models would be sufficient considering the 18S only. In contrast, for 28S stem regions, homogeneity was received only after exclusion of both, Diptera and Diplura (547 characters, $P = 0.99$, Tab. 3.3). Since *PHASE-2.0* does not allow to vary the number of chosen sub-models among partitions, finally three sub-models have been fitted to each data partition.

Excluded groups, Diptera, Diplura and (Diplura + *Peripatus* + *Sinentomon*) were tested for compositional base heterogeneity without considering stems and loops. All tests reject homogeneity among these single taxa. Since parsimony uninformative sites were excluded, only 25-30 sites for each group were considered. χ^2-tests for such data might be statistically not valid due to a low number of included sites.

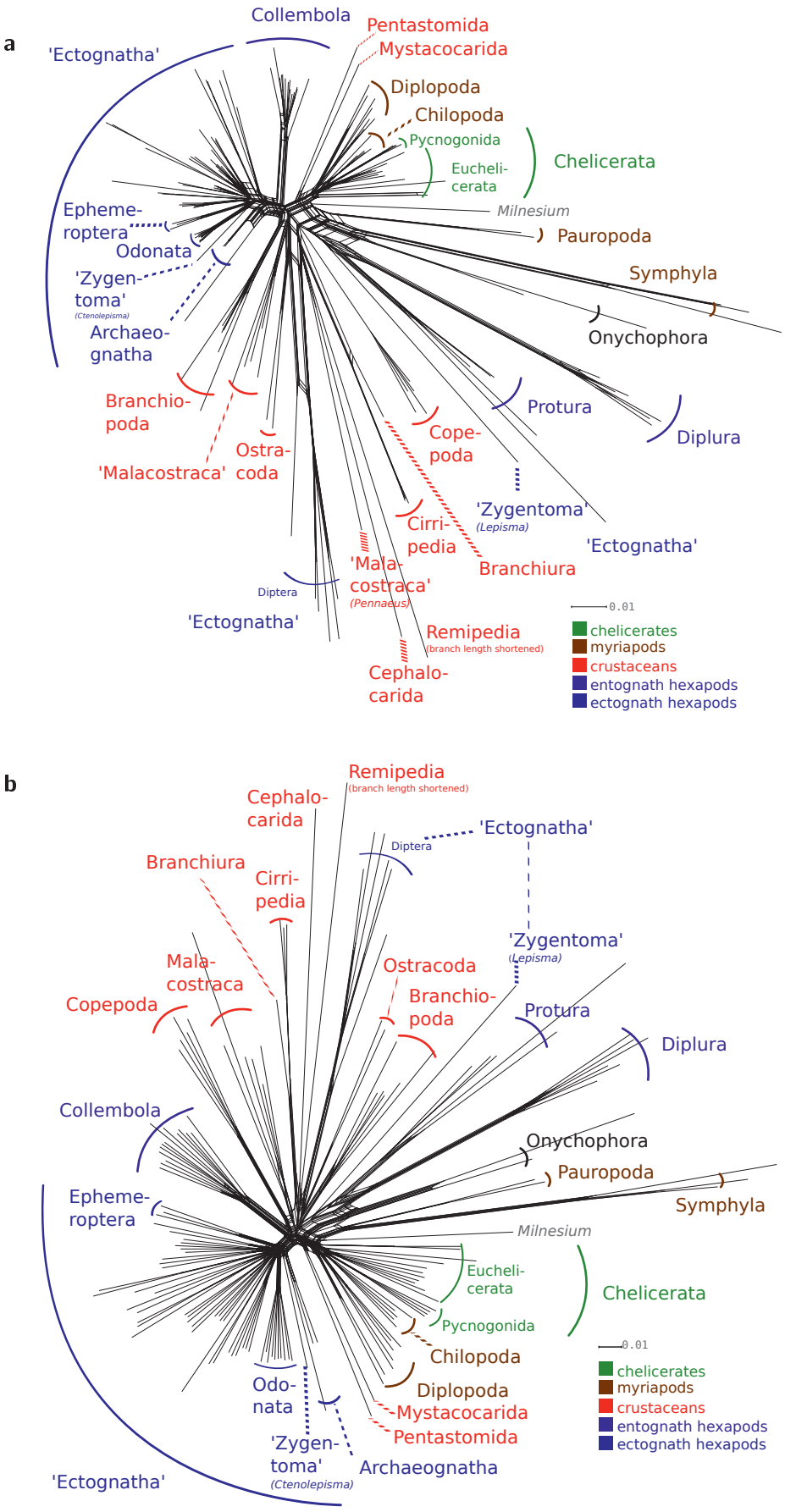

Figure 3.3.: Neighbornet graphs from the concatenated masked alignment (4,102 characters). Neighbornet graphs (using uncorrected p-distances and LogDet analysis with 30.79% invariable sites estimated in PAUP) were calculated with SplitsTree 4.10. **a** Neighbornet (uncorrected p-distances) of the concatenated masked alignment. **b** LogDet neighbornet of the concatenated masked alignment. Monophyly of clades is not necessarily supported in given neighbornet graphs. Quotation marks indicate obviously polyphyletic clades.

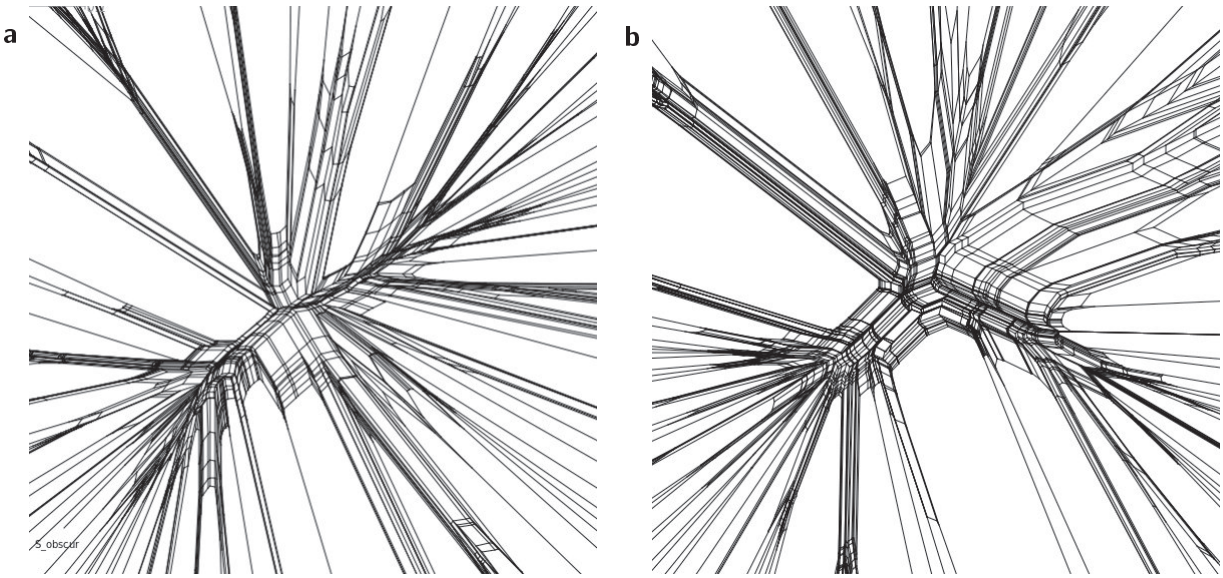

Figure 3.4.: Inner sections of LogDet neighbornet graphs based on the unmasked (11,678 characters) and masked (4,102 characters) concatenated alignment. Proportion of invariable sites: 1,866% (218 sites), unmasked alignment; 5.29% (217 sites), masked alignment.**a** Section of the unmasked LogDet neighbornet graph **b** Section of the masked LogDet neighbornet. Within deep splits, patterns are slightly more distinct compared with patterns in Fig 3.4a.

3.3.4. Phylogenetic reconstructions

Mixed DNA/RNA models

Three combinations of mixed models (REV + Γ & RNA16I + Γ, TN93 + Γ & RNA16J + Γ and HKY85 + Γ & RNA16K + Γ) were compared to select the best model set. The overall *ln* likelihood converged for all tested models (burn-in: 250,499 generations, pre-run I). Most parameters did not converge for combined REV + Γ & RNA16I + Γ models; this model mixed was excluded from further analyses (see Fig. 3.2). For each of the remaining model sets, one chain was initiated for 3 million generations (burn-in: 299,999 generations). The Bayes Factor Test (Kaas and Raftery, 1995; Nylander et al., 2004), strongly favored model set TN93 + Γ & RNA16J + Γ ($2lnB_{10} = 425.39$, harmonic mean lnL_0(TN93 + Γ & RNA16J + Γ)= 79791.08; harmonic mean lnL_1(HKY85 + Γ & RNA16K + Γ)= 80003.78). Subsequently to final stationary and non-stationary analyses, all single chains which passed a threshold value of $2lnB_{10} < 10$ in a BFT were assembled to a metachain (Fig. 3.2). Support values (Tab. 3.4, columns RNA/DNA) derived from 56,000 sampled trees for the non-stationary approach (Fig. 3.5) and from 18,000 sampled trees for the stationary approach (Fig. 3.6). Harmonic means of *ln* likelihoods from included non-stationary chains were compared against all *ln* likelihoods of included stationary chains (burn-in discarded) in a final BFT. Modeling non-stationary processes was strongly favored ($2lnB_{10} = 1362.13$).

DNA models

Analogue to the mixed model approach, chains of DNA runs (TN93 + Γ) with a threshold value of $2lnB_{10} < 10$ (BFT) were merged into one metachain. Three non-stationary chains were combined to

infer the consensus tree. For the stationary approach, the consensus tree was derived from only one chain. Support values were deduced from 15,000 sampled trees (3 chains) for the non-stationary set and from 5,000 sampled trees (one chain) for the stationary set (Tab. 3.4, columns DNA). Harmonic means of *ln* likelihoods of included non-stationary chains were compared against *ln* likelihoods of included stationary chains (burn-in discarded) in a final BFT. Again, modeling non-stationarity was strongly favored ($2lnB_{10} = 1544.12$). Comparing non-stationary DNA models *versus* mixed models, however, standard DNA modeling was favored ($2lnB_{10} = 3296.55$).

Table 3.3.: Results of the base frequency test conducted with PAUP. The data set was tested for compositional base heterogeneity with parsimony uninformative positions excluded. Gaps were treated as missing data. Tests were separately conducted for both genes (18S, 28S). Masked with Aliscore: randomly similar aligned sections were excluded in PAUP. Masked with Aliscore, disrupted stems treated as unpaired: remaining counterparts of disrupted stems by alignment masking were treated as unpaired (18S: 34 sites, 28S: 57 sites). no. incl.: number of included (taxa or sites, respectively). P-values indicating compositional base homogeneity are marked in blue. *Per.* sp.: *Peripatus* sp. (Onychophora); *S. erythr.*: *Sinentomon erythranum* (Hexapoda, Protura). o/o: a base frequency test was not conducted because compositional base homogeneity has already been supported by excluding another taxa group.

considered characters (- excluded taxa)		**masked with Aliscore**		**masked with Aliscore, disrupted stems treated as unpaired**	
	no. incl. taxa	no. incl. sites	P-value	no. incl. sites	P-value
18S rRNA					
all	148	901	0.00000000	901	0.00000000
all - Diptera	142	864	0.00000000	864	0.00000000
all - Diplura	142	859	0.00023869	859	0.00023869
all - Diptera, Diplura	136	813	0.98948593	813	0.98948593
loops	148	477	0.00000000	506	0.00000000
loops - Diptera	142	457	0.00000000	486	0.00000000
loops - Diplura	142	445	0.98223690	474	0.97572435
loops - Diptera, Diplura	136	418	0.99984320	o/o	o/o
stems	148	424	0.00013940	395	0.00017406
stems - Diptera	142	407	0.61606681	378	0.66350184
stems - Diplura	142	414	0.99999618	386	0.99999997
stems - Diptera, Diplura	136	395	1.00000000	o/o	o/o
28S rRNA					
all	148	1152	0.00000000	1152	0.00000000
all - Diptera	142	1111	0.00000000	1111	0.00000000
all - Diplura	142	1111	0.00000000	1111	0.00000000
all - Diptera, Diplura	136	1067	0.00005175	1067	0.00005175
all - Diptera, Diplura, *Per.* sp., *S. erythr.*	134	1045	0.10365597	1045	0.10365597
loops	148	515	0.00000000	567	0.00000000
loops - Diptera	142	497	0.00000032	546	0.00000000
loops - Diplura	142	489	0.29727210	541	0.06842099
loops - Diptera, Diplura	136	o/o	o/o	o/o	o/o
loops - Diptera, Diplura, *Per.* sp., *S. erythr.*	134	o/o	o/o	o/o	o/o
stems	148	637	0.00000000	585	0.00000000
stems - Diptera	142	614	0.00000005	565	0.00058741
stems - Diplura	142	622	0.00000007	570	0.00298339
stems - Diptera, Diplura	136	596	0.99973324	547	0.99999936
stems - Diptera, Diplura, *Per.* sp., *S. erythr.*	134	o/o	o/o	o/o	o/o

3.3.5. Resulting topologies

Representatives of Symphyla and Pauropoda show unorthodox positions in all trees. Their placement is clearly incongruent with morphological evidence and results obtained from other genes. Symphyla are proposed as a sister group of all remaining arthropod clades. Pauropoda build one clade with Onychophora. Consequently, myriapods are polyphyletic in all inferred trees. Tab. 3.4 lists support

values of selected clades inferred from 'mixed model trees' and 'DNA trees' for both, non-stationary and stationary tree reconstructions.

Topologies inferred from the mixed model approach

Both trees (Figs. 3.5, 3.6) support monophyletic Chelicerata (Tab. 3.4) with sea spiders (Pycnogonida) as sister group to remaining chelicerates. Pycnogonida shows maximal support in both trees. Eucheliceratа are maximally supported in the stationary tree while this clade shows moderate support (pP 0.89) in the non-stationary tree. The horseshoe crab *Limulus polyphemus* is nested within arachnids. However, some internal relationships within Eucheliceratа show low support. Chilopoda are proposed as sister group of monophyletic millipedes (Diplopoda) in both analyses with high support. Within Diplopoda, the most ancient split occurs between Penicillata and Helminthomorpha. This clade – Myriapoda "partim" – is sister group of Chelicerata, thus giving support to the Myriochelata hypothesis (Mallatt et al., 2004), respectively Myriochelata "partim". Pancrustacea are maximally supported. Monophyly of Malacostraca and Branchiopoda is maximally supported in both trees while their position varies. Branchiopoda are suggested as a sister group of a clade (Copepoda + 'Hexapoda') in the stationary tree (Fig. 3.6). The cephalocarid *Hutchinsoniella* nests within hexapods and consequently 'smashes' Hexapoda. Among hexapods, monophyly is unambiguously supported for Protura, Diplura, Collembola, Archaeognatha, Odonata, Ephemeroptera, Phasmatodea, Mantophasmatodea, Mantodea, Plecoptera, Hemiptera, Coleoptera, Hymenoptera, Lepidoptera and Diptera. Nonoculata are maximally supported in all trees. Pterygota are inferred in both topologies, well supported in the non-stationary tree and with moderate support in stationary tree. Within winged insects, both trees suggest Odonata as sister group to a well supported Chiastomyaria (Ephemeroptera + Neoptera, Boudreaux, 1979; Kjer, 2004; Misof et al., 2007; Whitfield and Kjer, 2008). Blattodea are always paraphyletic with respect to the isopteran representative. This assemblage shows a sister group relationship with Mantodea, supporting monophyletic Blattopteroidea or Dictyoptera. The position of Dictyoptera among hemimetabolan insects varies. Dermaptera always cluster with Plecoptera. Hemiptera (Heteroptera,Homoptera) form a clade with remaining orthopterans + ((*Acheta*,Mantophasmatodea)Phasmatodea) with low support in both trees. Orthopteran insects are always polyphyletic due to the placement of *Acheta*. Within monophyletic Endopterygota (pP 1.0), Hymenoptera branch off first and are suggested as sister group of remaining endopterygote insects.

While the non-stationary and the stationary mixed model tree correspond in overall topologies, they differ in a number of remarkable details: 1) Hexapoda, Entognatha, Ectognatha and Dicondylia are only resolved in the non-stationary tree. 2) The cephalocarid *Hutchinsoniella* clusters among crustaceans as a sister group to Branchiopoda only in the non-stationary tree. This clade is proposed to be a sister group to (Copepoda,Hexapoda) although weakly supported. 3) The stationary tree show highly supported (Malacostraca,Ostracoda) as sister group of ((Mystacocarida,Pentastomida) + (Branchiura,Cirripedia)).

Table 3.4.: Bayesian support values for selected clades modeling non-stationarity and stationarity. pP: posterior probability; RNA/DNA: mixed model; DNA: DNA model. DNA trees are given in Fig. A.12.

Selected clades	**pP; non-stationary**		**pP; stationary**	
	RNA/DNA	DNA	RNA/DNA	DNA
Symphyla	1.0	1.0	1.0	1.0
(Pauropoda,Onychophora)	0.97	1.0	1.0	1.0
Pauropoda	1.0	1.0	1.0	1.0
Onychophora	1.0	1.0	1.0	1.0
Chelicerata	0.91	0.99	1.0	1.0
Pycnogonida	1.0	1.0	1.0	1.0
Euchelicerata (excl. Pycnogonida)	0.89	1.0	1.0	1.0
Myriapoda "partim": (Diplopoda,Chilopoda) (excl. Symphyla, Pauropoda)	0.97	1.0	0.98	0.94
Diplopoda	0.99	1.0	1.0	0.94
Chilopoda	1.0	1.0	1.0	1.0
Myriochelata "partim": ((Diplopoda,Chilopoda)(Euchelicerata,Pycnogonida))	0.97	0.97	1.0	0.98
(Myriochelata "partim",Pancrustacea)	0.95	0.85	0.98	0.89
Pancrustacea	1.0	0.99	1.0	1.0
(*Derocheilocaris*,Ostracoda)	0.62	0.57	-	-
(*Derocheilocaris*,*Raillietiella*)	-	-	0.75	-
(*Speleonectes*,*Hutchinsoniella*)	-	0.57	-	0.93
(Ostracoda,Malacostraca)	-	-	0.99	1.0
Malacostraca	1.0	1.0	1.0	1.0
(*Raillietiella*((*Hutchinsoniella*,Branchiopoda)(Copepoda,Hexapoda)))	0.60	-	-	-
((*Hutchinsoniella*,Branchiopoda)(Copepoda,Hexapoda))	0.65	-	-	-
(*Hutchinsoniella*,Branchiopoda)	0.65	-	-	-
(*Raillietiella*(Branchiopoda,(Copepoda,Hexapoda)))	-	0.38	-	0.93
Branchiopoda	1.0	1.0	1.0	1.0
(Copepoda,Hexapoda)	0.67	0.61	-	0.58
((Copepoda((*Lepisma*,*Hutchinsoniella*)(remaining hexapod taxa)))	-	-	0.70	-
((*Lepisma*,*Hutchinsoniella*)(remaining hexapod taxa))	-	-	0.58	-
Hexapoda	0.96	0.96	-	0.94
Entognatha: ((Protura,Diplura)(Collembola))	0.98	1.0	-	1.0
Nonoculata: (Protura,Diplura)	0.98	1.0	1.0	1.0
((*Lepisma*,*Hutchinsoniella*)(Protura,Diplura))	-	-	0.72	-
(*Lepisma*,*Hutchinsoniella*)	-	-	0.72	-
Protura	1.0	1.0	1.0	1.0
Diplura	1.0	1.0	1.0	1.0
Collembola	1.0	1.0	1.0	1.0
Ectognatha: (Archaeognatha(Zygentoma,Pterygota)	1.0	1.0	-	1.0
(Archaeognatha(*Ctenolepisma*,Pterygota))	-	-	1.0	-
Archaeognatha	1.0	1.0	1.0	1.0
Zygentoma	0.98	1.0	-	1.0
Dicondylia: (Zygentoma,Pterygota)	0.99	0.98	-	0.99
(*Ctenolepisma*,Pterygota)	-	-	0.99	-
Pterygota	0.97	0.95	0.94	0.96
Chiastomyaria: (Ephemeroptera,Neoptera)	0.96	0.98	0.97	0.99
Neoptera	0.98	1.0	1.0	1.0
Hemiptera	1.0	1.0	1.0	1.0
((*Acheta*,Mantophasmatodea)(Phasmatodea))	0.82	0.99	1.0	0.97
(*Acheta*,Mantophasmatodea)	0.81	0.93	0.99	0.93
Phasmatodea	1.0	1.0	1.0	1.0
Mantophasmatodea	1.0	1.0	1.0	1.0
'Orthoptera' (excl: *Acheta*)	0.99	0.98	1.0	1.0
((Dermaptera,Plecoptera)(Dictyoptera))	0.42	0.94	-	0.83
Dictyoptera	1.0	1.0	1.0	1.0
((Mantodea(*Blattella*,*Gromphadorhina*))(*Ectobius*,Isoptera))	1.0	1.0	1.0	1.0
(Mantodea(*Blattella*,*Gromphadorhina*))	0.53	0.73	0.55	0.63
(*Ectobius*,Isoptera)	0.89	0.83	0.94	0.95
Mantodea	1.0	1.0	1.0	1.0
Blattaria	-	-	-	-
(((Dermaptera,Plecoptera)(Dictyoptera))(Endopterygota))	0.39	0.88	-	0.84
((Dermaptera,Plecoptera)(Endopterygota))	-	-	0.38	-
(Dermaptera,Plecoptera)	1.0	1.0	1.0	1.0
Endopterygota	1.0	1.0	1.0	1.0
(Hymenoptera, remaining endopterygote insects)	1.0	-	1.0	-
Hymenoptera	0.80	-	0.90	-
((Lepidoptera,Trichoptera)Diptera)	-	-	0.90	-
(Lepidoptera,Trichoptera)	1.0	1.0	1.0	1.0

In contrast, in the non-stationary tree Malacostraca are placed more terminal and are sister group of (Pentastomida((Cephalocarida,Branchiopoda) + (Copepoda,Hexapoda))). However, the position of Pentastomida is low supported. 4) In the stationary tree, *Hutchinsoniella* is a sister taxon to *Lepisma* with low support (pP 0.72). This clade is nested within remaining hexapods (Fig. 3.6). Thus, Hexapoda are resolved only in the non-stationary tree with high support (pP 0.96, Fig. 3.5). Here, Copepoda are sister group to Hexapoda, but with low support (pP 0.69). 5) Likewise, Copepoda are sister group of ((*Lepisma,Hutchinsoniella*) + remaining hexapods) in the time-homogeneous tree (Fig. 3.6). Again, this sister group relationship shows low support. 6) Consequently, Entognatha (pP 0.98), Ectognatha (pP 1.0) and Dicondylia (pP 0.99) are monophyletic only in the time-heterogeneous tree. Primary wingless hexapods are paraphyletic as was expected: Archaeognatha are inferred as sister group to Dicondylia. 7) Within pterygote insects, (Dermaptera,Plecoptera) are sister group of Dictyoptera in the non-stationary tree. In contrast, they are placed as a sister group of Endopterygota in the stationary tree. However, both scenarios are negligibly supported.

Non-stationary DNA versus non-stationary mixed model topology

Since the non-stationary DNA model tree is strongly favored over the stationary tree in a BFT ($2lnB_{10} = 1544.12$), only topological differences between the non-stationary mixed model and the non-stationary DNA tree are summarized (Tab. 3.4). In many instances, the non-stationary DNA topology resembles the non-stationary mixed model tree. Major differences mainly concern relationships between crustaceans orders and the placement of hymenopterans. Minor differences occur within Protura, Collembola, Odonata and Ephemeroptera. Topological differences between the non-stationary mixed model and the DNA tree are visualized in the consensus network (Fig. 3.7). Black lines indicate contradictory relationships that are only present in the non-stationary DNA tree.

Topological differences within crustaceans

Most obvious are differences within crustaceans (Fig. 3.7). The cephalocarid *Hutchinsoniella* is suggested as a sister group to Branchiopoda in the mixed model tree (Fig. 3.5). In contrast, *Hutchinsoniella* is placed as sister group to the remipede *Speleonectes* in the DNA tree. Both taxa show long branches. This clade (*Hutchinsoniella,Speleonectes*) is inferred as sister group to (Branchiura,Cirripedia). In contrast, the mixed model tree suggests *Speleonectes* as sister group to (Branchiura,Cirripedia). (*Speleonectes* (Branchiura,Cirripedia)) shows a sister group relationship with (Mystacocarida,Ostracoda). Malacostraca are sister group to (Mystacocarida,Ostracoda) in the DNA tree (Fig. 3.5). The placement of *Raillietiella* is similar in both trees. All scenarios are negligibly supported (Tab. 3.4).

Topological differences within Protura and Collembola

Within proturans, *Sinentomon* (Sinentomidae) is nested within Eosentomidae with weak support in the non-stationary DNA tree. Thus, Eosentomidae are paraphyletic. The position of *Sinentomon* as a sister group of Acerentomidae also shows weak support in the mixed model tree while Eosentomidae are monophyletic with high support.

Within Collembola, *Gomphiocephalus*, a member of the family Hypogastruridae, is a sister group to (*Podura*,Neanuridae) in the DNA tree. This contradicts the mixed model topology: here *Podura* and *Gomphiocephalus* are sister taxa. Both scenarios are weakly supported. In both trees, Hypogastruridae are polyphyletic.

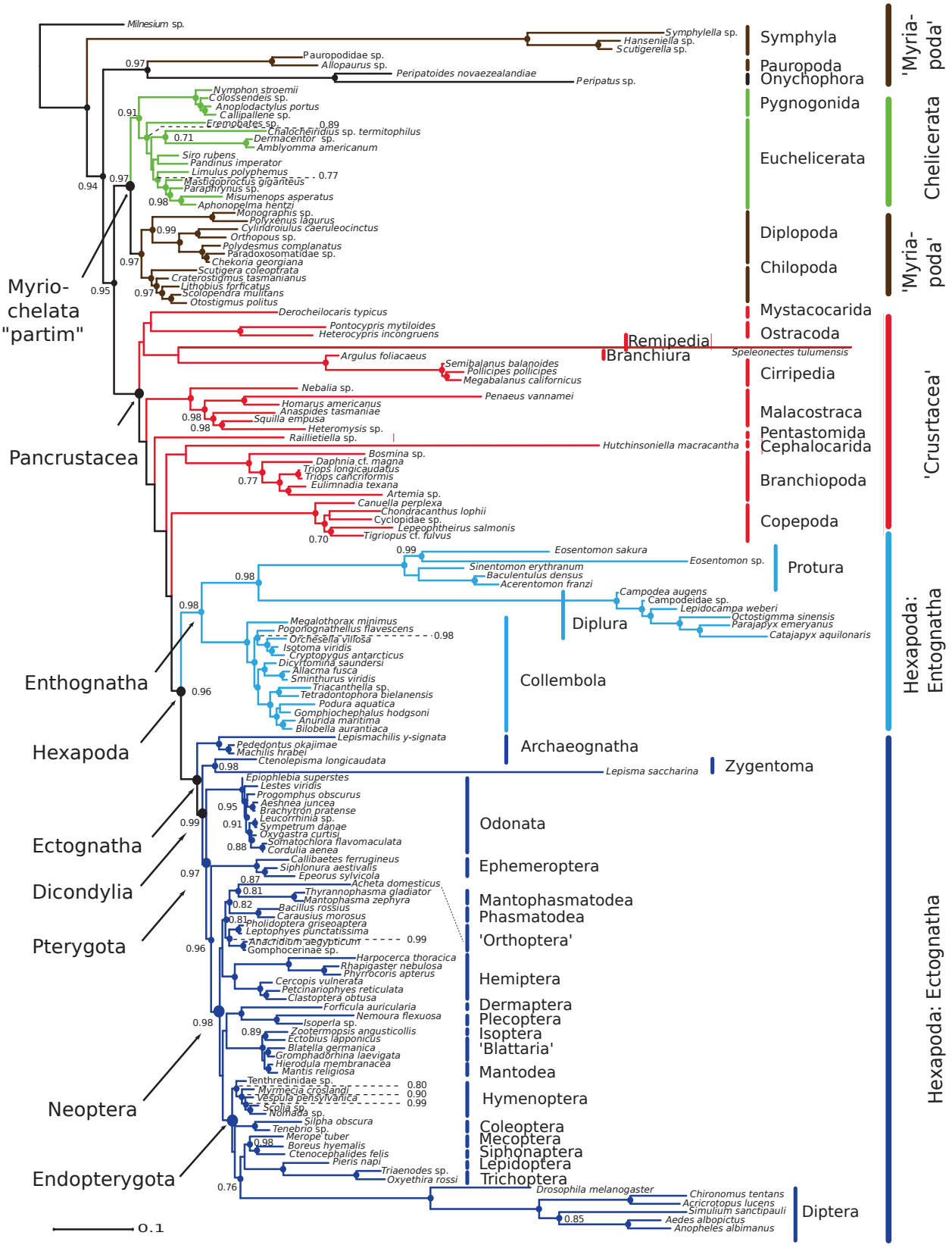

Figure 3.5.: Mixed model, non-stationary (time-heterogeneous) consensus tree deduced from 56,000 sampled trees inferred with *PHASE-2.0*. Support values below 0.70: nodes without dots, values not shown; pP = 1.0: represented by a dot only. Quotation marks indicate that monophyly is not supported. Color code is specified in Fig. 3.3.

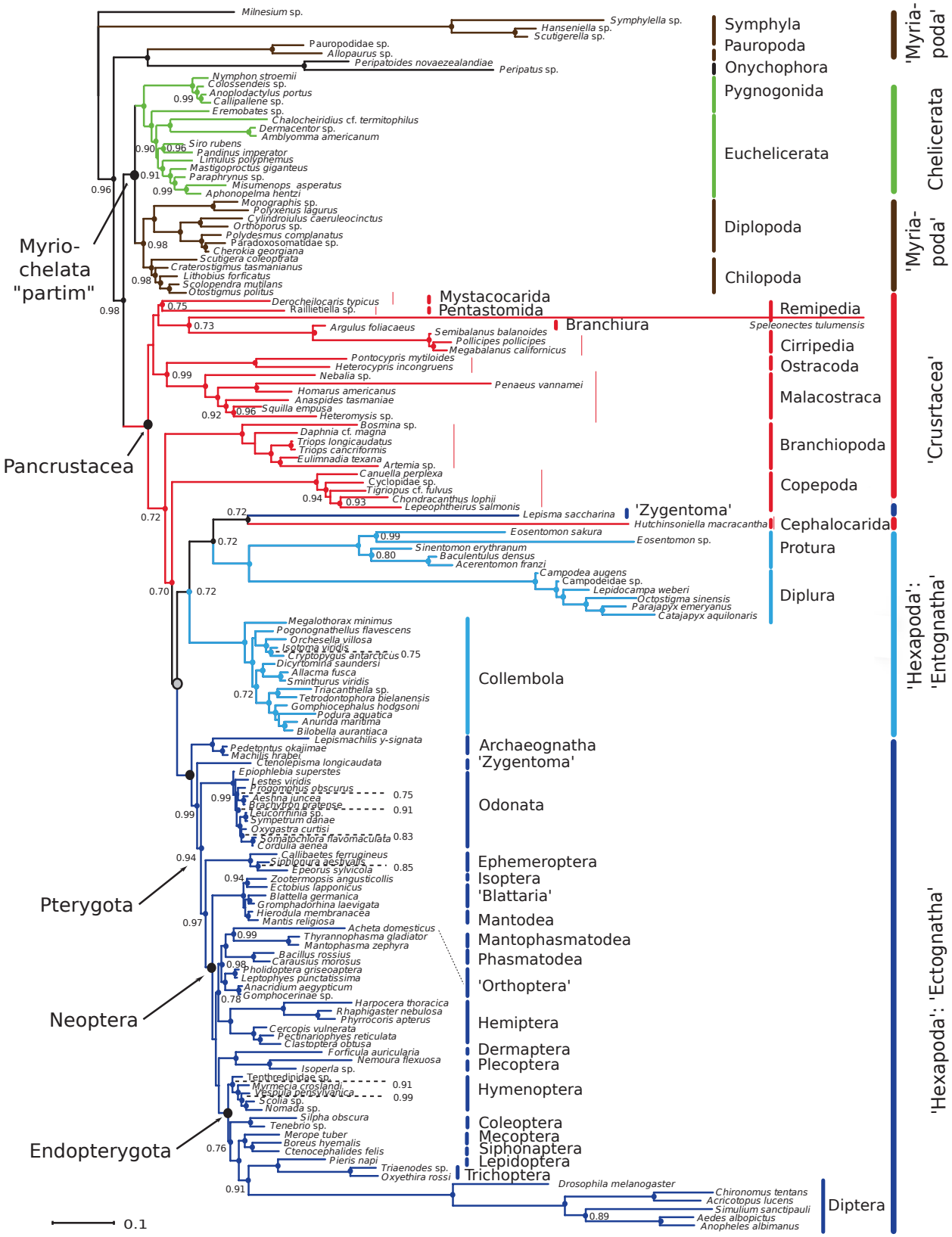

Figure 3.6.: Mixed model, stationary (time-homogeneous) consensus tree deduced from 18,000 sampled trees inferred with *PHASE-2.0*. Support values below 0.70: nodes without dots, values not shown; pP = 1.0: represented by a dot only. Gray dot: clade containing all hexapods including *Hutchinsoniella* (Cephalocarida, Crustacea) + *Lepisma* (Hexapoda, Zygentoma); pP 0.58. Quotation marks indicate that monophyly is not supported. Color code is specified in Fig. 3.3.

Topological differences within Odonata and Ephemeroptera

Within odonates, the damselfly *Lestes* branches off as most basal split (pP 1.0); the anisozygopteran *Epiophlebia* branched off as second split (pP 0.93) in the DNA topology while it is vice versa (pP. 1.0, pP 0.95) in the mixed model tree.

The mayfly *Eporus* is suggested as sister group to (*Callibaetes,Siphlonura*). Latter clade is negligibly supported (pP 0.56) in the DNA tree. In contrast, *Callibaetes* branches off first in the mixed model tree and is sister group of (*Eporus,Siphlonura*). Here, all splits show maximal support.

Figure 3.7.: Consensus network inferred from both non-stationary topologies (DNA and RNA/DNA tree). The consensus network was calculated with SplitsTree 4.10 (Huson and Bryant, 2006) (conflict threshold: 0.01). Black lines indicate contradictory relationships that are only present in the DNA tree. Taxa printed in black indicate a different placement compared with the RNA/DNA tree. Green lines indicate a similar topology in the DNA and RNA/DNA tree.

Topological differences considering hymenopteran insects

Hymenopterans are polyphyletic in the DNA topology: Tenthredinidae sp. branches off first within Endopterygota. Coleoptera are nested between Tenthredinidae sp. and remaining hymenopterans. This 'mixed' clade is negligibly supported (pP 0.33). The sister group relationship of Coleoptera + remaining hymenopterans also shows negligible support (pP 0.3). In contrast, Hymenoptera are monophyletic and branch off as most basal split within Endopterygota in the mixed model tree. Coleoptera are sister group to remaining endopterygote insects. Here, all splits show maximal support.

Branch lengths and support values: non-stationary DNA versus RNA/DNA tree

No remarkable differences are recognized comparing branch lengths of both trees. Branch lengths slightly differ in the 3rd or 4th decimal place, but this is not apparent 'by eye'. There is no tendency to shorter or longer branch lengths perceptible comparing both trees. With respect to node support, there is a slight tendency towards an increase of support values in the DNA topology for several clades (e.g. Chelicerata, Euchelicerata, Entognatha or Nonoculata). The most drastic support increase (Tab. 3.4) occurs for following clades: ((Dermaptera,Plecoptera)(Dictyoptera)), (((Dermaptera,Plecoptera)(Dictyoptera))(Endopterygota)), and (Mantodea(Blattella,Gromphadorhina)). Support values slightly decrease for e.g. Pancrustacea, (Copepoda,Hexapoda), Dicondylia or Pterygota (Tab. 3.4). Most strong decreases occur, for example, for (Myriochelata "partim",Pancrustacea) and for (Copepoda,Hexapoda).

3.4. Discussion

Nuclear rRNA genes show the densest taxon coverage of characterized sequences. Recent studies including both nuclear rRNA genes have shown an exhaustive taxon sampling, but none of them address arthropods in particular (Paps et al., 2009; Mallatt et al., 2010). Studies addressing arthropod relationship often use only a single nuclear rRNA marker or fragments of the 18S and 28S (Edgecombe and Giribet, 2002; D'Haese, 2002b; Yamaguchi and Endo, 2003; Giribet et al., 2004; Kjer, 2004; Luan et al., 2005; Kjer et al., 2006; Misof et al., 2007; Dell'Ampio et al., 2009), but in contrast see Mallatt et al. (2004), Gai et al. (2006), Mallatt and Giribet (2006), and Gao et al. (2008). The reliability of phylogenetic reconstructions based on rRNA markers is unclear (see e.g. Gillespie et al., 2005a; Misof et al., 2007; Jordal et al., 2008). A major issue is modeling of covariation patterns in sequence alignments (e.g. Buckley et al., 2000; Misof et al., 2006; Letsch et al., 2010) and in tree inference (e.g. Telford et al., 2005; Erpenbeck et al., 2007; Kjer and Honeycutt, 2007; Voigt et al., 2008; Letsch et al., 2010; Mallatt et al., 2010). Another important point is the impact of modeling non-stationarity among evolutionary rates. Covariation patterns and non-stationarity quickly lead to an erosion of phylogenetic signal and therefore, should be considered in modeling data sets (Foster, 2004; Simon et al., 2006; Misof et al., 2007; Kjer and Honeycutt, 2007; Gowri-Shankar and Rattray, 2007; Mallatt et al., 2010). Currently, there are no studies published on rRNA genes, except from this[3], which implement both, mixed RNA/DNA models and modeling non-stationarity.

3.4.1. Methodological Aspects

Alignment strategies and alignment masking

Present analyses included automated structure guided alignment procedures and automated alignment masking. Structure guided alignments have been advocated to improve the accuracy of rRNA alignments (Kjer, 1995, 2004; Misof et al., 2006, 2007; Dell'Ampio et al., 2009). For these purposes automated alignment software is available, for example RNAsalsa (Stocsits et al., 2009), which was applied in the present study, MXSCARNA (Tabei et al., 2008) or MAFFT (*Q-INSI, X-INSI,* Katoh and Toh, 2008). Structure guided alignments have mostly been constructed manually (Kjer and Honeycutt, 2007; Dohrmann et al., 2008; Mallatt et al., 2010). Currently, automated tools are rarely used in phylogenetic studies on real data sets. Yet, most software cannot handle large sequences ($>$ 1500 bp), except from RNAsalsa. Thus, for the present data set only RNAsalsa was applicable. The suitability of a structure guided alignment strategy for the current data set has been recently corroborated by simulation studies and real data of Letsch et al. (2010) where RNAsalsa performed best.

Phylogenetic signal in sequence data can get noisy due to multiple substitution processes (saturation) and erroneous homology hypotheses caused by ambiguous sequence alignments. These potentially bias tree reconstruction (Philippe et al., 2005a; Susko et al., 2005; Rodríguez-Ezpeleta et al., 2007; Wägele and Mayer, 2007). In general, studies predominantly include a manual alignment check

[3] A part of present analyses has been published in von Reumont et al. (2009): von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits R, Luan YX, Wägele JW, Pass G, Hadrys H, Misof B (2009): Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. BMC Evolutionary Biology (2009) 9:119.

for untrustworthy regions (Friedrich and Tautz, 1995; Yamaguchi and Endo, 2003; Kjer, 2004; Carapelli et al., 2005, 2007; Luan et al., 2005; Gai et al., 2006; Kjer et al., 2006; Mallatt and Giribet, 2006; Misof et al., 2007; Dohrmann et al., 2008; Mallatt et al., 2010). Since 2007, studies addressing arthropod relationships have used automated tools (Podsiadlowski et al., 2007; Aleshin et al., 2009; Comandi et al., 2009; Letsch et al., 2009; Simon et al., 2009). In the present study, alignments were masked with Aliscore (Misof and Misof, 2009; Kück et al., 2010) and ALICUT (Kück, 2009). Compared with the software Gblocks (Castresana, 2000), Aliscore is not dependent on the specification of an arbitrary threshold (Misof and Misof, 2009). Masking alignments increased signal of the present data sets (see chapter 2 and this chapter, section 3.3, Fig. 3.3). Therefore, alignment masking is strongly recommended for all kinds of molecular studies prior to tree reconstruction. However, parameter settings should be cautiously chosen, depending on selected genes and taxon composition.

Implementing mixed models and non-stationary processes

One aim of present analyses was to implement background knowledge by mixed models due to a biologically realistic consideration. For the same reason, non-stationarity was modeled.

Paired sites (stems) do not evolve independently (Wheeler and Honeycutt, 1988; Dixon and Hillis, 1993; Swofford et al., 1996; Jow et al., 2002; Hudelot et al., 2003; Felsenstein, 2004; Galtier, 2004). Treating stem regions as independently evolving in phylogenetic reconstructions is a clear simplification of the real situation. In theory, this leads to biased support (Jow et al., 2002; Hudelot et al., 2003). In less supported nodes, where signal is at the edge of resolution, negligence theoretically might turn the balance between two competing hypotheses. In the present analyses, however, mixed models statistically perform less well than DNA models. This is not in line with other studies on real data: it is suggested that mixed models are superior to standard DNA models in stem regions (e.g. Kjer, 2004; Telford et al., 2005; Erpenbeck et al., 2007; Dohrmann et al., 2008; Ware et al., 2008). The outperformance of the non-stationary DNA model over the non-stationary mixed model for the present data set is astonishing, but possibly correlated to parameter-richness of the selected RNA model within the mixed model approach (see below).

DNA-corresponding 16–state RNA models (Gowri-Shankar and Jow, 2006) have been selected within the present mixed model approaches. 16–state models might be problematic because it is difficult to fit these models to real data due to parameter-richness. Dohrmann et al. (2008) compared 6–state and 7–state RNA models. They observed that parameter-rich models might have been overparameterized and thus led to biased likelihood scores. A Bayes Factor Test clearly favored less complex RNA models. Nevertheless, RNA models outperformed DNA models in stem regions (Dohrmann et al., 2008; Letsch et al., 2010). Even the 'best choice' of a consensus secondary structure can only capture predominantly conserved structural features among sequences. This implies that applied RNA models must be able to cope with mismatches in base-pairing. Less complex RNA models (6– and 7–state) completely ignore mismatches or pool these into a single character state which produces artificial synapomorphies. According to Schöniger and von Haeseler (1994), it is more likely that covariation is a multiple-step process. It allows an intermediate existence of instable (non Watson-Crick) pairs. Intermediate states are only described in 16–state RNA models. In the present study, however, the DNA model outperforms the mixed model while non-stationarity is considered (see section 3.3.4). This is not in line with Dohrmann et al. (2008) and Letsch et al. (2010). Present results might be not

directly comparable with these studies, since Dohrmann et al. (2008) and Letsch et al. (2010) ignore non-stationarity. Nevertheless, the non-stationary DNA topology suggests some suspicious clades with negligible support, for example polyphyletic hymenopterans (section 3.3.4). If 16–state mixed models indeed lead to biased likelihood scores due to parameter-richness or over-parametrization, a poorer performance might be a consequent result in the present study. Letsch (pers. comm.) stated further, that mixed models outperform DNA models if loops are not saturated and perform less well than DNA models if loop regions are saturated. Therefore, it would be worthwhile to test loop regions of the present data for saturation.

Shifts in base composition of rRNA genes have been observed in Diptera, Diplura, Protura and Symphyla (Friedrich and Tautz, 1997; Luan et al., 2005; Gai et al., 2006; Mallatt and Giribet, 2006; Misof et al., 2007; Dell'Ampio et al., 2009). This is also true for the present data set (section 3.3.3) and has been considered by modeling non-stationarity. Ignoring base compositional heterogeneity within the data can mislead phylogenetic reconstructions (Foster, 2004; Jermiin et al., 2004; Blanquart and Lartillot, 2006; Gowri-Shankar and Rattray, 2007). Still, the effect of modeling of non-stationarity on real data is relatively unexplored (Blanquart and Lartillot, 2006; Gowri-Shankar and Rattray, 2007). Several studies discuss non-stationary processes as possible explanations for misplacements of particular taxa (Mallatt et al., 2004; Luan et al., 2005; Gai et al., 2006; Dell'Ampio et al., 2009; Hassanin et al., 2005). Nevertheless, these studies lack modeling of non-stationarity. Instead, LogDet methods have been applied to compensate for variations of base frequencies (Mallatt et al., 2004; Luan et al., 2005; Mallatt and Giribet, 2006), which leads to some independence of non-stationarity but, among site rate variation (ASRV) cannot be handled with LogDet methods.

In present rRNA analyses, using traditional measures of goodness-of-fit tests, non-stationary approaches clearly outperform stationary approaches. Therefore, they should be preferred in tree reconstructions (see section 3.3.4). Furthermore, only one chain passed defined criteria and was finally included for the stationary DNA model tree. It seems likely that this chain has been trapped into a local optimum during tree search which can lead to biased tree inference. Additionally, the non-stationary mixed model topology particularly suggests more 'plausible' clades which can, of course, only serve as ample argument (see section 3.4.2).

The present mixed model and DNA trees show no perceptible differences between branch lengths (see section 3.3.5). This contradicts findings of Schöniger and von Haeseler (1994, 1995) and simulation studies of Letsch et al. (2010). Both studies propose shorter branch lengths when mixed models are applied. Therefore it is worthwhile, to examine real data sets. Present branch length estimations have been separately conducted with a fixed topology (see section 3.2.7) which might be due to similar branch lengths.

Mallatt et al. (2010) and Letsch et al. (2010) reported that support values increase when only DNA models are applied on rRNA genes. Already Tillier and Collins (1998) observed reduced statistical confidence for mixed model trees: if paired-sites models are applied, there are 50% less characters with respect to stem regions than in DNA models, where paired sites are treated independently (Jow et al., 2002; Mallatt et al., 2010). Jow et al. (2002) and Telford et al. (2005) argue for the superiority of mixed models due to reduced support for biologically implausible clades. In present analyses, Chelicerata, Euchelicerata, Entognatha or Nonoculata show slightly increased support in the non-stationary DNA tree (Tab. 3.4). A drastic increase of support is obtained for the clade

((Dermaptera,Plecoptera)(Dictyoptera)) (non-stationary RNA/DNA tree: pP 0.42; non-stationary DNA tree: pP 0.94). It is unclear, whether this is only due to the enlarged number of effective sites in the DNA model approach. A decrease of support, e.g. for Myriochelata "partim" + Pancrustacea (non-stationary RNA/DNA tree: pP 0.95; non-stationary DNA tree: pP 0.85) cannot be explained by an increased number of effective sites in the DNA model approach. Assuming that unpaired regions show more signal for this clade than paired regions, could explain a support decrease: unpaired regions which show more signal for a clade might have less weight than paired regions in a DNA model approach (since they are counted twice) with low or no support for the respective clade. Assuming that support varies between unpaired and paired regions for a clade, and support in paired regions is low and in unpaired regions moderate: if then DNA models are applied, stem regions are counted twice due to independent site modeling, thus more sites show low support and might have more weight than loop sites showing high support. Taken support of all sites together, a general shift from more support towards less support might be possible. This assumption certainly requires much more detailed analyses or simulation studies. If this scenario can be corroborated, again it gives more arguments that application of mixed models infer more 'realistic' support in tree reconstructions.

Impact of mixed models and non-stationarity on taxa with long branches

Hendy and Penny (1989) and Zwickl and Hillis (2002) suggested breaking long branches by adding more taxa. Heath et al. (2008a) recently stated: "Thorough taxon sampling is thus one of the most practical ways to improve the accuracy of phylogenetic estimates, as well as the accuracy of biological inferences that are based on these phylogenetic trees". Nevertheless, data accumulation (adding taxa) is not a panacea to cure biasing effects due to long branch taxa (Aguinaldo et al., 1997; Brinkmann et al., 2005; Lartillot et al., 2007; Brinkmann and Philippe, 2008). Excluding taxa showing long branches from data sets has been suggested by Brinkmann and Philippe (2008) but, this is no solution when phylogenetic relationships of these species are addressed.

For the present data set, an exhaustive taxon sampling and cautious modeling obviously reduce biasing effects. Suspicious clades are present in the stationary mixed model tree, e.g. (*Lepisma*,*Hutchinsoniella*). This clade is not present anymore by modeling non-stationarity (Fig. 3.5 and section 3.4.2). *Speleonectes* (Remipedia, Crustacea), *Hutchinsoniella* (Cephalocarida, Crustacea) and *Lepisma* (Zygentoma, Hexapoda) show very long branches. A clade (*Speleonectes*,*Hutchinsoniella*) is present in both DNA trees (Tab. 3.4). Still, phylogenetic relationships of remipedes and cephalocarids are hard to address. Currently, there is not enough known from both groups and data availability is sparse. A clade (*Speleonectes*,*Hutchinsoniella*) has been proposed in previous studies (Shultz and Regier, 2000; Regier et al., 2005, 2008; Koenemann et al., 2010; Regier et al., 2010). Koenemann et al. (2010) argue that a sister group relationship of Remipedia and Cepahlocarida is based on artifacts. Regier et al. (2010) suggest a sister group relationship of Remipedia and Cephalocarida ("Miracrustacea = surprising crustaceans"). They propose that Miracrustacea are sister group of Hexapoda. Data on remipedian larvae contradict this scenario (Koenemann et al., 2007, 2009). If this clade is a result of a biasing effect in present DNA trees, modeling site-interdependence helps to impede bias with respect to both species (Fig. 3.5). The most promising approach in the present study is modeling site-interdependence plus non-stationarity. Applying mixed models alone does not improve phylogenetic reconstruction with respect to the placement of *Lepisma* and *Hutchinsoniella*. The effect of modeling non-stationarity is apparent in the mixed model tree. Deep hexapod splits are disrupted (hexapods,

ectognaths, dicondylian insects) and most hexapod clades show low support if non-stationarity is not considered (Fig. 3.6). In contrast, accepted hexapod clades are inferred with high support modeling non-stationarity.

In future rRNA analyses, paired positions, previously identified as randomly similar and excluded from the present data set, should be retained, because Aliscore currently does not consider base pairings. For tree inference, there is a high demand to work on methods that accurately identify model misspecification which can lead to biased tree inference, especially, when proportions of variable sites change across the tree (see Grievink et al., 2010). Current model testing software (e.g. jModeltest Posada, 2008) does not consider mixed models or non-stationarity; furthermore, they are also not applicable to large data sets (current software crashed on the present rRNA data set, see chapter 2, section 2.4.1). An accurate model estimation also might avoid over-parametrization of the particular data set. Branch length estimations should be refined by new or improved algorithms which enable branch length estimation as accurate possible during tree inference, and not separated from a tree search. Kolaczkowski and Thornton (2009) argue that Bayesian reconstructions might be heavily biased in favor of topologies that group taxa showing long branches together. Therefore, it is worthwhile to infer phylogenetic trees from the present data set in a ML framework taking mixed models and non-stationarity into account. Recently, mixed models have been implemented in RAxML 7.1.0 (Exelixis Lab, A. Stamatakis 2009) but currently, only stems can be partitioned. Since there are differences in evolutionary rates between 18S and 28S, partitioning of loops is necessary for a comparative analysis of the present data set within a ML framework. It might be worthwhile to test loop regions for saturation, because mixed models perform less well compared to DNA models if loops are saturated (H. Letsch, unpublished data, pers. comm).

3.4.2. Conflicting phylogenetic hypotheses: Modeling non-stationarity versus stationarity with mixed models

The comparison of the non-stationary and the stationary approach indicate inconsistencies between analyses that may be explained by the adding of non-stationary processes during evolution of rRNA genes.

Cephalocarida and Remipedia

A sister group relationship of *Hutchinsoniella* (Cephalocarida) and *Lepisma* (Zygentoma, Hexapoda) is suggested by the stationary, mixed model tree. Except from low support, this clade lacks any evidence from morphological, developmental and other molecular studies. This is also true for the sister group relationship of (*Hutchinsoniella*,*Zygentoma*) and Nonoculata (Fig. 3.6 and section 3.4.1). The placement of *Hutchinsoniella* and *Lepisma* consequently leads to poly- or paraphyly of several major groups (Hexapoda, Entognatha, Ectognatha, Dicondylia). This clade might be an artifact and a result of ignoring non-stationarity that might have lead to an attraction of both taxa due to signal erosion (Wägele and Mayer, 2007). In contrast, Cephalocarida are sister group to Branchiopoda in the non-stationary, mixed model tree. Although this clade receives low support, it is congruent with morphological data, fossil data, neuroanatomical studies, molecular and combined analyses (Walossek, 1993, 1999; Giribet et al., 2001; Fanenbruck, 2003; Wheeler et al., 2004). Most recent molecular studies on rRNA data lack Cephalocarida (e.g. Mallatt et al., 2004; Mallatt and Giribet, 2006; Mallatt

et al., 2010). Several studies propose a sister group relationship of Remipedia and Cephalocarida (likewise represented by *Hutchinsoniella*) with moderate or high support (Regier et al., 2005, 2008, 2010; Koenemann et al., 2010). This clade has also been proposed as an artifact in recent studies of Koenemann et al. (2010) and Jenner (2010). However, it is also suggested in present non-stationary and stationary DNA trees (Tab. 3.4) but, the support in the non-stationary DNA topology is low. Since the non-stationary DNA tree is assumed to reflect a more realistic picture and a sister group relationship of Remipedia and Cephalocarida is not supported by both RNA trees, there is some doubt on the reliability of resolution of this clade in the stationary DNA tree. From the present trees, a plausible conclusion cannot be drawn addressing phylogenetic positions of remipedes and cephalocarids.

Hexapoda, Ectognatha and Entognatha

Monophyly of Hexapoda, Entognatha, Ectognatha, Dicondylia including *Lepisma* was supported by all non-stationary trees. Likewise, an explanation is given for the misplacement of *Lepisma* in the stationary approach, which cannot be accomplished by alternatively excluding the taxon. Entognatha with Nonoculata as sister group to Collembola and Entognatha as sister group to Ectognatha is supported (Fig. 3.5). This result is in line with multiple molecular studies (Kjer, 2004; Luan et al., 2005; Mallatt and Giribet, 2006; Gao et al., 2008; Dell'Ampio et al., 2009; Mallatt et al., 2010). The monophyly of hexapods (Dohle, 2001; Bitsch and Bitsch, 2004; Harzsch et al., 2005; Harzsch, 2006; Ungerer and Scholtz, 2008) has been challenged seriously by mitochondrial data: reciprocal paraphyly of hexapods has been affirmed, Collembola occurred outside from hexapods as sister group to other pancrustacean taxa (Nardi et al., 2003a; Carapelli et al., 2005, 2007). This would imply that features of the hexapod bauplan have evolved at least twice. Reanalyses of the same data set, however, contradict this scenario (Delsuc et al., 2003). In the present analyses, both non-stationary trees strongly corroborate Ectognatha (Tab. 3.4), which is unambiguously suggested by molecular and non-molecular data and has been widely accepted (Grimaldi, 2010).

Support for the monophyly of Entognatha (Hennig, 1953; Tuxen, 1959) has been generally considered as ambiguous. The polarization of morphological characters is difficult due to extreme adaptations to subterranean or cryptic life styles. The presence of many possible plesiomorphic characters, for example the presence of fully muscular antennae, abdominal appendages and anameric development in Protura, gives entognathous hexapods an important role in understanding evolution of hexapods (see chapter 1). The clade has been unambiguously supported by analyses of e.g. Kjer (2004) or Gao et al. (2008). Again, Entognatha is corroborated by the present non-stationary trees. In contrast, Dell'Ampio et al. (2009) showed low support for Entognatha. Inconsistencies are reported with respect to different methods of reconstruction (e.g. Mallatt and Giribet, 2006; Mallatt et al., 2010, ML analyses show weak support, Bayesian reconstructions show high support). Except from springtails, non-rRNA data are still sparse. A more thorough taxon sampling is desirable, especially for phylogenomic data. First phylogenomic analyses presented in this thesis (chapter 2) do not provide consistent support for Entognatha.

Internal relationships of Protura

Protura are classically divided into Acerentomata, Eosentomata and Sinentomata (Yin, 1996) due to three different types of *pseudoculi* (Yin et al., 2002). Eosentomata possess spiracles with a primitive

tracheal system, while members of Acerentomata lack these structures (see discussion in Luan et al., 2005; Gao et al., 2008). The position of Sinentomata (with Fujientomidae and Sinentomidae) is unclear. Sinentomata have been discussed as paraphyletic: Sinentomidae have a tracheal system which strongly differs from Eosentomata. Fujientomidae have no tracheal system at all. Spermatological studies show markedly different spermatozoans between proturan species which might reflect their long evolutionary course (Yin and Xue, 1993). Sinentomata have been discussed as 'intermediate' (Imadaté, 1966; Yin, 1996). Alternatively, they have been placed within Acerentomata (Tuxen, 1977). The placement of Sinentomata is also inconsistent between molecular studies. The only study to date that included species of Acerentomata, Eosentomata and Sinentomata covering both families has been published by Luan et al. (2005). The authors propose a paraphyly of Sinentomata: Fujientomidae branch off first and Sinentomidae are a sister group of (Acerentomata,Eosentomata) with weak support. In the data set of Gao et al. (2008) and in the present study, only Sinentomidae are included. Gao et al. (2008) suggest Sinentomidae as sister group to Acerentomata with high support. All studies, except from the present thesis, address compositional heterogeneity, but do not implement either mixed models or non-stationarity. In present analyses, both mixed model trees are in line with findings of Gao et al. (2008), but the support of Sinentomidae (*Octostigma*) as sister group of Acerentomata is low (pP 0.6). The position of Sinentomidae remains unclear. Although DNA trees are statistically preferred, the mixed model topologies seem from a taxonomic point of view more plausible: in DNA trees, *Octostigma* is nested within Eosentomidae as sister group to *Eosentomon sakura*, albeit weakly supported. This is considered as a suspicious clade because a paraphyly of Eosentomata seems unlikely and has never been supported by previous studies.

Relationships within Poduromorpha (Collembola)

Internal relationships of Collembola are still not resolved (see chapter 1). Especially, monophyletic Hypogastruidae have been contradicted (D'Haese, 2002b) du to the instability of some species, e.g. *Gomphiocephalus* or *Podura*. In the non-stationary mixed model and in the stationary DNA tree, *Gomphiocephalus* is sister group to (*Podura*,Neanuridae). In both remaining trees, *Podura* and *Gomphiocephalus* are proposed as sister taxa. Therefore, approaches of the present thesis do not allow plausible conclusions about the position of both poduromorphs *Gomphiocephalus* (Hypogastruridae) and *Podura* (Poduridae). In contrast, all trees propose a polyphyly of Hypogastruridae which has been suggested earlier (D'Haese, 2002b, 2003; Luan et al., 2005; Xiong et al., 2008; Dell'Ampio et al., 2009).

Endopterygote insects

Endopterygota are inferred in all present trees with high support (see below) but, placement and monophyly of Hymenoptera is inconsistent comparing mixed model and DNA topologies. In both mixed model trees, Hymenoptera are monophyletic with high support. They branch off first and are sister group of remaining endopterygotes. In contrast, within both DNA trees, Coleoptera are nested within hymenopterans with weak support. This leads to a polyphyly of hymenopteran insects. This can be argued as highly unlikely with respect to all morphological and molecular studies considering Hymenoptera and Coleoptera (e.g. Wiegmann et al., 2009). Assuming that hymenopterans are monophyletic, there is a preference towards the present mixed model approach. Ignoring non-stationarity does not affect the topology. Altogether, the effect of implementing mixed models and /

or non-stationarity is more or less present, depending on the considered taxon.

3.4.3. Clades not affected by non-stationary processes and mixed model approaches

Symphyla and Pauropoda

Currently, the monophyly of myriapods is the most recent working hypothesis (Shear and Edgecombe, 2010) and has been critically considered in Edgecombe (2010). Due to the placement of Symphyla and Pauropoda, myriapods are polyphyletic in all trees of the present study. Suggested placements of symphylans (sister group to euarthropods including onychophorans) and pauropods (sister taxon of onychophorans) that are highly supported in trees (Figs. 3.5, 3.6, Tab. 3.4) contradict any evidence from morphological, developmental, and partly from molecular studies. Therefore, these clades seems rather unlikely. Although one aim was to break long branches by dense taxon sampling, problems apparently persist for particular taxa. With respect to the position and the branch lengths of symphylans, pauropods and onychophorans, saturation by multiple substitution might cause signal erosion (Wägele and Mayer, 2007).

Mandibulata versus Myriochelata

Analyses of rRNA sequences have favored Myriochelata over Mandibulata (Friedrich and Tautz, 1995; Mallatt et al., 2004; Mallatt and Giribet, 2006). The problem of positioning myriapods has recently been addressed in Mallatt et al. (2010) and Edgecombe (2010). Present trees do not provide a final conclusion: the position of Pauropoda and Symphyla is unorthodox and myriapods are polyphyletic. The most recent study of Regier et al. (2010) should be treated with caution. The authors propose strongly supported Mandibulata based on a multi-gene analysis (non-ribosomal nuclear genes) and neglect putative phylogenetic signal from other data. Nevertheless, they point out that Myriochelata might be supported mainly by ribosomal genes. A reliable reconstruction of the position of myriapods within the Euarthropoda demands more cautioned analyses, for example considering 'gene groups' (e.g. ribosomal *versus* non-ribosomal genes).

Sister group of Hexapoda

Most molecular studies support paraphyly of crustaceans with respect to hexapods. A sister group relationship between Branchiopoda and Hexapoda was proposed for the first time by Regier and Shultz (1997). Shultz and Regier (2000) and Regier et al. (2005) have corroborated this relationship. Likewise, this clade is favored by rRNA-based studies (Mallatt et al., 2004; Mallatt and Giribet, 2006). Both studies, however, neglect that their trees show Cyclopidae (Copepoda) as a sister group of Hexapoda, albeit with weak support. Again, topologies of present analyses suggest Copepoda as a sister group, but the negligible support might indicate low signal or conflicts in signal. This clade, however, lacks any support from morphological studies and is also rejected by present phylogenomic results (chapter 2).

Nonoculata versus Ellipura

Ellipura (Börner, 1910; Hennig, 1981; Kristensen, 1991; Koch, 1997) are supported only by few molecular mt rRNA studies (e.g. Carapelli et al., 2000). Instead, mostly Nonoculata are supported

(Giribet et al., 2001, 2004; Kjer, 2004; Luan et al., 2004). All topologies of the present rRNA data show Nonoculata within a monophyletic Entognatha. This is congruent with recent studies (Kjer, 2004; Luan et al., 2005; Mallatt and Giribet, 2006; Gao et al., 2008; Dell'Ampio et al., 2009; Mallatt et al., 2010). Following Luan et al. (2005), Dell'Ampio et al. (2009) cautioned that Nonoculata may be artificial caused by a shared nucleotide bias and long branch attraction. This argument can be rejected by present analyses: Nonoculata again show high support modeling non-stationarity. Since both Protura and Diplura show long branches (Figs. 3.5, 3.6), artifacts may still be present. On the other hand, Nonoculata are corroborated with maximal support by phylogenomic EST analyses, earlier presented (chapter 2); here branch lengths are short. However, any non-molecular evidence for Nonoculata is ambiguous (Machida, 2006; Szucsich and Pass, 2008, see chapter 1). Possibly, advanced methods, for example spectra of split-supporting (*SAMS*, Wägele and Mayer, 2007) or tools to detect long branch artifacts, may corroborate current results.

Monophyletic Diplura and internal relationships

The monophyly of Diplura has been debated and a dipluran paraphyly has been proposed by several morphological, and developmental studies. Studies on sperm- and ovariole structures supported a possible paraphyly. Either Campodeoidea or Japygoidea were proposed to be closer related to Ectognatha, depending on respective genital character complexes (see chapter 1, Jamieson, 1987; Štys and Bilinski, 1990; Štys et al., 1993; Bilinski, 1994; Štys and Zrzavý, 1994; Jamieson et al., 1999). Koch (1997) argued for the monophyly of diplurans due to entognathous conditions (see chapter 1). In particular, Diplura have been supported by molecular studies (e.g. Giribet and Ribera, 2000; Giribet et al., 2004; Kjer, 2004; Luan et al., 2004, 2005; Mallatt and Giribet, 2006; Gao et al., 2008; Dell'Ampio et al., 2009; Mallatt et al., 2010; Regier et al., 2010). Few studies propose paraphyly (e.g. Shultz and Regier, 2000) or polyphyly (e.g. Carapelli et al., 2007, based on mt markers). Present rRNA topologies corroborate monophyletic Diplura with maximal support.

Even though Projapygoidea take up a key position, only a handful morphological and even less molecular studies included this taxon. Many morphological characters are considered as intermediate between Campodeoidea and Japygoidea (Rusek, 1982). In morphological studies, they group either with Campodeoidea or with Japygoidea (Rusek, 1982; Štys and Bilinski, 1990; Štys et al., 1993; Pagés, 1997; Bitsch and Bitsch, 2000, 2004). Rusek (1982) argued for a most basal split: he found more morphological variation within Projapygoidea than in Campodeoidea and Japygoidea. Therefore, he concluded that they must be older than both remaining suborders. Molecular studies which included Projapygoidea, found a sister group relationship of Projapygoidea and Japygoidea with high support (Luan et al., 2004, 2005; Gao et al., 2008). Trees of the present reconstructions show paraphyletic Campodeoidea with *Lepidocampa* as sister group of (Projapygoidea,Japygoidea), maximally supported in all instances. Present analyses unambiguously support the monophyly of Japygoidea and reject earlier results of e.g. Kjer (2004) and Dell'Ampio et al. (2009).

Addressing dipluran relationships, in further research it is worthwhile to include more molecular markers (e.g. from EST data) for all three suborders. ESTs are currently only available for *Campodea*, which have been generated for the present thesis (chapter 2).

Relationships between collembolan suborders

Contemporary, the classification of springtails into Poduromorpha, Entomobryomorpha, Symphypleona and Neelipleona (Deharveng, 2004) is accepted, whereas relationships between collembolan suborders are unclear. Gao et al. (2008) and Dell'Ampio et al. (2009) support the monophyly of all suborders, but earlier studies inferred paraphyletic Symphypleona (e.g. D'Haese, 2002b; Luan et al., 2005). Currently, Neelipleona is the smallest order with one single family, Neelidae (short-horned springtails). They have been described at the end of the 19th century (Folsom, 1896). Neelipleona have been separated from Symphypleona by Massoud (1971) as they strongly differ in their globular bodies. Originally, they were proposed as a sister group of Symphypleona (see discussion in D'Haese, 2003; Dell'Ampio et al., 2009). Gao et al. (2008) firstly included Neelipleona in an rRNA study, followed by Xiong et al. (2008) and Dell'Ampio et al. (2009). These studies provide inconsistent results: Xiong et al. (2008) suggest a sister group relationship of Neelipleona and Symphypleona, whereas Dell'Ampio et al. (2009) propose Neelipleona to be polytomous with Entomobryomorpha and (Poduromorpha,Symphypleona). Present topologies are in line with findings of Gao et al. (2008). Neelipleona (*Megalothorax*) always split off first with high support.

Monophyly of all suborders is corroborated in the present rRNA reconstructions with high support (pP $>$ 90), except from the stationary, mixed model tree. Here, Entomobryomorpha are only weakly supported. A sister group relationship of Entomobryomorpha and (Poduromorpha,Symphypleona) is well supported (pP $>$ 90), again except from the stationary mixed model tree (pP around 0.7). D'Haese (2003) also inferred a sister group relationship of Entomobryomorpha and (Poduromorpha,Symphypleona). Porco and Deharveng (2009) recently suggested that epicuticular chemical compounds provide good resolutions for terminal branches but not for internal collembolan splits. Increased data availability of Neelipleona (ESTs) and detailed analyses of slow and fast evolving genes might be helpful to resolve relationships within springtails.

Ancient splits within pterygote insects

Present rRNA data suggest Chiastomyaria as most likely (but see discussion in Misof et al., 2007). For all three possible arrangements of Odonata, Ephemeroptera and Neoptera likewise morphological support is present. This might be best explained by an early "explosive radiation" (Whitfield and Kjer, 2008). All present topologies are in line with the rRNA study of Letsch et al. (2010) and the phylogenomic study of Simon et al. (2009). These topologies are inconsistent with results of Mallatt et al. (2010) and Regier et al. (2010). The present phylogenomic approach (chapter 2) does not provide distinct results addressing early pterygote splits. A re-analysis of morphological mandible muscle structures with a thorough taxon sampling may be crucial: the Metapterygota hypothesis (Ephemeroptera(Odonata,Pterygota) (morphologically supported by Staniczek, 2000, 2003), relies on a very sparse odonate taxon sampling (Blanke, ZFMK, pers. comm.).

Endopterygote insects: position of Hymenoptera

The placement of Hymenoptera within Endopterygota is still unclear. Several scenarios have been suggested: for example, beetles + neuropteridans branch off first or hymenopterans are sister group to all other endopterygote insects (Kristensen, 1999; Kukalová-Peck and Lawrence, 2004; Beutel and Pohl, 2006; Wiegmann et al., 2009). Present analyses show strong support for most endopterygote

orders. Hymenoptera are proposed as sister group to remaining intricateness insects in present mixed model topologies with maximal support. In rRNA analyses of Letsch et al. (2010), this scenario is weakly supported, but only in their DNA and not in their mixed model topology. In contrast to present analyses, the authors used the software MrBayes (Ronquist and Huelsenbeck, 2003) with the DNA model GTR + Γ for inference. Topological differences might indicate that reconstruction of trees based on nuclear rRNA data, at least for subgroups, is highly dependent on reconstruction methods. Therefore, conclusions based on nuclear rRNA genes should be drawn with caution for Hymenoptera. This is also true for early pterygote splits (see above). If taxa underly a fast radiation or differentiation (e.g. Coleoptera or Hymenoptera), signal of selected genes might be low. If the placement of these taxa turns out to be highly dependent on the reconstruction method, it is worthwhile to model present data by incorporating background knowledge. Alternatively one should go for genes that provide more signal for tree reconstructions. A sister group relationship of Hymenoptera and remaining endopterygotes is suggested by Savard et al. (2006). Wiegmann et al. (2009) and Simon et al. (2009) corroborate this scenario. Phylogenomic results (see chapter 2) also propose that Hymenoptera branch off first within Endopterygota. However, analyses based on complete mitochondrial genomes contradict this scenario (e.g. Castro and Dowton, 2005).

3.5. Conclusions

The tree resolution of this arthropod rRNA data set demonstrates that data quality assessment and a cautious modeling incorporating background knowledge as well as an increased computational effort pays off. Applied approaches also show that the impact of biasing effects is not well understood according to different modeling. Tools are required to identify and characterize these effects properly and treat long branch taxa adequately in analyses if we want to address phylogenetic relationships in future research. Present analyses clearly show that the implementation of base compositional heterogeneity (non-stationarity) is crucial to reliably reconstruct phylogenies (see placement of *Hutchinsoniella* (Cephalocarida, Crustacea), and *Lepisma* (Zygentoma, Hexapoda)).

From a statistical point of view, non-stationary DNA models outperform non-stationary mixed models. Still, DNA topologies show some suspicious results with negligible support, for example a paraphyly of Eosentomata (Protura) or a polyphyly of hymenopteran insects (Endopterygota) which lacks any morphological, developmental and molecular evidence. Therefore, it is suggested that modeling site-interdependence plus non-stationarity increases the chance to infer robust topologies.

4. Future Prospects

Methodological considerations towards phylogenomics

Extension of MARE

Extended geometry mapping (eGM) implemented in MARE is a technically convenient way of estimating potential information. In its current application, it is a rough assessment of potential signal among taxa, as it does not consider different phylogenetic taxa within a data set. An extension of a hypothesis-driven assessment of potential signal opens the possibility to a hypothesis-driven "balanced" taxon/gene selection and to reductions of edge-weighted matrices accordingly. Implementation of different empirical matrices and a refined MARE version for a nucleotide level is should be taken into account. For both, the evaluation of potential information content and reduction heuristics, not only a particular gene overlap but, additionally a certain overlap of positions between taxa should be monitored. Further, implementing measures of phylogenetic information content from additional characters from a new taxon added to a phylogenetic quartet proposed by Townsend and Lopez-Giraldez (2010) or measuring of information content based on invariants (Allman and Rhodes, 2007; Casanellas and Fernández-Sánchez, 2007) are considerable. Techniques of maximal cliques and/or quasi-bicliques (see Discussion in chapter 2) could be implemented to select optimal submatrices.

Nucleotide versus amino acid analyses

Nucleotide analyses seem worthwhile for the present arthropod data sets, as proposed by Regier et al. (2010). However, suggested methods are currently not applicable on data sets of this thesis. The unreduced nucleotide data set span up to 1 million bp, the SOS 100.000 bp. Even with parallelized software and high performance computing (HPC), ML analyses will probably exeed currently available computational resources. Phylogenetic reconstruction with PhyloBayes or MrBayes are simply unrealistic with respect to parameter convergence. Therefore, extended algrorithms which require less computation time and less RAM should be developed.

Establishement of a primertoolbox for genomic DNA

The establishment of a primertoolbox for amplifying genes possibly compensates the sparse data availability for particular taxa. Thereby, primers that work on genomic DNA is favorable because DNA can easily be preserved and remains stable. Recently, an automated primer design tool has been developed by J. Borner, University of Hamburg (Borner, unpublished). The software can generate degenerated primers from amino acid alignments that work on genomic DNA, while possible secondary structures are considered. A check of the primer location with respect to introns and exons is included against closely related whole genome species. This seems an alternative way to amplify many genes for species, where an EST approach is not possible.

Processing ESTs from 454 sequencing and check of contamination
In future, ESTs will be exclusively generated by 454 sequencing techniques. A check for sequence contamination is highly demanded. The problem of sequence contamination in published data is known, but only addressed in Wägele et al. (e.g. 2009). A check for EST data derived from 454 techniques cannot be manually handled and bioinformatical tools are necessary to exclude possible contaminations.

Methods beyond Maximum Likelihood and Bayesian approaches
Recently, Criscuolo et al. (2006) and Criscuolo and Gascuel (2008) proposed a superdistance tree approach, where a tree is generated from a superdistance matrix by combining submatrices where most taxa are still present. This seems apparently more accurate than reconstructing trees from unreduced raw-matrices. Such an approach requires little computational effort without the need of HP computing Another approach to infer a "mega-phylogeny" is based on data based sequences as well as taxonomic hierarchies (Smith et al., 2009). Furthermore, alignment free methods are still matter of interest for phylogenetic inference, although this idea is not new (Thorne and Kishino, 1992; Vinga and Almeida, 2003; Höhl and Tagan, 2007; Dai et al., 2008; Daskalakis and Roch, 2010). Since phylogenomic data will have the potential of mega-phylogenomic data sets, considerations about alignment free methods for phylogenomics (Huiguang, 2010) should be seriously taken up. Invariant techniques bear alternative approaches to infer phylogeny (Cipra, 2007; Casanellas and Fernández-Sánchez, 2007; Matsen, 2009; Casanellas and Fernández-Sánchez, 2010) independent from e.g. heterogeneous processes and should be taken into account in future studies.

Methodological considerations rRNA genes

An implementation of mixed models into ML tree-reconstruction software is desirable to account for character interdependence within ribosomal RNA genes. Thereby, partitioning into paired and unpaired sites and gene partitioning should be implemented (see chapter 3). Furthermore, a proper model selection that accounts for saturated loops (Letsch et al., 2010) is demanded. Testing the performance of mixture models (Venditti et al., 2008) is at least worthwhile to overcome a possible over-parametrization applying mixed RNA/DNA models.

Proper modeling to avoid artifacts

Slow-fast analyses may be a simple and effective method to reduce the influence of substitution saturation, one of the causes of phylogenetic noise and long branch artifacts Kostka et al. (2008). Long branch artifacts should be not relevant for non-parsimonious analyses. Recent studies, however, showed that such artifacts occur due to a bad performance of MCMC processes, or to unproper branch length estimates in Bayesian approaches (Kolaczkowski and Thornton, 2009; Brown et al., 2010). The risk, that (hidden) long branch artifacts (Wägele and Mayer, 2007) occur, might also be due to model misspecification and/or an unproper application of the GAMMA parameter and invariant sites (I) in Bayesian and ML analyses (Yang, 1996; Waddell et al., 1997; Minin et al., 2003; Mayrose et al., 2005). Currently, there are no methods to identify such artifacts. In addition, a proper evaluation and handling of heterotachy effects (Grievink et al., 2010) is a challenge in (phylogenomic) real-world data sets.

References

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, K. A. R., W. R. McCombie, and J. C. Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656.

Aguinaldo, A. M. A., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493.

Aleshin, V. V., K. V. Mikhailov, A. V. Konstantinova, M. A. Nikitina, L. Y. Rusinc, D. A. Buinova, O. S. Kedrova, and N. B. Petrov. 2009. On the phylogenetic position of insects in the Pancrustacea clade. Mol Biol (Mosk) 43:804–818.

Alexe, G., S. Alexe, Y. Crama, S. Foldes, P. L. Hammer, and B. Simeone. 2002. Consensus algorithms for the generation of all maximal bicliques. DIMACS Technical Reports 2002-52 Rutgers University Piscataway, NJ, USA.

Allman, E. S. and J. A. Rhodes. 2007. Phylogenetic invariants. chap. 4, Pages 108–146 *in* Reconstructing evolution: new mathematical and computational advances (O. Gascuel and M. Steel, eds.). Oxford University Press, Oxford.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Anderson, F. E. and D. L. Swofford. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol Phylogenet Evol 33:440–451.

Averof, M. and M. Akam. 1995. Insect-Crustacean Relationships: Insights from Comparative Developmental and Molecular Studies. Philos Trans R Soc Lond, B, Biol Sci 347:293–303.

Ax, P. 1990. Das System der Metazoa. Ein Lehrbuch der phylogenetischen Systematik. Gustav Fischer Verlag, Stuttgart.

Bacetti, R. and R. Dallai. 1973. The spermatozoon of Arthropoda. XXII. The "12+0", "14+0" or aflagellate sperm of Protura. J Cell Sci 13:321–333.

Bäcker, H., M. Fanenbruck, and J. W. Wägele. 2008. A forgotten homology supporting the monophyly of Tracheata: The subcoxa of insects and myriapods re-visited. Zool Anz 247:185–207.

Ban, N., P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å Resolution. Science 289:905–920.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Duruflé, T. Gaasterland, P. Lopez, M. Müller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci USA 99:1414–1419.

Baurain, D., H. Brinkmann, and H. Philippe. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol 24:6–9.

Beiko, R. G., J. M. Keith, T. J. Harlow, and M. A. Ragan. 2006. Searching for convergence in phylogenetic Markov Chain Monte Carlo. Syst Biol 55:553–565.

Berglund, A.-C., E. Sjölund, G. Ostlund, and E. L. L. Sonnhammer. 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucl Acids Res 36:D263–266.

Beutel, R. G. ., F. Friedrich, T. Hörnschemeyer, H. Pohl, F. Hünefeld, F. Beckmann, R. Meier, B. Misof, M. Whiting, and L. Villhemsen. 2010. Morphological and molecular evidence converging upon a robust phylogeny of the megadiverse Holometabola. *In press*. doi: 10.1111/j.1096-0031.2010.00305.x.

Beutel, R. G. and S. N. Gorb. 2001. Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny. J Zool Syst Evol Research 39:177–207.

Beutel, R. G. and S. N. Gorb. 2006. A revised interpretation of the evolution of attachment structures in Hexapoda with special emphasis on Mantophasmatodea. Arthropod Syst Phylogeny 64:3–25.

Beutel, R. G. and H. Pohl. 2006. Endopterygote systematics – where do we stand and what is the goal (Hexapoda, Arthropoda)? Syst Entomol 31:202–219.

Bilinski, S. 1994. The ovary of Entognatha. Pages 7–30 *in* The Insect Ovary (J. Büning, ed.). Chapman and Hall, London.

Bininda-Emonds, O. R. P. 2004. The evolution of supertrees. Trends Ecol Evol (Amst) 19:315–322.

Bininda-Emonds, O. R. P., J. Gittleman, and M. A. Steel. 2002. The (super) tree of life: Procedures, problems and prospects. Annu Rev Ecol Syst 33:265–289.

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and Genomewise. Genome Res 14:988–995.

Bitsch, C. and J. Bitsch. 2000. The phylogenetic interrelationships of the higher taxa of apterygote hexapods. Zool Scr 29:131–156.

Bitsch, J. and C. Bitsch. 2004. Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. Zool Scr 33:511–550.

Blanquart, S. and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol 23:2058–2071.

Bleidorn, C., L. Podsiadlowski, M. Zhong, I. Eeckhaut, S. Hartmann, K. M. Halanych, and R. Tiedemann. 2009. On the phylogenetic position of Myzostomida: Can 77 genes get it wrong? BMC Evol Biol 9:150.

Boguski, M. S. 1995. The turning point in genome research. Trends Biochem Sci 20:295–296.

Boguski, M. S., T. M. J. Lowe, and C. M. Tolstoshev. 1993. dbEST - database for "expressed sequence tags". Nat Genet 4:332–333.

Boore, J. L., T. M. Collins, D. Stanton, L. L. Daehler, and W. M. Brown. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376:163–165.

Boore, J. L., D. V. Lavrov, and W. M. Brown. 1998. Gene translocation links insects and crustaceans. Nature 392:667–668.

Börner, C. 1909. Neue Homologien zwischen Crustaceen und Hexapoden. Die Beissmandibel der Insecten und ihre phylogenetische Bedeutung. Archi- und Metapterygota. Zool Anz 34:100–125.

Börner, C. 1910. Die phylogenetische Bedeutung der Protura. Biol. Zbl. 30:633–641.

Bouck, A. and T. Vision. 2007. The molecular ecologist's guide to expressed sequence tags. Mol Ecol 16:907–924.

Boudreaux, B. H. 1979. Arthropod phylogeny: with special reference to insects. John Wiley and Sons Inc., New York, Chichester, Brisbane, Toronto.

Bourlat, S. J., C. Nielsen, A. D. Economou, and M. J. Telford. 2008. Testing the new animal phylogeny: A phylum level molecular analysis of the animal kingdom. Mol Phylogenet Evol 49:23–31.

Brinkmann, H. and H. Philippe. 2008. Animal phylogeny and large-scale sequencing: progress and pitfalls. J Syst Evol 46:274–286.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743–757.

Brown, J. M., S. M. Hedtke, A. R. Lemmon, and E. M. Lemmon. 2010. When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates. Syst Biol 59:145–161.

Brown, J. M. and A. R. Lemmon. 2007. The importance of data partitioning and the utility of bayes factors in bayesian phylogenetics. Syst Biol 56:643–655.

Bryant, D. and V. Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265.

Buckley, T. R., C. Simon, and G. K. Chambers. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst Biol 50:67–86.

Buckley, T. R., C. M. Simon, P. K. Flook, and B. Misof. 2000. Secondary structure and alignment of domains IV and V of the insect mitochondrial large subunit rRNA gene. Insect Mol Biol 9:565–580.

Budd, G. E. and M. J. Telford. 2009. The origin and evolution of arthropods. Nature 457:812–817.

Carapelli, A., S. Comandi, P. Convey, F. Nardi, and F. Frati. 2008. The complete mitochondrial genome of the Antarctic springtail *Cryptopygus antarcticus* (Hexapoda: Collembola). BMC Genomics 9:S12.

Carapelli, A., F. Frati, F. Nardi, R. Dallai, and C. Simon. 2000. Molecular phylogeny of the apterygotan insects based on nuclear and mitochondrial genes. Pedobiologia (Jena) 44:361–373.

Carapelli, A., P. Liò, F. Nardi, E. van der Wath, and F. Frati. 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. BMC Evol Biol 7, Suppl 2:S8.

Carapelli, A., F. Nardi, R. Dallai, J. L. Boore, P. Liò, and F. Frati. 2005. Relationships between hexapods and crustaceans based on four mitochondrial genes. Pages 295–306 *in* Crustacean and Arthropod Relationships (S. Koenemann and R. A. Jenner, eds.) vol. 16 of *Crustacean Issues*. CRC Press.

Carapelli, A., F. Nardi, R. Dallai, and F. Frati. 2006. A review of molecular data for the phylogeny of basal hexapods. Pedobiologia (Jena) 50:191–204.

Casanellas, M. and J. Fernández-Sánchez. 2007. Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees. Mol Biol Evol 24:288–293.

Casanellas, M. and J. Fernández-Sánchez. 2010. Relevant phylogenetic invariants of evolutionary models. *In press*. doi: 10.1016/j.matpur.2010.11.002.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552.

Castro, L. R. and M. Dowton. 2005. The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera: Symphyta: Pergidae). Mol Phylogenet Evol 34:469–479.

Chalwatzis, N., A. Baur, E. Stetzner, R. Kinzelbach, and F. K. Zimmermann. 1995. Strongly expanded 18S rRNA genes correlated with a peculiar morphology in the insect order of Strepsiptera. Zoology (Jena) 98:115–126.

Chen, F., A. J. Mackey, and D. S. Vermunt, Jeroen K. Ross. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2:e383.

Chou, H.-H. and M. H. Holmes. 2001. DNA sequence quality trimming and vector removal. Bioinformatics 17:1093–1104.

Chou, Q., M. Russell, D. E. Birch, J. Raymond, and W. Bloch. 1992. Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-NUMBER amplifications. Nucleic Acids Res 20:1717–1723.

Cipra, B. A. 2007. Algebraic geometers see ideal approach to biology. SIAM News 40:4 `http://www.siam.org/news/news.php?id=1146`, access 27th Nov 2010.

Comandi, S., A. Carapelli, L. Podsiadlowski, F. Nardi, and F. Frati. 2009. The complete mitochondrial genome of Atelura formicaria (Hexapoda: Zygentoma) and the phylogenetic relationships of basal insects. Gene 439:25–34.

Cook, C. E., Q. Yue, and M. Akam. 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. Proc R Soc Lond B Biol Sci 272:1295–1304.

Cotton, J. A. and M. Wilkinson. 2009. Supertrees join the mainstream of phylogenetics. Trends Ecol Evol (Amst.) 24:1–3.

Criscuolo, A., V. Berry, E. J. P. Douzery, and O. Gascuel. 2006. SDM: a fast distance-based approach for (super) tree building in phylogenomics. Syst Biol 55:740–755.

Criscuolo, A. and O. Gascuel. 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics 9:166.

Dai, Q., Y. Yang, and T. Wang. 2008. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics 24:2296–2302.

Dallai, R. 1991. Are Protura really insects? Pages 263–269 *in* The Early Evolution of the Metazoa and the Significance of Problematic Taxa (A. M. Simonetta and S. C. Morris, eds.). Cambridge University Press, Cambridge, UK.

Dallai, R. 1994. Recent findings on apterygotan sperm structure. Acta Zool Fenn 195:23–27.

Dallai, R. and B. Afzelius. 1999. Accessory microtubles in insect spermatozoa: structure, function and phylogenetic significance. Pages 333–350 *in* The Male Gamete: From Basic Knowledge to Clinical Applications (C. Cagnon, ed.). Cache River Press, Vienna, Austria, IL.

Dallai, R., D. Mercati, Y. Bu, Y. W. Yin, G. Callaini, and M. G. Riparbelli. 2010. The spermatogenesis and sperm structure of *Acerentomon microrhinus* (Protura, Hexapoda) with considerations on the phylogenetic position of the taxon. Zoomorphology 129:61–80.

Dallai, R., L. Xu'e, and W. Yin. 1990. Aflagellated spermatozoa of *Huhentomon* and *Acerella* (Protura, Apterygota). Int J Insect Morphol Embryol 19:211–217.

Dallai, R., L. Xu'e, and W. Yin. 1992. Flagellate spermatozoa of Protura (Insecta, Apterygota) are motile. Int J Insect Morphol Embryol 21:137–148.

Daskalakis, C. and S. Roch. 2010. Alignment-free phylogenetic reconstruction. Pages 123–137 *in* Research in Computational Molecular Biology (B. Berger, ed.) vol. 6044 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg.

Dawande, M., P. Keskinocak, J. Swaminathan, and S. Tayur. 2001. On bipartite and multipartite clique problems. Journal of Algorithms 41:388–403.

De Queiroz, A. and J. Gatesy. 2006. The supermatrix approach to systematics. Trends Ecol Evol (Amst) 22:34–41.

De Salle, R., J. Gatesy, W. Wheeler, and D. Grimaldi. 1992. DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. Science 257:1933–1936.

Deharveng, L. 2004. Recent advances in Collembola systematics. Pedobiologia (Jena) 48:415–433.

Dell'Ampio, E., N. U. Szucsich, A. Carapelli, F. Frati, G. Steiner, A. Steinacher, and G. Pass. 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.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375.

Delsuc, F., M. P. Phillips, and D. Penny. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science 301:1482.

Delsuc, F., G. Tsagkogeorga, N. Lartillot, and H. Philippe. 2008. Additional molecular support for the new chordate phylogeny. Genesis 46:592–604.

D'Haese, C. A. 2002a. Phylogeny of the apterygote hexapods. Abstracts, 20th Annual Meeting of the Willi Hennig Society. Cladistics 18:220.

D'Haese, C. A. 2002b. Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and optimization alignment. Proc R Soc Lond B Biol Sci 269:1143–1151.

D'Haese, C. A. 2003. Morphological appraisal of Collembola phylogeny with special emphasis on Poduromorpha and a test of the aquatic origin hypothesis. Zool Scr 32:563–586.

Dias, V. M., C. M. de Figueiredob, and J. L. Szwarcfiter. 2007. On the generation of bicliques of a graph. Discrete Applied Mathematics 155:1826–1832.

Dixon, M. T. and D. M. Hillis. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analyses. Mol Biol Evol 10:256–267.

Dohle, W. 2001. Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name "Tetraconata" for the monophyletic unit Crustacea + Hexapoda. Ann Soc Entomol Fr (New Series) 37:85–103.

Dohrmann, M., D. Janussen, J. Reitner, A. G. Collins, and G. Wörheide. 2008. Phylogeny and evolution of glass sponges (porifera, hexactinellida). Syst Biol 57:388–405.

Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, B. C. O'Meara, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis. Cambridge University Press, Cambridge, UK.

Ebersberger, I., S. Strauss, and A. von Haeseler. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9:157.

Edgar, R. C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar, R. C. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

Edgecombe, G. D. 2009. Palaeontological and molecular evidence linking arthropods, onychophorans, and other Ecdysozoa. Evo Edu Outreach 2:178–190.

Edgecombe, G. D. 2010. Arthropod phylogeny: An overview from the perspectives of morphology, molecular data and the fossil record. Arthropod Struct Dev 39:74–87.

Edgecombe, G. D. and G. Giribet. 2002. Myriapod phylogeny and the relationships of Chilopoda. Pages 143–168 *in* Biodiversidad, Taxonomía y Biogeografia de Artrópodos de México: Hacia una Síntesis de su Conocimiento (J. L. Bousquets and J. Morrone, eds.) vol. III. Aula-Verlag.

Edgecombe, G. D., G. D. F. Wilson, D. J. Colgan, M. R. Gray, and G. Cassis. 2000. Arthropod Cladistics: Combined analysis of histone H3 and U2 snRNA sequences and morphology. Cladistics 16:155–203.

Eigen, M., R. Winkler-Oswatitsch, and A. Dress. 1988. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. Proc Natl Acad Sci USA 85:5913–5917.

Erpenbeck, D., S. A. Nichols, O. Voigt, M. Dohrmann, B. M. Degnan, J. N. A. Hooper, and G. Wörheide. 2007. Phylogenetic analyses under secondary structure-specific substitution models outperform traditional approaches: case studies with diploblast LSU. J Mol Evol 64:543–557.

Ertas, B., B. M. von Reumont, J.-W. Wägele, B. Misof, and T. Burmester. 2009. Hemocyanin suggests a close relationship of Remipedia and Hexapoda. Mol Biol Evol 26:2711–2718.

Fanenbruck, M. 2003. Die Anatomie des Kopfes und des cephalen Skelett-Muskelsystems der Crustacea, Myriapoda und Hexapoda: Ein Beitrag zum phylogenetischen System der Mandibulata und zur Kenntnis der Herkunft der Remipedia und Tracheata. Ph.D. thesis Ruhr-Universtät Bochum, Germany, Fakultät für Biologie.

Fanenbruck, M. and S. Harzsch. 2005. A brain atlas of *Godzilliognomus frondosus* Yager, 1989 (Remipedia, Godzilliidae) and comparison with the brain of *Speleonectes tulumensis* Yager, 1987 (Remipedia, Speleonectidae): implications for arthropod relationships. Arthropod Struct Dev 34:343–378.

Fanenbruck, M., S. Harzsch, and J. W. Wägele. 2004. The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. Proc Natl Acad Sci USA 101:3868–3873.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19:99–113.

Fitch, W. M. 2000. Homology a personal view on some of the problems. Trends Genet 16:227–231.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst Biol 53:485–495.

Fox, G. E. and C. R. Woese. 1975. The architecture of 5S rRNA and its relation to function. J Mol Evol 6:61–76.

Friedrich, F. and R. G. Beutel. 2010. Goodbye Halteria? The thoracic morphology of Endopterygota (Insecta) and its phylogenetic implications. Cladistics 26:1–34.

Friedrich, M. and D. Tautz. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165–167.

Friedrich, M. and D. Tautz. 1997. An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Mol Biol Evol 14:644–653.

Friedrich, M. and D. Tautz. 2001. Arthropod rDNA phylogeny revisited: A consistency analysis using Monte Carlo simulation. Ann Soc Entomol Fr (New Series) 37:21–40 origin of Hexapoda ed. by T. Deuve.

Gai, Y.-H., D.-X. Song, H.-Y. Sun, and K.-Y. Zhou. 2006. Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences. Zool Sci 23:1101–1108.

Galtier, N. 2004. Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites. Syst Biol 53:38–46.

Galtier, N. and V. Daubin. 2008. Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond, B, Biol Sci 363:4023–4029.

Galtier, N. and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA 92:11317–11321.

Gao, Y., Y. Bu, and Y.-X. Luan. 2008. Phylogenetic relationships of basal hexapods reconstructed from nearly complete 18S and 28S rRNA gene sequences. Zoolog Sci 25:1139–1145.

Gillespie, J. J., J. S. Johnston, J. J. Cannone, and R. R. Gutell. 2005a. Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of Apis mellifera (Insecta: Hymenoptera): structure, organization, and retrotransposable elements. Insect Mol Biol 15:657–686.

Gillespie, J. J., J. B. Munro, J. M. Heraty, M. J. Yoder, A. K. Owen, and A. E. Carmichael. 2005b. A secondary structural model of the 28S rRNA expansion segments D2 and D3 for chalcidoid wasps (Hymenoptera: Chalcidoidea). Mol Biol Evol 22:1593–1608.

Giribet, G. 2003. Molecules, development and fossils in the study of metazoan evolution; Articulata versus Ecdysozoa revisited. Zoology (Jena) 106:303–326.

Giribet, G., S. Carranza, J. Baguñà, M. Riutort, and C. Ribera. 1996. First molecular evidence for the existence of a Tardigrada + Arthropoda clade. Mol Biol Evol 13:76–84.

Giribet, G., G. D. Edgecombe, J. M. Carpenter, C. A. D'Haese, and W. C. Wheeler. 2004. Is Ellipura monophyletic? A combined analysis of basal hexapod relationships with emphasis on the origin of insects. Org Divers Evol 4:319–340.

Giribet, G., G. D. Edgecombe, and W. C. Wheeler. 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature 413:157–161.

Giribet, G. and C. Ribera. 2000. A review of arthropod phylogeny: New data based on ribosomal DNA sequences and direct character optimization. Cladistics 16:204–231.

Giribet, G., S. Richter, G. D. Edgecombe, and W. C. Wheeler. 2005. The position of crustaceans within the Arthropoda - evidence from nine molecular loci and morphology. Pages 307–352 *in* Crustacean and Arthropod Relationships (S. Koenemann and R. A. Jenner, eds.) vol. 16 of *Crustacean Issues*. CRC Press, Boca Raton.

Glenner, H., P. F. Thomsen, M. B. Hebsgaard, M. V. Sørensen, and E. Willerslev. 2006. Evolution: The origin of insects. Science 314:1883–1884.

Gowri-Shankar, V. and H. Jow. 2006. *PHASE*: a software package for *Ph*ylogenetics *A*nd *S*equence *E*volution. University of Manchester 2.0 ed.

Gowri-Shankar, V. and M. Rattray. 2006. On the correlation between composition and site-specific evolutionary rate: Implications for phylogenetic inference. Mol Biol Evol 23:352–364.

Gowri-Shankar, V. and M. Rattray. 2007. A reversible jump method for bayesian phylogenetic inference with a nonhomogeneous substitution model. Mol Biol Evol 24:1286–1299.

Green, P. 1996. Crossmatch. University of Montreal available from `http://www.incogen.com/public_documents/vibe/details/crossmatch.html` ed.

Grievink, L. S., D. Penny, M. D. Hendy, and B. R. Holland. 2010. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. Syst Biol 59:288–297.

Grimaldi, D. and M. S. Engel. 2005. Evolution of Insects. Cambridge University Press, Cambridge, UK.

Grimaldi, D. A. 2010. 400 million years on six legs: On the origin and early evolution of Hexapoda. Arthropod Struct Dev 39:191–203.

Grünewald, S., K. Forslund, A. Dress, and V. Moulton. 2007. QNet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. Mol Biol Evol 24:532–538.

Gutell, J. C., Robin R .and Lee and J. J. Cannone. 2002. The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol 12:301–310.

Hall, T. A. 1999. BioEdit: a user-friendly biological alignment sequence EDITOR and analysis program for Windows95/98/NT. Nucleic Acids Symp Ser 41:95–98.

Hancock, J. M., D. Tautz, and G. A. Dover. 1988. Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. Mol Biol Evol 5:393–414.

Hartmann, S. and T. J. Vision. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol 8:95:S13.

Harzsch, S. 2006. Neurophylogeny: Architecture of the nervous system and a fresh view on arthropod phyologeny. Integr Comp Biol 46:162–194.

Harzsch, S., C. H. G. Müller, and H. Wolf. 2005. From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and "Myriapoda" but favour the Mandibulata concept. Dev Genes Evol 215:53–68.

Hassanin, A. 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol Phylogenet Evol 38:100–16.

Hassanin, A., N. Léger, and J. Deutsch. 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences. Syst Biol 54:277–298.

Hausdorf, B., M. Helmkampf, A. Meyer, H. Witek, Alexander Herlyn, I. Bruchhaus, T. Hankeln, T. Struck, and B. Lieb. 2007. Spiralian phylogenomics supports the resurrection of bryozoa comprising Ectoprocta and Entoprocta. Mol Biol Evol 24:2723–2729.

Heath, T. D., S. M. Hedtke, and D. M. Hillis. 2008a. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257.

Heath, T. D., D. J. Zwickl, J. Kim, and D. M. Hillis. 2008b. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst Biol 57:160–166.

Helmkampf, M., I. Bruchhaus, and B. Hausdorf. 2008. Multigene analysis of lophophorate and chaetognath phylogenetic relationships. Mol Phylogen Evol 46:206–214.

Hendy, M. D. and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool 38:297–309.

Hennig, W. 1953. Kritische Bemerkungen zum phylogenetischen System der Insekten. Beitr Ent 3:1–85 sonderheft.

Hennig, W. 1969. Die Stammesgeschichte der Insekten. Verlag von Waldemar Kramer in Frankfurt am Main, Frankfurt a. M.

Hennig, W. 1981. Insect Phylogeny. John Wiley and Sons, New York.

Heymonds, R. 1901. Die Entwicklungsgeschichte der Scolopender. Zoologica H. 33:1–244, Taf.I–VIII.

Hickson, R. E., C. Simon, and S. W. Perrey. 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Mol Biol Evol 17:530–539.

Hillis, D. M. and M. T. Dixon. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. The Quarterly Review of Biology 66:411–453.

Hirst, S., S. amd Maulik. 1926. On some arthropod remains from the Rhynie chert (Old Red Sandstone). Geological Magazine 63:69–71.

Höhl, M. and M. A. Tagan. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol 56:206–221.

Holland, B. and V. Moulton. 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. Pages 165–176 *in* Workshop on Algorithms in Bioinformatics (WABI) 2812, Proceedings LNBI, Springer, Berlin / Heidelberg.

Holland, B. R., G. Conner, K. Huber, and V. Moulton. 2007. Imputing supertrees and supernetworks from quartets. Syst Biol 56:57–67.

Holmes, D. S. and J. Bonner. 1973. Preparation, molecular weight, base composition, and secondary strucdture of giant nuclear ribonucleic acid. Biochemistry 12:2330–2338.

Hou, X. and J. Bergström. 1995. Cambrian lobopodians-ancestors of extant onychophorans? Zool J Linn Soc 114:3–19.

Hudelot, C., V. Gowri-Shankar, H. Jow, M. Rattray, and P. G. Higgs. 2003. RNA-based phylogenetic methods: Application to mammalian mitochonmdrial RNA sequences. Mol Phylogenet Evol 45:241–252.

Hudson, M. E. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8:3–17.

Hughes, J., S. J. Longhorn, A. Papadopoulou, K. Theodorides, A. de Riva, M. Mejia-Chang, p. G. Foster, and A. P. Vogler. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Mol Biol Evol 23:268–278.

Huiguang, Y. 2010. Hashtree: An ultrafast Alignment free and assembly free phylogenomics approach.

Huson, D. H. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267.

Hwang, U. W., M. Friedrich, D. Tautz, C. J. Park, and W. Kim. 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature 413:154–157.

Ikeda, Y. and R. Machida. 1998. Embryogenesis of the Dipluran *Lepidocampa weberi* Oudemans (Hexapoda, Diplura, Campodeidae): External Morphology. J Morhol 237:101–115.

Imadaté, G. 1966. Taxonomic arrangement of Japanese Protura (VI). The proturan chaetotaxy and its meaning to phylogeny. Bull. Natn. Sci. Mus., Ser A (Zoology) (Tokyo) (VI) 9:277–315.

Jamieson, B. G. M. 1987. The ultrastructure and phylogeny of insect spermatozoa chap. 4, Pages 81–89. Science Publishers.

Jamieson, B. G. M., R. Dallai, and B. A. Afzelius. 1999. Insects: their Spermatozoa and Phylogeny chap. 4, Pages 60–80. Science Publishers.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231.

Jenner, R. A. 2010. Higher-level crustacean phylogeny: Consensus and conflicting hypotheses. Arthropod Struct Dev 39:143–153.

Jermiin, L. S., S. Y. W. Ho, F. Ababneh, J. Robinson, and A. W. D. Larkum. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol 53:638–643.

Jordal, B., J. J. Gillespie, and A. I. Cognato. 2008. Secondary structure alignment and direct optimization of 28S rDNA sequences provide limited phylogenetic resolution in bark and ambrosia beetles (Curculionidae: Scolytinae). Zool Scr 37:43–56.

Jow, H., C. Hudelot, M. Rattray, and P. G. Higgs. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol Biol Evol 19:1591–1601.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467.

Kaas, R. E. and A. E. Raftery. 1995. Bayes Factors. Journal of the American Statistical Association 90:773–795.

Kadner, D. and A. Stollewerk. 2004. Neurogenesis in the chilopod *Lithobius forficatus* suggests more similarities to chelicerates than to insects. Dev Genes Evol 214:367–379.

Katoh, K. and H. Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298.

Kelchner, S. A. and M. A. Thomas. 2007. Model use in phylogenetics: nine key questions. Trends Ecol Evol (Amst.) 22:87–94.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. Mol Phylogenet Evol 4:314–330.

Kjer, K. M. 2004. Aligned 18S and insect phylogeny. Syst Biol 53:506–514.

Kjer, K. M., F. L. Carle, J. Litman, and J. Ware. 2006. A Molecular Phylogeny of Hexapoda. Arthropod Syst Phylogeny 64:35–44.

Kjer, K. M. and R. L. Honeycutt. 2007. Site specific rates of mitochondrial genomes and the phylogeny of Eutheria. BMC Evol Biol 7:8.

Klass, K. and N. P. Kristensen. 2001. The ground plan and affinities of hexapods: recent progress and open problems. Ann Soc Entomol Fr (New Series) 37:265–298.

Klass, K.-D. 2007. Die Stammesgeschichte der Hexapoden: eine kritische Diskussion neuerer Daten und Hypothesen. Pages 413–450 *in* Evolution – Phänomen Leben (O. L. Linz, ed.) vol. 20 of *Denisia*. Eigenverlag, Linz, Austria.

Koch, M. 1997. Monophyly and phylogenetic position of the Diplura (Hexapoda). Pedobiologia (Jena) 41:9–12.

Koch, M. 2001. Mandibular mechanisms and the evolution of hexapods. Ann Soc Entomol Fr (New Series) 37:129–174.

Koenemann, S., R. A. Jenner, M. Hoenemann, T. Stemm, and B. M. von Reumont. 2010. Arthropod phylogeny revisited, with a focus on crustacean relationships. Arthropod Struct Dev 39:88–110.

Koenemann, S., J. Olesen, F. Alwes, T. M. Iliffe, M. Hoenemann, P. Ungerer, C. Wolff, and G. Scholtz. 2009. The post-embryonic development of Remipedia (Crustacea) – additional results and new insights. Dev Genes Evol 219:131–45.

Koenemann, S., F. R. Schram, A. Bloechl, and T. Iliffe. 2007. Post-embryonic development of remipede crustaceans. Evol Dev 9:117–121.

Kolaczkowski, B. and J. W. Thornton. 2009. Long-Branch Attraction Bias and Inconsistency in Bayesian Phylogenetics. PLoS One 4:e7891.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Gene 39:309–338.

Kostka, M., M. Uzlikova, I. Cepicka, and J. Flegr. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of *Blastocystis*. BMC Bioinformatics 9:341.

Kristensen, N. P. 1975. The phylogeny of hexapod "orders" . A critical review of recent accounts. J Zoolog Syst Evol Res 13:1–44.

Kristensen, N. P. 1981. Phylogeny of insect orders. Ann Rev Entomol 26:135–157.

Kristensen, N. P. 1991. Phylogeny of extant hexapods. Pages 125–140 *in* The Insects of Australia: A Textbook for Students and Research Workers (I. D. Naumann, J. F. Lawrence, E. S. Nielsen, J. P. Spradberry, R. W. Taylor, M. J. Whitten, and M. J. Littlejohn, eds.) vol. 1. CSIRO, Melbourne Univ. Press, Carlton, Victoria.

Kristensen, N. P. 1998. The groundplan and basal diversification of the hexapods. Pages 281–293 *in* Arthropod Relationships (R. A. Fortey and R. H. Thomas, eds.) vol. 55 of *The Systematics Association Special Volume Series*. Chapman and Hall, London.

Kristensen, N. P. 1999. Phylogeny of endopterygote insects, the most successful lineage of living organisms. Eur J Entomol 96:237–253.

Kück, P. 2009. ALICUT: a Perlscript which cuts ALISCORE identified RSS. Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany version 2.0 ed.

Kück, P., K. Meusemann, J. Dambach, B. Thormann, B. von Reumont, J. W. Wägele, and B. Misof. 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. 1983. Origin of the insect wing and wing articulation from the arthropodan leg. Can J Zool 61:1618–1669.

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. Pages 141–179 *in* The Insects of Australia: A Textbook for Students and Research Workers (I. D. Naumann, J. F. Lawrence, E. S. Nielsen, J. P. Spradberry, R. W. Taylor, M. J. Whitten, and M. J. Littlejohn, eds.) vol. 1. CSIRO, Melbourne Univ. Press, Carlton, Victoria.

Kukalová-Peck, J. and J. F. Lawrence. 2004. Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters. Eur J Entomol 101:95–144.

Lang, A. 1888. Lehrbuch der vergleichenden Anatomie. Gustav Fischer, Jena.

Lartillot, N., S. Blanquart, and T. Lepage. 2008. PhyloBayes 2.3 - a Bayesian software for phylogenetic reconstruction using mixture models. University of Montreal 2.3c, current versions available from `http://www.phylobayes.org` ed.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7:S4.

Lartillot, N. and H. Philippe. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109.

Lartillot, N. and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond, B, Biol Sci 363:1463–1472.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. M. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Syst Biol 58:130–145.

Letsch, H. O., C. Greve, P. Kück, G. Fleck, and B. Stocsits, Roman R.and Misof. 2009. Simultaneous alignment and folding of 28S rRNA sequences uncovers phylogenetic signal in structure variation. Mol Phylogenet Evol 53:758–771.

Letsch, H. O., P. Kück, R. R. Stocsits, and B. Misof. 2010. The impact of rRNA secondary structure consideration in alignment and tree reconstruction: simulated data and a case study on the phylogeny of hexapods. Mol Biol Evol. .

Li, G., M. Steel, and L. Zhang. 2008a. More taxa are not necessarily better for the reconstruction of ancestral character states. Syst Biol 57:647–653.

Li, J., K. Sim, G. Liu, and L. Wong. 2008b. Maximal quasi-bicliques with balanced noise tolerance: concepts and co-clustering applications. *in* Proceedings of the SIAM International Conference on Data Mining SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA SIAM.

Li, L., C. J. Stoeckert, J. Ross, and D. S. R. Ross. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189.

Li, W. and Y. Liu. 2007. Modeling species-genes data for efficient phylogenetic inference. Pages 429–440 *in* Proceedings LSS Computational Systems Bioinformatics Conference, August, 2007. vol. 6 LSS - Life Sciences Society.

Luan, Y.-x., J. M. Mallatt, R.-d. Xie, Y.-m. Yang, and W.-y. Yin. 2005. The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on on ribosomal RNA gene sequences. Mol Biol Evol 22:1579–1592.

Luan, Y.-x., Y.-G. Yao, R.-D. Xie, Y.-M. Yang, Y.-P. Zhang, and W.-Y. Yin. 2004. Analysis of 18S rRNA gene of *Octostigma sinensis* (Projapygoidea: Octostigmatidae) supports the monophyly of Diplura. Pedobiologia (Jena) 48:453–459.

Luan, Y.-x., Y. Zhang, Y. Qiao-yun, J. Pang, R.-d. Xie, and W.-y. Yin. 2003. Ribosomal DNA gene and phylogenetic relationships of Diplura and lower hexapods. Sci China, C, Life Sci 46:67–76.

Machida, R. 2006. Evidence from embryology for reconstructing the relationships of hexapod basal clades. Arthropod Syst Phylogeny 64:95–104.

Machida, R., Y. Ikeda, and K. Toro. 2002. Evolutionary changes in developmental potentials of the embryo proper and embryonic membranes in Hexapoda. Pages 1–11 *in* Proceedings of Arthropodan Embryonical Society of Japan vol. 37 AESJ.

Maddison, D. R., M. D. Baker, and K. A. Ober. 1999. Phylogeny of carabid beetles as inferred from 18S ribosomal DNA (Coleoptera: Carabidae). Syst Entomol 24:103–138.

Mallatt, J., C. W. Craig, and M. J. Yoder. 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.

Mallatt, J. and G. Giribet. 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, J. M., J. R. Garey, and J. W. Shultz. 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol Phylogenet Evol 31:178–191.

Manton, S. M. 1977. The Arthropoda, habits, functional morphology and evolution. Clarendon Press, Oxford.

Marshall, D. C. 2010. Cryptic failure of partitioned bayesian phylogenetic analyses: lost in the land of long trees. Syst Biol 59:108–117.

Massoud, Z. 1971. Contribution à la connaissance morphologique et systématique des Collemboles Neelidae. Rev Écol Biol Sol 8:195–198.

Matsen, F. A. 2009. Fourier transform inequalities for phylogenetic trees. IEEE/ACM Trans Comput Biol Bioinformatics 6:89–95.

Mayer, G. and P. M. Whintington. 2009. Velvet worm development links myriapods with chelicerates. Proc R Soc Lond B Biol Sci 276:3571–3579.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21, Suppl 2:ii151–ii158.

McKenna, D. D. and B. D. Farrell. 2010. 9-genes reinforce the phylogeny of holometabola and yield alternate views on the phylogenetic placement of strepsiptera. PLoS ONE 5:e11887.

Meusemann, K., B. M. von Reumont, S. Simon, F. Roeding, P. Kück, S. Strauss, I. Ebersberger, M. Walzl, G. Pass, S. Breuers, V. Achter, A. von Haeseler, T. Burmester, H. Hadrys, J. W. Wägele, and B. Misof. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol Biol Evol 27:2451–2464.

Meyer, B. 2009. Reduced superalignments can improve the accuracy of large phylogenetic trees. Master's thesis Biozentrum Grindel & Zoologisches Museum, University of Hamburg, Germany 55 p.

Michot, B., J.-P. Bachellerie, and F. Raynal. 1983. Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA. Nucleic Acids Res 11:3375–3391.

Minelli, A. 2009. Perspectives in Animal Phylogeny and Evolution. Oxford University Press, Oxford, UK.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52:674–683.

Misof, B. and K. Misof. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst Biol 58:21–34.

Misof, B., O. Niehuis, I. Bischoff, A. Rickert, D. Erpenbeck, and A. Staniczek. 2006. A hexapod nuclear SSU rRNA secondary-structure model and catalog of taxon-specific structural variation. J Exp Zool B Mol Dev Evol 306B:70–88.

Misof, B., O. Niehuis, I. Bischoff, A. Rickert, D. Erpenbeck, and A. Staniczek. 2007. Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. Zoology (Jena) 110:409–429.

Nardi, F., G. Spinsanti, J. L. Boore, A. Carapelli, R. Dallai, and F. Frati. 2003a. Hexapod origins: monophyletic or paraphyletic? Science 299:1887–1889.

Nardi, F., G. Spinsanti, J. L. Boore, A. Carapelli, R. Dallai, and F. Frati. 2003b. Response to Comment on "Hexapod origins: monophyletic or paraphyletic?". Science 301:1482.

Nieselt-Struwe, K. 1997. Graphs in sequence spaces: a review of statistical geometry. Biophys Chem 66:111–131.

Nieselt-Struwe, K. and A. von Haeseler. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol Biol Evol 18:1204–1219.

Noller, H. F. 2005. RNA structure: reading the ribosome. Science 309:1508–1514.

Nosek, J. 1967. The new species of Protura from central Europe. Zeitschrift der Arbeitsgemeinschaft österreichischer Entomologen 19:76–88.

Nosek, J. 1973. The European Protura: their taxonomy, ecology and distribution with keys for determination. Muséum d'Histoire Naturelle, Genève, Switzerland.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst Biol 53:47–67.

O'Brien, K. P., M. Remm, and E. L. L. Sonnhammer. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 1:D476–480.

Odronitz, F., S. Becker, and M. Kollmar. 2009. Reconstructing the phylogeny of 21 completely sequenced arthropod species based on their motor proteins. BMC Genomics 10:173.

Ogden, T. H. and M. F. Whiting. 2003. The problem with "the Paleoptera Problem": sense and sensitivity. Cladistics 19:432–442.

Ott, M., J. Zola, S. Aluru, and A. Stamatakis. 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. *in* Proceedings of ACM/IEEE Supercomputing conference 2007.

Pagés, J. 1959. Remarques sur la classification des Diploures. Trav. Lab. Zool. Station Aquicole Grimaldi Fac. Sci. Dijon 26:1–25.

Pagés, J. 1997. Notes sur les Diploures Rhabdoures (Insectes, Aptérygotes) No. 1 – Diplura Genavensia XXIII. Rev Suisse Zool 69:869–896.

Paps, J., J. Baguñà, and M. Riutort. 2009. Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. Proc R Soc Lond B Biol Sci 276:1245–1254.

Parkinson, J. and M. Blaxter. 2009. Expressed sequence tags: an overview. Pages 1–12 *in* Expressed Sequence Tags (ESTs): Generation and Analysis (J. Parkinson, ed.) vol. 533. Humana Press.

Paulus, H. F. 1979. Eye structure and the monophyly of the Arthropoda. chap. 6, Pages 299–383 *in* Arthropod Phylogeny (A. P. Gupta, ed.). Van Nostrand Reinhold Co., New York.

Penny, D., P. J. Lockhart, M. A. Steel, and M. D. Hendy. 1994. The role of models in reconstructing evolutionary trees. Pages 211–230 *in* Models in phylogeny reconstruction (R. W. Scotland, D. J. Diebert, and D. M. Williams, eds.) The Systematics Association Special Volume Series. Oxford University Press, U.S.A.

Pertea, G. 2005-2006. SeqClean. Dana-Farber Cancer Institute available from `http://compbio.dfci.harvard.edu/tgi/software/` ed.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–652.

Peterson, K. J., J. B. Lyons, K. S. Nowak, C. M. Takacs, M. J. Wargo, and M. A. McPeek. 2004. Estimating metazoan divergence times with a molecular clock. Proc Natl Acad Sci USA 101:6536–6541.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005a. Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36:541–562.

Philippe, H., R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Quéinnec, C. Da Silva, P. Wincker, H. Le Guyader, S. Leys, D. J. Jackson, F. Schreiber, D. Erpenbeck, B. Morgenstern, G. Wörheide, and M. Manuël. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19:706–712.

Philippe, H., A. Germot, and D. Moreira. 2000. The new phylogeny of eukaryotes. Curr Opin Genet Dev 10:596–601.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol Biol Evol 22:1246–1253.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. H. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. Mol Biol Evol 21:1740–1752.

Philippe, H. and M. J. Telford. 2006. Large-scale sequencing and the new animal phylogeny. Trends Ecol Evol (Amst) 21:614–620.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Pisani, D., L. Poling, M. Lyons-Weiler, and S. B. Hedges. 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol 2:1.

Pocock, R. I. 1893. On the classification of the tracheate Arthropoda. Zool Anz 16:271–275.

Podsiadlowski, L. 2006. The mitochondrial genome of the bristletail *Petrobius brevistylis* (Archaeognatha: Machilidae). Insect Mol Biol 15:253–258.

Podsiadlowski, L., A. Carapelli, F. Nardi, R. Dallai, M. Koch, J. L. Boore, and F. Frati. 2006. The mitochondrial genomes of *Campodea fragilis* and *Campodea lubbocki* (Hexapoda: Diplura): High genetic divergence in a morphologically uniform taxon. Gene 381:49–61.

Podsiadlowski, L., H. Kohlhagen, and M. Koch. 2007. The complete mitochondrial genome of *Scutigerella causeyae* (Myriapoda: Symphyla) and the phylogenetic position of Symphyla. Mol Phylogenet Evol 45:251–260.

Porco, D. and L. Deharveng. 2009. Phylogeny of Collembola based on cuticular compounds: inherent usefulness and limitation of a character type. Naturwissenschaften 96:943–954.

Posada, D. 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256.

Putney, S. D., W. C. Herlihy, and P. Schimmel. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature 302:718–721.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. The R Foundation for Statistical Computing Vienna, Austria version 2.9 ed. ISBN 3-900051-07-0.

Rannala, B. and Y. Ziheng. 2008. Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet 9:217–231.

Regier, J. C. and J. W. Shultz. 1997. Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. Mol Biol Evol 14:902–913.

Regier, J. C. and J. W. Shultz. 2001. Elongation factor-2: A useful gene for arthropod phylogenetics. Mol Phylogenet Evol 20:136–148.

Regier, J. C., J. W. Shultz, A. R. D. Ganley, A. Hussey, D. Shi, B. Ball, A. Zwick, J. E. Stajich, M. P. Cummings, J. W. Martin, and C. W. Cunningham. 2008. Resolving arthropod phylogeny: Exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst Biol 57:920–938.

Regier, J. C., J. W. Shultz, and R. E. Kambic. 2004. Phylogeny of basal hexapod lineages and estimates of divergence times. Ann Entomol Soc Am 97:411–419.

Regier, J. C., J. W. Shultz, and R. E. Kambic. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc R Soc Lond B Biol Sci 272:395–401.

Regier, J. C., J. W. Shultz, A. Zwick, A. Hussey, B. Ball, R. Wetzer, J. W. Martin, and C. W. Cunningham. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–83.

Remm, M., C. E. V. Storm, and E. L. L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052.

Rodríguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56:389–399.

Roeding, F., S. Hagner-Holler, H. Ruhberg, I. Ebersberger, von Haeseler Arndt, M. Kube, R. Reinhardt, and T. Burmester. 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Mol Phylogenet Evol 45:942–951.

Ronquist, F. and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rota-Stabelli, O. and M. L. Telford. 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.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol 7 (Suppl1):52:S2.

Rusek, J. 1982. *Octostigma herbivora* n. gen. & sp. juring plant roots in the Tonga Islands. New Zealand J Zool 9:25–32.

Sanderson, M. J. 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol Evol (Amst.) 13:105–109.

Sanderson, M. J. 2007. Construction and annotation of large phylogenetic trees. Aust Syst Bot 20:287–301.

Sanderson, M. J. and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci 8:374–379.

Sanderson, M. J., A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol Biol Evol 20:1036–1042.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467.

Savard, J., D. Tautz, S. Richards, G. M. Weinstock, R. A. Gibbs, J. H. Werren, H. Tettelin, and M. J. Lercher. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects. Genome Res 16:1334–1338.

Schierwater, B., M. Eitel, W. Jakob, H.-J. Osigus, H. Hadrys, S. L. Dellaporta, S.-O. Kolokotronis, and R. De Salle. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "Urmetazoon" hypothesis. PLoS Biol 7:e1000020.

Scholtz, G. and G. Edgecombe. 2006. The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. Dev Genes Evol 216:395–415.

Scholtz, G., B. Mittmann, and M. Gerberding. 1998. The pattern of Distal-less expression in the mouthparts of crustaceans, myriapods and insects: new evidence for a gnathobasic mandible and the common origin of Mandibulata. Int J Dev Biol 42:801–810.

Schöniger, M. and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol Phylogenet Evol 3:240–247.

Schöniger, M. and A. von Haeseler. 1995. Performance of the Maximum Likelihood, Neighbor Joining, and Maximum Parsimony methods when sequence sites are not independent. Syst Biol 44:533–547.

Schram, F. R. and S. Koenemann. 2004. Are the crustaceans monophyletic? Pages 319–329 *in* Assembling the tree of life (J. Cracraft and M. J. Donoghue, eds.). Oxford University Press, New York.

Schreiber, F., G. Wörheide, and B. Morgenstern. 2009. OrthoSelect: a web server for selecting orthologous gene alignments from EST sequences. Nucleic Acids Res 37:W185–188.

Shear, W. A. and G. D. Edgecombe. 2010. The geological record and phylogeny of the Myriapoda. Arthropod Struct Dev 39:174–190.

Shoemaker, J. S. and W. M. Fitch. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol Biol Evol 6:270–289.

Shultz, J. W. and J. C. Regier. 2000. Phylogenetic analysis of arthropods using two nuclear protein-encoding gene ssupports a crustacean + hexapod clade. Proc R Soc Lond B Biol Sci 267:1011–1019.

Silvestri, F. 1907. Descrizione di un nuovo genere di insetti Apterigoti, rappresentante di un nuovo ordine. Bollett Labor Zool gen ed agraria di Portici 1:296–311.

Simon, C., T. R. Buckley, F. Frati, J. B. Stewart, and A. T. Beckenbach. 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved Polymerase Chain Reaction primers for animal mitochondrial DNA. Annu Rev Ecol Evol Syst 37:545–579.

Simon, S., S. Strauss, A. von Haeseler, and H. Hadrys. 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol Biol Evol 26:2719–2730.

Smit, A. F. A., R. Hubley, and P. Green. 1996-2004. RepeatMasker. University of Montreal open-3.0, available from `http://www.repeatmasker.org` ed.

Smith, S. A., J. M. Beaulieu, and M. J. Donoghue. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol Biol 9:37.

Smith, S. A. and C. W. Dunn. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716.

Snodgrass, R. E. 1935. Principles of insect morphology. McGraw-Hill Book Co., Inc., New York and London.

Snodgrass, R. E. 1938. Evolution of the Annelida, Onychophora, and Arthropoda. Smithsonian Miscellaneous Collection 97:1–159.

Sonnhammer, E. L. L. and E. V. Koonin. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18:619–620.

Stamatakis, A. 2006a. Phylogenetic models of rate heterogeneity: A high performance computing perspective. *in* 20th International Parallel and Distributed Processing Symposium (IPDPS 2006), Proceedings, 25-29 April 2006, Rhodes Island, Greece IEEE.

Stamatakis, A. 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Staniczek, A. H. 2000. The mandible of silverfish (Insecta: Zygentoma) and mayflies (Ephemeroptera): its morphology and phylogenetic significance. Zool Anz 239:147–178.

Staniczek, A. H. 2003. Die Mandibel der dicondylen Insekten: neue Erkenntnisse zur Klärung der basalen Spaltungsereignisse in der Phylogenie der Fluginsekten. Pages 89–91 *in* Verhandlungen der Westdeutschen Entomologischen Tagung, 2002, Löbbeke Museum, Düsseldorf.

Staniczek, A. H. and G. Bechly. 2007. "Apterygota": Primarily wingless insects. Pages 149–154 *in* The Crato Fossil Beds of Brazil: Window into an Ancient World (D. Martill, G. Bechly, and S. Heads, eds.). Cambridge University Press.

Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst Biol 49:225–232.

Steel, M. and A. Rodrigo. 2008. Maximum likelihood supertrees. Syst Biol 57:243–250.

Steel, M. and M. J. Sanderson. 2010. Characterizing phylogenetically decisive taxon coverage. Applied Mathematics Letters 23:82–86.

Steel, M. A. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Appl Math Lett 7:19–24.

Stephan, W. 1996. The rate of compensatory evolution. Genetics 144:419–426.

Stocsits, R. R., H. Letsch, J. Hertel, B. Misof, and P. F. Stadler. 2008. RNAsalsa. Zoologisches Forschungsmuseum A. Koenig, Bonn version 0.7.3, current versions available from `http://rnasalsa.zfmk.de` ed.

Stocsits, R. R., H. Letsch, J. Hertel, B. Misof, and P. F. Stadler. 2009. Accurate and Efficient Reconstruction of Deep Phylogenies from Structured RNAs. Nucleic Acids Res 37:6184–6193.

Stollewerk, A. and A. D. Chipman. 2006. Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships. Integr Comp Biol 46:195–206.

Stollewerk, A. and P. Simpson. 2005. Evolution of early development of the nervous system: a comparison between arthropods. Bioessays 27:874–883.

Strausfeld, N., I. Sinakevitch, S. M. Brown, and S. M. Farris. 2009. Ground plan of the insect mushroom body: functional and evolutionary implications. J Comp Neurol 513:265–291.

Strausfeld, N. J. 2009. Brain organization and the origin of insects: an assessment. Proc R Soc Lond B Biol Sci 276:1929–1937.

Sullivan, J. and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol 50:723–729.

Susko, E., M. Spencer, and A. J. Roger. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. J Mol Evol 61:351–359.

Swofford, D. L. 2003. *PAUP**. Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer Associates Sunderland, Massachusetts version 4 ed.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular Systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland.

Szucsich, N. U. and G. Pass. 2008. Incongruent phylogenetic hypotheses and character conflicts in morphology: The root and early branches of the hexapodan tree. Mitt Dtsch Ges Allg Angew Entomol 16:415–429.

Tabei, Y., H. Kiryu, K. T., and K. Asai. 2008. A fast structural multiple alignment method for long RNA sequences. BMC Bioinformatics 9:33.

Tarrío, R., F. Rodríguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol Biol Evol 18:1464–1473.

Tatusova, T. A. and T. L. Madden. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247–250.

Telford, M. J., M. J. Wise, and V. Gowri-Shankar. 2005. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the Bilateria. Mol Biol Evol 22:1129–1136.

Thomson, R. C. and H. B. Shaffer. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst Biol 59:42–58.

Thorley, J. L. and R. D. M. Page. 2000. RadCon: phylogenetic tree comparison and consensus. Bioinformatics 16:486–487.

Thorley, J. L. and M. Wilkinson. 1999. Testing the phylogenetic stability of early tetrapods. J Theor Biol 200:343–344.

Thorne, J. L. and H. Kishino. 1992. Freeing phylogenies from artifacts of alignment. Mol Biol Evol 9:1148–1162.

Tillier, E. R. M. and R. A. Collins. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 148:1993–2002.

Timmermans, M. J., D. D Roelofs, J. Mariën, and N. M. van Straalen. 2008. Revealing pancrustacean relationships: Phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. BMC Evol Biol 8:83.

Townsend, J. P. and F. Lopez-Giraldez. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst Biol 59:446–457.

Tuxen, S. L. 1959. The phylogenetic significance of entognathy in entognathous apterygotes. Smithsonian Miscellaneous Collection 137:379–416.

Tuxen, S. L. 1964. The Protura. A Revision of the species of the world with keys for determination. Hermann, Paris.

Tuxen, S. L. 1977. The systematical position of Sinentomon (Insecta, Protura). Bull. Natn. Sci. Mus., Ser A (Zoology) (Tokyo) 3:25–36.

Ungerer, P. and G. Scholtz. 2008. Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata. Proc R Soc Lond B Biol Sci 275:369–376.

Štys, P. and S. Bilinski. 1990. Ovariole types and the phylogeny of hexapods. Biol Rev 65:401–429.

Štys, P. and J. Zrzavý. 1994. Phylogeny and classification of extant Arthropoda: review of hypotheses and nomenclature. Eur J Entomol 91:257–275.

Štys, P., J. Zrzavý, and F. Weyda. 1993. Phylogeny of the Hexapoda and ovarian metamerism. Biol Rev 68:365–379.

Van de Peer, Y., P. De Rijk, J. Wuyts, T. Winkelmans, and R. De Wachter. 2000. The European Small Subunit Ribosomal RNA database. Nucleic Acids Res 28:175–176.

Venditti, C., A. Meade, and M. Pagel. 2008. Phylogenetic mixture models can reduce node-density artifacts. Syst Biol 57:286–293.

Vinga, S. and J. Almeida. 2003. Alignment-free sequence comparisonâĂŤa review. Bioinformatics 19:513–523.

Voigt, O., D. Erpenbeck, and G. Wörheide. 2008. Molecular evolution of rDNA in early diverging Metazoa: first comparative analysis and phylogenetic application of complete SSU rRNA secondary structures in Porifera. BMC Evol Biol 8:69.

von Reumont, B. M., K. , N. U. Szucsich, E. Dell'Ampio, D. Bartel, S. Simon, H. O. Letsch, R. R. Stocsits, Y.-x. Luan, J. W. Wägele, G. Pass, H. Hadrys, and B. Misof. 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, B. M. 2010. Molecular insights to crustacean phylogeny. A status quo of past, present and perspective prospects also covering phylogenomics. Ph.D. thesis Rheinische Friedrich-Wilhelms-Universität Bonn, Germany, Zool. Forschungsmuseum A. Koenig, Molecular Lab 269 p.

Waddell, P. J., D. Penny, and T. Moore. 1997. Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol Phylogenet Evol 8:33–50.

Wägele, J. W. 1993. Rejection of the 'Uniramia' hypothesis and implications of the Mandibulata concept. Zoologische Jahrbücher. Abteilung für Systematik, Ökologie und Geographie der Tiere 120:253–288.

Wägele, J. W., H. Letsch, A. Klussmann-Kolb, C. Mayer, B. Misof, and H. Wägele. 2009. Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny). Front Zool 6:12.

Wägele, J. W. and C. Mayer. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol Biol 7:147.

Walossek, D. 1993. The upper Cambrian Rehbachiella and the phylogeny of Branchiopoda and Crustacea vol. 32 of *Fossils and Strata*. Scandinavian University Press.

Walossek, D. 1999. On the Cambrian diversity of Crustacea. Pages 3–27 *in* Crustaceans and the Biodiversity Crisis, Proceedings of the Fourth International Crustacean Congress, Amsterdam, The Netherlands, July 20-24, 1998 (F. R. Schram and J. C. von Vaupel Klein, eds.) vol. 1 Brill Academic Publishers, Leiden.

Waloszek, D. 2003. Cambrian 'Orsten'-type preserved arthropods and the phylogeny of Crustacea. . Pages 69–87 *in* The new panorama of animal evolution, Proceedings of the 18th International Congress of Zoology, Athen, Greece, 2000 (A. Legakis, R. Sfenthourakis, R. Polymeni, and M. Thessaqlou-Legaki, eds.) Pensoft Publishers, Sofia, Moscow, Russia.

Ware, J. L., S. Y. W. Hob, and K. Kjer. 2008. Divergence dates of libelluloid dragonflies (Odonata: Anisoptera) estimated from rRNA using paired-site substitution models. Mol Phylogenet Evol 47:426–432.

Westheide, W. and R. Rieger. 2007. Spezielle Zoologie, Bd.1 : Einzeller und Wirbellose Tiere. Elsevier, München.

Wheeler, W. C., G. Giribet, and G. D. Edgecombe. 2004. Arthropod systematics. The comparative study of genomic, anatomical, and paleontological information. Pages 281–295 *in* Assembling the tree of life (J. Cracraft and M. J. Donoghue, eds.). Oxford University Press, New York.

Wheeler, W. C. and R. L. Honeycutt. 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. Mol Biol Evol 5:90–96.

Wheeler, W. C., M. Whiting, Q. D. Wheeler, and J. M. Carpenter. 2001. The phylogeny of extant hexapod orders. Cladistics 17:113–169.

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

Whitfield, J. B. and K. M. Kjer. 2008. Ancient rapid radiations of insects: Challenges for phylogenetic analysis. Annu Rev Entomol 53:449–472.

Wiegmann, B., M. Trautwein, J.-W. Kim, B. Cassel, M. Bertone, S. Winterton, and D. Yeates. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol 7:34.

Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst Biol 47:625–640.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528–538.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol 54:731–742.

Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. J Biomed Inform 39:34–42.

Wiens, J. J. and D. S. Moen. 2008. Missing data and the accuracy of Bayesian phylogenetics. J. Syst. Evol. 46:307–314.

Wilkinson, M., J. A. Cotton, C. Creevey, O. Eulenstein, S. R. Harris, F.-J. Lapointe, C. Levasseur, J. O. Mcinerney, D. Pisani, and J. L. Thorley. 2005. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst Biol 54:419–431.

Wuyts, J., P. De Rijk, Y. Van de Peer, G. Pison, P. Rousseeuw, and R. De Wachter. 2000. Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. Nucleic Acids Res 28:4698–4708.

Wuyts, J., G. Perrière, and Y. Van de Peer. 2004. The European ribosomal RNA database. Nucleic Acids Res 32:D101–103.

Wuyts, J., Y. Van de Peer, T. Winkelmans, and R. De Wachter. 2002. The European database on small subunit ribosomal RNA. Nucleic Acids Res 30:183–185.

Xiong, Y., Y. Gao, W.-y. Yin, and Y.-x. Luan. 2008. Molecular phylogeny of Collembola inferred from ribosomal RNA genes. Mol Phylogenet Evol 49:728–735.

Yamaguchi, S. and K. Endo. 2003. Molecular phylogeny of Ostracoda (Crustacea) inferred from 18S ribosomal DNA sequences: implication for its origin and diversification. Mar Biol 143:23–38.

Yan, C., J. G. Burleigh, and O. Eulenstein. 2005. Identifying optimal incomplete phylogenetic data sets from sequence databases. Mol Phylogenet Evol 30:528–535.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol (Amst.) 11:367–372.

Yin, W.-Y. 1996. New considerations on systematics of Protura. Page 60 p. *in* Proceedings of XX International Congress of Entomology, 25âĂŞ31 August, 1996. Firenze, Italy.

Yin, W.-Y., R.-d. Xie, Y.-m. Yang, Q.-y. Yue, and Y.-x. Luan. 2002. Analysis of the main characters for regrouping the class protura. Acta Zootaxonomica Sinica 27:649–658.

Yin, W.-y. and L.-z. Xue. 1993. Comparative spermatology of Protura and its significance on proturan systematics. Sci China, B, Chemistry 36:575–586.

Zantke, J., C. Wolff, and G. Scholtz. 2008. Three-dimensional reconstruction of the central nervous system of *Macrobiotus hufelandi* (Eutardigrada, Parachela): implications for the phylogenetic position of Tardigrada. Zoomorphology 127:21–36.

Zhang, J., C. Zhou, Y. Gai, D. Song, and K. Zhou. 2008. The complete mitochondrial genome of *Parafronurus youi* (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera. Gene 424:18–24.

Zrzavý, J. and P. Štys. 1997. The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. J Evol Biol 10:653–367.

Zwickl, D. J. and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51:588–598.

Zwickl, D. J. and M. T. Holder. 2004. Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics. Syst Biol 53:877–888.

5. Acknowledgements

Herzlich bedanke ich mich bei Bernhard Misof, Thomas Bartolomaeus, Wolfgang Alt, und Jes Rust für die Übernahme der Gutachtertätigkeit und Teilnahme an der Prüfungkommission. Ein besonderer Dank gilt Herrn J. Wolfgang Wägele für die Ermöglichung, Förderung und die konstruktive Begleitung dieser Arbeit am ZFMK und für sein in mich gesetztes Vertrauen.

Mein allergrößter Dank gilt Bernhard Misof, meinem Doktorvater und Betreuer, der mich für die molekulare Phylogenie und Methoden (und Kritik) begeistert und mich immer mit Ideen, Konstruktivität und Vertrauen in jeglicher Hinsicht und mit vielem mehr in allen Lebenslagen unterstützt hat. Bernhard, DANKE! Ich freue mich sehr auf die weitere Zusammenarbeit, das wird genial!

Ein besonderer Dank gilt der besten Kooperationsgruppe, die ich kenne: Günther Pass, Emiliano Dell'Ampio, Nikola Szucsich & Daniela Bartel, sowie Manfred Walzl und Thomas Schwaha, ohne die ich einen Großteil der Arbeit nicht hätte umsetzen können, die mir viel beigebracht haben und mir sehr lieb geworden sind.

Ein Dank von Herzen von der "Kurzen" an Claudia Etzbauer als Labormanagerin und für die Hilfe bei der Bewältigung von Hürden des Instituts-, und sonstigen Alltags. Dem gesamten Molekularlabor-Team des ZFMK danke ich für die großartige Unterstützung, den Autausch und die tolle Atmosphäre.

Danken möchte ich ebenso allen Sammlern (AG Pass, Uni Wien; AG Frati, Uni Siena; R. Machida, Japan; K. Schütte, Uni Hamburg) und den Kooperationspartnern innerhalb des SPP 1174, T. Burmester & Arbeitsgruppe, Uni Hamburg; H. Hadrys & S. Simon, ITZ Hannover; M. Wiens, Uni Mainz, A. v. Haeseler, I. Ebersberger und insbsondere Sascha Strauss, CIBIV, Uni Wien sowie M. Kube & S. Klages, MPIMG, Berlin. Ein besonderer Dank an Anke Braband, mit Vorfreude auf die zukünftige Zusammenarbeit.

Meinem langjährigen Kollegen Björn v. Reumont gilt ein spezieller Dank von der "tiny tante" für die enge, wertvolle, sehr ereignis- und facettenreiche Zusammenarbeit, und die gute Zeit im Rahmen der "crustacean–hexapod–connection", Tagungen, Exkursionen, Aufenthalt in den USA, der Fahrten nach Hause, Schlüsselsuche, Kräuterlikör, Käse etc. mit eingeschlossen.

An Carola Greve, Julia Schwarzer und insbesondere Patrick Kück ("Lieblingskollege") & Daniela Bartel ein riesiger Dank für Hilfe bei der Strukturierung, Formulierung, für Korrekturvorschläge zu jeder Tages- und Nachtzeit, Diskussionen und offene Ohren in jeglicher Hinsicht, besonders im letzten Jahr. Ohne Euch hätte ich vielleicht noch Jahre gebraucht; Revanche kommt.

I thank Karl Kjer, New Brunswick, USA, for the great cooperation, all usefull hints for my work and papers and thanks also to Karl and his family for the wonderful time we had in New Brunswick before and after the SMBE Meeting 2008. Hope to see you soon again!

Herzlicher Dank an Berit Ullrich für die Labor-Einarbeitung und Unterstützung, an "Ex-Diplomanden" Joe Dambach für alles; an Alex Blanke und Manu Thelen für alle Strahlzeiterlebnisse (ich hoffe es kommen noch einige); an Alexi Stamatakis für alle RAxML- und sonstigen Tips; an Katrin Langen &

Jonas Astrin für's Korrekturlesen und an Thomas Stamm & Hubertus Becker für die LaTeX Unterstützung. Ein spezielles Dankeschön an die Korrekturfee Birthe Thormann für das Finden jeglicher Fehler, die andere übersehen haben.

Der Lotusblume Benjamin Meyer möchte ich insbesondere für die großartige (Cliquen–, Matrix– & sonstige) Zusammenarbeit in den letzten zwei Jahren danken, ebenso für die Einführung ins Hamburger Nachtleben. Viel Glück bei der Angstforschung; ich hoffe noch auf viele weitere gemeinsame Cliquen! Janus Borner und Ralph Peters danke ich für den fachlichen Austausch und das gesellige Zusammensein der sehr angenehmen Hamburg-Aufenthalte!

Thomas Stamm gilt mein besonderer Dank für Rückhalt in jedem Gemütszustand.

Schlussendlich bedanke ich mich von Herzen bei meinem Eltern für die Unterstützung und Förderung meiner wissenschaftlicher Begeisterung, und besonders meiner Mutter und Schwester für Rücksicht und Verständnis im vergangenen Jahr. Meinen Katern danke ich für das morgendliche Wecken und das beruhigende Schnurren. Sollte ich jemanden vergessen haben, bitte ich hiermit um Nachsicht.

A. Supplementary Information

A.1. Arthropod Phylogenomics

Some Figures and Tables have been published in Meusemann et al. (2010).

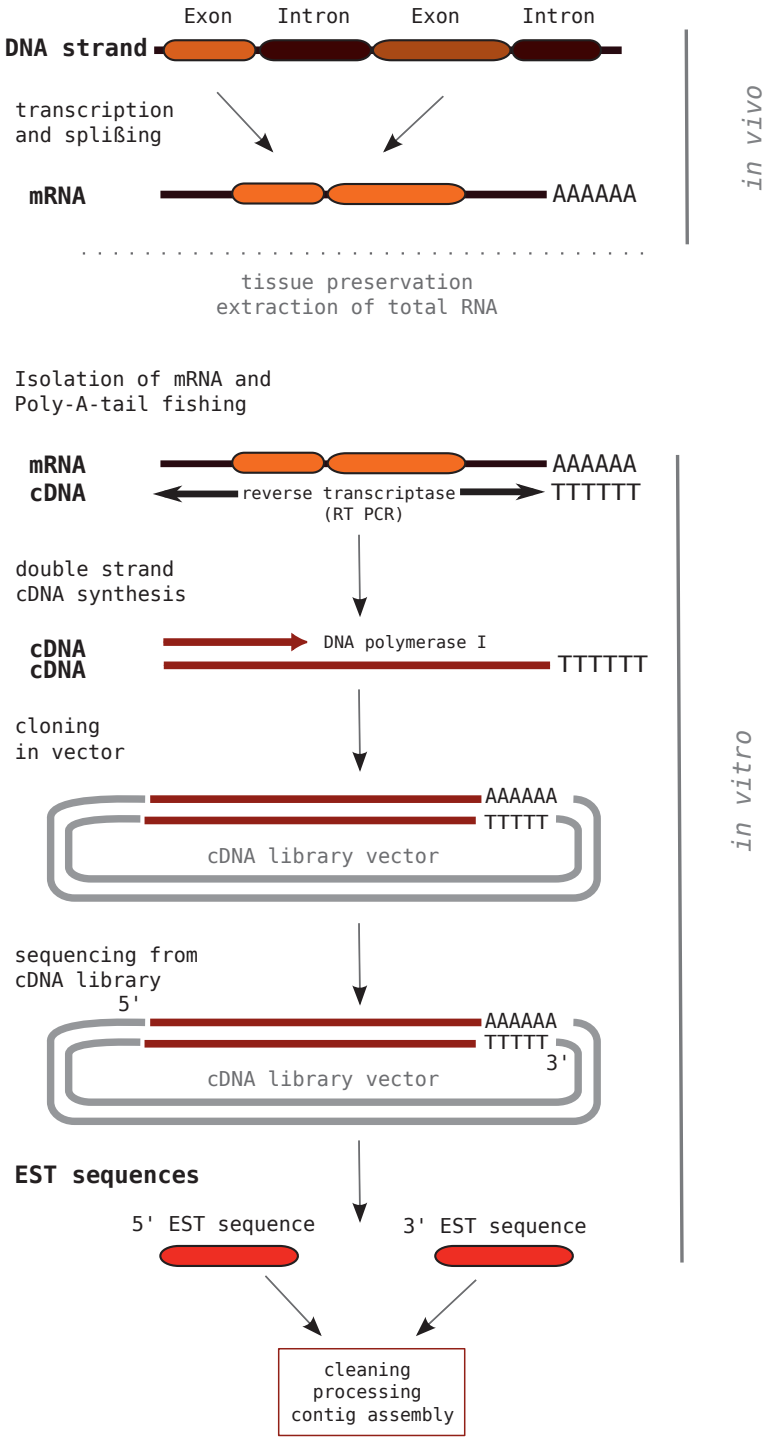

Figure A.1.: Generating ESTs: cloning and sequencing modified after Bouck and Vision (2007). Protein coding DNA regions (exons) of a DNA strand are spliced and transcribed to mRNA. The mRNA is mostly characterized by a poly-A tail and a 5'-end cap with a modified guanine nucleotide. After isolation of total RNA (including mRNA), the poly-A tail is targeted for a Reverse Transcriptase PCR. A cDNA strand from the mRNA template is sythesized. mRNA is digested by adding RNase H. Then, DNA Polymerase 1 synthesizes a second strand, resulting in a double stranded cDNA. The double stranded cDNA is inserted into cloning vectors. Vectors are subsequently inserted into competent cells for the generation of a cDNA library. From this library, Expressed Sequence Tags (EST) sequences can be generated in 5' or 3' end direction.

Table A.1.: New EST data used in this study. Accession no.: Accession numbers; No. of EST rawdata: number of EST reads before processing. Proc. EST sequences: number of ESTs after processing; No. of EST contigs: number of EST contigs after assembling of processed reads into clusters (contigs). EST data except for apterygote hexapods were kindly provided by cooperation partners from the priority programme SPP 1174. RNA extraction with Urea-phenol conducted after Holmes and Bonner (1973). RNA extraction for crustaceans and apterygote hexapods was conducted at the MPI for Genetics (Berlin, Germany) as well as generation of all cDNA libraries, normalization and sequencing. Own data are marked in blue.

Species	Group	Accession no.	RNA extraction	cDNA library construction	No. of EST rawdata	Proc. EST sequences	No. of EST contigs
Peripatopsis sedgwicki	ON	FN232766-FN243241	Urea-phenol	CloneMiner	10,611	10,476	3,452
Endeis spinosa	CH, Pycnogonida	FN211278-FN215339	Urea-phenol	CloneMiner	4,063	4,062	2,672
Limulus polyphemus	CH, Xiphosura	FN224411-FN232765	Urea-phenol	Creator SMART	8,435	8,355	4,050
Archispirostreptus gigas	MY, Diplopoda	FN194820-FN198827	Urea-phenol	Creator SMART	4,032	4,008	2,299
Pollicipes pollicipes	CR, Cirripedia	FN243242-FN247432	Absolutely RNA (Strategene)	CloneMiner	4,224	4,191	1,721
Tigriopus californicus	CR, Copepoda	FN247433-FN252183	Trizol (Invitrogen)	Creator SMART	5,024	5,006	2,598
Triops cancriformis	CR, Branchiopoda	FM868344-FM872274	Trizol (Invitrogen)	Creator SMART	3,981	3,930	2,542
Acerentomon franzi	HE, Protura	FN186135-FN190445	Absolutely RNA (Strategene)	CloneMiner	4,600	4,565	1,995
Campodea cf. *fragilis*	HE, Diplura	FN203025-FN211277	Absolutely RNA (Strategene)	CloneMiner	8,375	8,253	6,407
Anurida maritima	HE, Collembola	FN190447-FN194819	Trizol (Invitrogen)	Creator SMART	4,391	4,373	3,504
Lepismachilis y-signata	HE, Archaeognatha	FN219557-FN224410	Absolutely RNA (Strategene)	CloneMiner	4,895	4,854	2,288
Ischnura elegans[a]	HE, Odonata	FN215340-FN219556	RNAeasy (Quiagen)	Creator SMART	4,219	4,217	3,194
Baetis sp.[a]	HE, Ephemeroptera	FN198828-FN203024	RNAeasy (Quiagen)	Creator SMART	4,225	4,197	3,035

ON: Onychophora; CH: Chelicerata; MY: Myriapoda; CR: Crustacea; HE: Hexapoda. [a] cDNA library: normalized, ESTs were published in (Simon et al., 2009)

Table A.2.: Taxa included in analyses. Species written in capitals – used proteome species; # Taxa included in the optimal data subset selected by reduction heuristics; § Taxa used to train Hidden Markov Models (HMMs) to predict putative orthologs (Ebersberger et al., 2009); Source: dbEST – `http://www.ncbi.nlm.nih.gov/dbEST`, Gene Index Project (gendix) – `http://compbi.dfci.harvard.edu/tgi/tgipage.html`, NCBI Trace Archive – `ftp://ftp.ncbi.nih.gov/pub/TraceDB`, JGI – `http://www.jgi.doe.gov`, InParanoid (Remm et al., 2001; Berglund et al., 2008), v6.14 – `http://inparanoid6.sbc.su.se`, VectorBase – `http://www.vectorbase.org`, BeetleBase – `http://beetlebase.org`, SilkDB – `http://silkworm.genomics.org.cn`, UniProt (integr8) – `http://www.ebi.ac.uk/integr8/`, UCSC – `http://hgdownload.cse.ucsc.edu`. Data of *C. pipiens quinquefasciatus* was kindly provided by the Broad Institute of MIT and Harvard (USA). No. of EST contigs: number of assembled EST contigs; No. of genes orig. data set: number of orthologous genes per taxon in the original data set; No. of genes data subset: number of orthologous genes per taxon in the optimal data subset (SOS) after performing MARE.

Species	Group	Source	No. of EST contigs	No. of genes orig. data set	No. of genes data subset
Hypsibius dujardini#	Panarthropoda, Tardigrada	dbEST	2,386	140	81
Richtersius coronifer#	Panarthropoda, Tardigrada	NCBI Trace Archive	1,537	99	52
Epiperipatus sp. TB-2001	Panarthropoda, Onychophora	dbEST	825	49	
Peripatopsis sedgwicki#	Panarthropoda, Onychophora	present study	3,452	142	72
Euperipatoides kanangrensis#	Panarthropoda, Onychophora	NCBI Trace Archive	1,449	110	53
Julida sp. APV-2005	Euarthropoda, Myriapoda	dbEST	231	13	
Archispirostreptus gigas#	Euarthropoda, Myriapoda	present study	2,299	117	58
Scutigera coleoptrata#	Euarthropoda, Myriapoda	NCBI Trace Archive	807	54	35
Anoplodactylus eroticus#	Euarthropoda, Chelicerata	NCBI Trace Archive	1,281	91	55
Endeis spinosa#	Euarthropoda, Chelicerata	present study	2,672	174	69
Limulus polyphemus#	Euarthropoda, Chelicerata	present study	4,050	210	89
Carcinoscorpius rotundicauda	Euarthropoda, Chelicerata	dbEST	512	21	
Mesobuthus gibbosus	Euarthropoda, Chelicerata	dbEST	587	38	
Loxosceles laeta	Euarthropoda, Chelicerata	dbEST	1,209	66	
Dysdera erythrina	Euarthropoda, Chelicerata	dbEST	279	22	
Cupiennius salei	Euarthropoda, Chelicerata	dbEST	208	30	
Araneus ventricosus	Euarthropoda, Chelicerata	dbEST	204	11	
Acanthoscurria gomesiana#	Euarthropoda, Chelicerata	dbEST	3,713	234	90
Chilobrachys jingzhao	Euarthropoda, Chelicerata	dbEST	230	22	
Ixodes scapularis#	Euarthropoda, Chelicerata	genidx	38,275	578	128
Ixodes ricinus	Euarthropoda, Chelicerata	dbEST	1,300	53	
Amblyomma variegatum#	Euarthropoda, Chelicerata	genidx	2,109	162	62
Amblyomma americanum#	Euarthropoda, Chelicerata	dbEST	2,798	88	44
Amblyomma cajennense	Euarthropoda, Chelicerata	dbEST	1,165	71	
Dermacentor andersoni#	Euarthropoda, Chelicerata	dbEST	752	63	38
Dermacentor variabilis	Euarthropoda, Chelicerata	dbEST	1,075	49	
Boophilus microplus#	Euarthropoda, Chelicerata	dbEST	14,507	425	112
Rhipicephalus appendiculatus#	Euarthropoda, Chelicerata	genidx	7,359	321	92
Argas monolakensis	Euarthropoda, Chelicerata	dbEST	1,620	51	
Ornithodoros porcinus porcinus	Euarthropoda, Chelicerata	dbEST	771	29	
Ornithodoros parkeri	Euarthropoda, Chelicerata	dbEST	689	37	
Ornithodoros coriaceus	Euarthropoda, Chelicerata	dbEST	702	19	
Glycyphagus domesticus#	Euarthropoda, Chelicerata	dbEST	2,511	97	56

Table A.2 continued

Species	Group	Source	No. of EST contigs	No. of genes orig. data set	No. of genes data subset
Blomia tropicalis[#]	Euarthropoda, Chelicerata	dbEST	1,331	80	37
Psoroptes ovis	Euarthropoda, Chelicerata	dbEST	281	18	
Sarcoptes scabiei	Euarthropoda, Chelicerata	dbEST	817	38	
Dermatophagoides pteronyssinus	Euarthropoda, Chelicerata	dbEST	1,258	67	
Dermatophagoides farinae	Euarthropoda, Chelicerata	dbEST	1,046	59	
Suidasia medanensis[#]	Euarthropoda, Chelicerata	dbEST	2,083	139	73
Tyrophagus putrescentiae	Euarthropoda, Chelicerata	dbEST	881	46	
Acarus siro	Euarthropoda, Chelicerata	dbEST	652	57	
Aleuroglyphus ovatus	Euarthropoda, Chelicerata	dbEST	1,440	58	
Gammarus pulex[#]	Euarthropoda, Crustacea	dbEST	4,241	102	63
Eurydice pulchra	Euarthropoda, Crustacea	dbEST	562	26	
Euphausia superba	Euarthropoda, Crustacea	dbEST	1,101	43	
Homarus americanus[#]	Euarthropoda, Crustacea	dbEST	14,147	383	111
Pacifastacus leniusculus	Euarthropoda, Crustacea	dbEST	175	14	
Petrolisthes cinctipes[#]	Euarthropoda, Crustacea	dbEST	27,086	416	119
Callinectes sapidus[#]	Euarthropoda, Crustacea	dbEST	2,239	114	56
Carcinus maenas[#]	Euarthropoda, Crustacea	dbEST	4,567	233	76
Cancer magister	Euarthropoda, Crustacea	dbEST	445	14	
Celuca pugilator	Euarthropoda, Crustacea	dbEST	1,482	64	
Gecarcoidea natalis	Euarthropoda, Crustacea	dbEST	656	23	
Ilyoplax pusilla	Euarthropoda, Crustacea	dbEST	251	2	
Eriocheir sinensis	Euarthropoda, Crustacea	dbEST	1,136	58	
Marsupenaeus japonicus[#]	Euarthropoda, Crustacea	dbEST	1,944	61	46
Fenneropenaeus chinensis[#]	Euarthropoda, Crustacea	dbEST	3,458	114	74
Penaeus monodon[#]	Euarthropoda, Crustacea	dbEST	4,097	129	81
Litopenaeus vannamei[#]	Euarthropoda, Crustacea	dbEST	3,774	126	75
Litopenaeus stylirostris	Euarthropoda, Crustacea	dbEST	314	12	
Litopenaeus setiferus	Euarthropoda, Crustacea	dbEST	642	50	
Tigriopus californicus[#]	Euarthropoda, Crustacea	present study	2,598	65	39
Calanus finmarchicus[#]	Euarthropoda, Crustacea	dbEST	4,906	189	49
Lepeophtheirus salmonis[#]	Euarthropoda, Crustacea	dbEST	5,102	339	98
Pollicipes pollicipes[#]	Euarthropoda, Crustacea	present study	1,721	107	59
Artemia franciscana[#]	Euarthropoda, Crustacea	dbEST	10,330	323	116
Triops cancriformis[#]	Euarthropoda, Crustacea	present study	2,542	115	54
Daphnia magna[#]	Euarthropoda, Crustacea	dbEST	5,307	207	85
DAPHNIA PULEX[#, §]	Euarthropoda, Crustacea	JGI	30,939	775	129
Acerentomon franzi[#]	Euarthropoda, Hexapoda	present study	1,995	99	52
Campodea cf. fragilis[#]	Euarthropoda, Hexapoda	present study	6,407	150	68
Folsomia candida[#]	Euarthropoda, Hexapoda	dbEST	5,955	143	41
Anurida maritima[#]	Euarthropoda, Hexapoda	present study	3,504	131	53
Onychiurus arcticus[#]	Euarthropoda, Hexapoda	dbEST	9,981	309	89
Lepismachilis y-signata[#]	Euarthropoda, Hexapoda	present study	2,288	123	66
Tricholepisma aurea	Euarthropoda, Hexapoda	dbEST	344	34	
Ischnura elegans[#]	Euarthropoda, Hexapoda	Simon et al. 2009	3,194	177	66
Baetis sp.[#]	Euarthropoda, Hexapoda	Simon et al. 2009	3,035	144	49
Locusta migratoria[#]	Euarthropoda, Hexapoda	dbEST	12,255	303	107
Allonemobius fasciatus	Euarthropoda, Hexapoda	dbEST	116	10	
Laupala kohalensis[#]	Euarthropoda, Hexapoda	dbEST	8,371	292	90
Gryllus bimaculatus[#]	Euarthropoda, Hexapoda	dbEST	3,945	238	93
Gryllus pennsylvanicus	Euarthropoda, Hexapoda	dbEST	338	30	

Table A.2 continued

Species	Group	Source	No. of EST contigs	No. of genes orig. data set	No. of genes data subset
Gryllus firmus	Euarthropoda, Hexapoda	dbEST	271	14	
Periplaneta americana[#]	Euarthropoda, Hexapoda	dbEST	1,577	84	58
Blattella germanica[#]	Euarthropoda, Hexapoda	dbEST	1,546	75	38
Diploptera punctata	Euarthropoda, Hexapoda	dbEST	666	20	
Hodotermopsis sjoestedti[#]	Euarthropoda, Hexapoda	dbEST	1,471	73	46
Reticulitermes flavipes	Euarthropoda, Hexapoda	dbEST	113	1	
Sphodromantis centralis	Euarthropoda, Hexapoda	dbEST	120	4	
PEDICULUS HUMANUS[#]	Euarthropoda, Hexapoda	VectorBase	11,198	636	122
Pediculus humanus corporis	Euarthropoda, Hexapoda	dbEST	472	55	
Pediculus humanus capitis	Euarthropoda, Hexapoda	dbEST	2,868	147	
Homalodisca coagulata[#]	Euarthropoda, Hexapoda	dbEST	5,661	237	96
Graphocephala atropunctata[#]	Euarthropoda, Hexapoda	dbEST	1,827	97	63
Oncometopia nigricans[#]	Euarthropoda, Hexapoda	dbEST	1,772	114	63
Lygus lineolaris	Euarthropoda, Hexapoda	dbEST	371	21	
Oncopeltus fasciatus	Euarthropoda, Hexapoda	dbEST	448	11	
Rhodnius prolixus	Euarthropoda, Hexapoda	dbEST	735	48	
Triatoma infestans	Euarthropoda, Hexapoda	dbEST	908	39	
Triatoma brasiliensis	Euarthropoda, Hexapoda	dbEST	1,897	33	
Bemisia tabaci[#]	Euarthropoda, Hexapoda	dbEST	4,548	61	40
Aleurothrixus sp. APV-2005	Euarthropoda, Hexapoda	dbEST	288	18	
Pachypsylla venusta[#]	Euarthropoda, Hexapoda	dbEST	4,631	118	56
Diaphorina citri	Euarthropoda, Hexapoda	dbEST	2,257	66	
Aphis gossypii[#]	Euarthropoda, Hexapoda	dbEST	3,716	210	88
Myzus persicae[#]	Euarthropoda, Hexapoda	dbEST	9,946	447	107
Acyrthosiphon pisum[#]	Euarthropoda, Hexapoda	dbEST	18,253	413	110
Rhopalosiphum padi	Euarthropoda, Hexapoda	dbEST	335	34	
Toxoptera citricida[#]	Euarthropoda, Hexapoda	dbEST	2,196	143	74
Sogatella furcifera	Euarthropoda, Hexapoda	dbEST	122	9	
Nilaparvata lugens	Euarthropoda, Hexapoda	dbEST	167	7	
Maconellicoccus hirsutus[#]	Euarthropoda, Hexapoda	dbEST	3,929	217	85
Nasonia giraulti[#]	Euarthropoda, Hexapoda	dbEST	6,764	277	101
Nasonia vitripennis[#]	Euarthropoda, Hexapoda	dbEST	2,999	160	86
Copidosoma floridanum	Euarthropoda, Hexapoda	dbEST	216	9	
Lysiphlebus testaceipes[#]	Euarthropoda, Hexapoda	dbEST	3,881	210	84
Microctonus hyperodae	Euarthropoda, Hexapoda	dbEST	545	22	
Vespula squamosa[#]	Euarthropoda, Hexapoda	dbEST	1,227	70	50
Solenopsis invicta[#]	Euarthropoda, Hexapoda	dbEST	12,252	297	95
Camponotus festinatus	Euarthropoda, Hexapoda	dbEST	149	8	
Lasius niger	Euarthropoda, Hexapoda	dbEST	347	3	
Bombus ignitus	Euarthropoda, Hexapoda	dbEST	213	22	
APIS MELLIFERA[#, §]	Euarthropoda, Hexapoda	InParanoid	13,448	775	129
Melipona quadrifasciata	Euarthropoda, Hexapoda	dbEST	321	2	
Eoxenos laboulbenei	Euarthropoda, Hexapoda	dbEST	345	32	
Mengenilla chobauti	Euarthropoda, Hexapoda	dbEST	297	27	
Micromalthus debilis	Euarthropoda, Hexapoda	dbEST	157	13	
Carabus granulatus	Euarthropoda, Hexapoda	dbEST	177	16	
Meladema coriacea	Euarthropoda, Hexapoda	dbEST	328	23	
Cicindela litorea	Euarthropoda, Hexapoda	dbEST	232	5	
Cicindela campestris	Euarthropoda, Hexapoda	dbEST	340	24	
Cicindela littoralis	Euarthropoda, Hexapoda	dbEST	236	12	
Sphaerius sp. APV-2005	Euarthropoda, Hexapoda	dbEST	396	29	

Table A.2 continued

Species	Group	Source	No. of EST contigs	No. of genes orig. data set	No. of genes data subset
Eucinetus sp. APV-2005	Euarthropoda, Hexapoda	dbEST	344	27	
Dascillus cervinus	Euarthropoda, Hexapoda	dbEST	354	28	
Georissus sp. APV-2005	Euarthropoda, Hexapoda	dbEST	408	33	
Trox sp. JH-2005	Euarthropoda, Hexapoda	dbEST	223	9	
Scarabaeus laticollis	Euarthropoda, Hexapoda	dbEST	328	30	
Julodis onopordi	Euarthropoda, Hexapoda	dbEST	337	24	
Hister sp. APV-2005	Euarthropoda, Hexapoda	dbEST	358	35	
Agriotes lineatus	Euarthropoda, Hexapoda	dbEST	452	22	
Tenebrio molitor	Euarthropoda, Hexapoda	dbEST	100	3	
TRIBOLIUM CASTANEUM[#, §]	Euarthropoda, Hexapoda	BeetleBase	16,421	775	129
Mycetophagus quadripustulatus	Euarthropoda, Hexapoda	dbEST	419	28	
Biphyllus lunatus	Euarthropoda, Hexapoda	dbEST	260	28	
Hypothenemus hampei	Euarthropoda, Hexapoda	dbEST	844	64	
Diaprepes abbreviatus[#]	Euarthropoda, Hexapoda	dbEST	1,921	65	42
Curculio glandium	Euarthropoda, Hexapoda	dbEST	241	25	
Sitophilus zeamais	Euarthropoda, Hexapoda	dbEST	82	8	
Ips pini	Euarthropoda, Hexapoda	dbEST	565	58	
Platystomus albinus	Euarthropoda, Hexapoda	dbEST	145	5	
Diabrotica virgifera virgifera[#]	Euarthropoda, Hexapoda	dbEST	7,871	336	114
Timarcha balearica	Euarthropoda, Hexapoda	dbEST	272	21	
Leptinotarsa decemlineata[#]	Euarthropoda, Hexapoda	dbEST	2,668	122	56
Callosobruchus maculatus	Euarthropoda, Hexapoda	dbEST	561	58	
Anoplophora glabripennis	Euarthropoda, Hexapoda	dbEST	386	31	
Limnephilus flavicornis	Euarthropoda, Hexapoda	dbEST	117	2	
Hydropsyche sp. T20	Euarthropoda, Hexapoda	dbEST	203	23	
Plutella xylostella[#]	Euarthropoda, Hexapoda	dbEST	1,048	72	55
Tineola bisselliella	Euarthropoda, Hexapoda	dbEST	188	7	
Danaus plexippus[#]	Euarthropoda, Hexapoda	dbEST	9,930	470	114
Bicyclus anynana[#]	Euarthropoda, Hexapoda	dbEST	5,575	165	68
Heliconius erato[#]	Euarthropoda, Hexapoda	dbEST	3,327	219	93
Heliconius melpomene[#]	Euarthropoda, Hexapoda	dbEST	1,820	104	64
Papilio dardanus	Euarthropoda, Hexapoda	dbEST	310	52	
Plodia interpunctella[#]	Euarthropoda, Hexapoda	dbEST	3,808	175	81
Ostrinia nubilalis	Euarthropoda, Hexapoda	dbEST	489	25	
Epiphyas postvittana[#]	Euarthropoda, Hexapoda	dbEST	2,895	154	88
Choristoneura fumiferana	Euarthropoda, Hexapoda	dbEST	589	17	
Trichoplusia ni	Euarthropoda, Hexapoda	dbEST	417	42	
Agrotis segetum	Euarthropoda, Hexapoda	dbEST	812	58	
Spodoptera litura	Euarthropoda, Hexapoda	dbEST	61	3	
Spodoptera frugiperda[#]	Euarthropoda, Hexapoda	dbEST	8,362	309	123
Heliothis virescens[#]	Euarthropoda, Hexapoda	dbEST	1,723	167	73
Helicoverpa armigera[#]	Euarthropoda, Hexapoda	dbEST	692	70	54
Euclidia glyphica	Euarthropoda, Hexapoda	dbEST	187	16	
Bombyx mandarina	Euarthropoda, Hexapoda	dbEST	207	12	
BOMBYX MORI[#, §]	Euarthropoda, Hexapoda	SilkDB	16,329	775	129
Manduca sexta[#]	Euarthropoda, Hexapoda	dbEST	2,197	120	68
Lonomia obliqua	Euarthropoda, Hexapoda	dbEST	610	58	
Samia cynthia ricini[#]	Euarthropoda, Hexapoda	dbEST	5,721	254	105
Antheraea yamamai	Euarthropoda, Hexapoda	dbEST	421	27	

Table A.2 continued

Species	Group	Source	No. of EST contigs	No. of genes orig. data set	No. of genes data subset
Antheraea assama[#]	Euarthropoda, Hexapoda	dbEST	8,927	292	108
Antheraea mylitta[#]	Euarthropoda, Hexapoda	dbEST	1,478	93	58
Panorpa cf. *vulgaris* APV-2005	Euarthropoda, Hexapoda	dbEST	322	21	
Ctenocephalides felis	Euarthropoda, Hexapoda	dbEST	1,775	82	
Xenopsylla cheopis	Euarthropoda, Hexapoda	dbEST	283	26	
Culicoides sonorensis[#]	Euarthropoda, Hexapoda	dbEST	1,405	90	62
Chironomus tentans[#]	Euarthropoda, Hexapoda	dbEST	3,445	216	97
ANOPHELES GAMBIAE[#]	Euarthropoda, Hexapoda	UniProt (integr8)	12,463	726	126
Anopheles aquasalis	Euarthropoda, Hexapoda	dbEST	121	4	
Anopheles darlingi	Euarthropoda, Hexapoda	dbEST	461	24	
Anopheles albimanus[#]	Euarthropoda, Hexapoda	dbEST	3,096	94	53
Anopheles anthropophagus	Euarthropoda, Hexapoda	dbEST	141	5	
Anopheles funestus	Euarthropoda, Hexapoda	dbEST	1,224	59	
AEDES AEGYPTI[#, §]	Euarthropoda, Hexapoda	InParanoid	15,419	654	112
Armigeres subalbatus[#]	Euarthropoda, Hexapoda	NCBI Trace Archive	7,770	329	97
CULEX PIPIENS QUINQUEFASCIATUS[#]	Euarthropoda, Hexapoda	Broad Institute	20,306	721	128
Culex pipiens pallens	Euarthropoda, Hexapoda	dbEST	76	3	
Toxorhynchites amboinensis	Euarthropoda, Hexapoda	dbEST	199	7	
Lutzomyia longipalpis[#]	Euarthropoda, Hexapoda	dbEST	19,739	478	126
Phlebotomus papatasi[#]	Euarthropoda, Hexapoda	dbEST	10,797	422	125
Rhynchosciara americana[#]	Euarthropoda, Hexapoda	dbEST	3,449	112	66
Mayetiola destructor[#]	Euarthropoda, Hexapoda	dbEST	1,482	81	48
Sitodiplosis mosellana	Euarthropoda, Hexapoda	dbEST	1,100	64	
Orseolia oryzae	Euarthropoda, Hexapoda	dbEST	976	29	
Glossina morsitans morsitans[#]	Euarthropoda, Hexapoda	dbEST	12,444	512	124
Musca domestica	Euarthropoda, Hexapoda	dbEST	296	14	
Stomoxys calcitrans	Euarthropoda, Hexapoda	dbEST	296	31	
Haematobia irritans	Euarthropoda, Hexapoda	dbEST	196	13	
Haematobia irritans irritans	Euarthropoda, Hexapoda	dbEST	189	13	
Ceratitis capitata[#]	Euarthropoda, Hexapoda	dbEST	11,132	475	123
Rhagoletis suavis	Euarthropoda, Hexapoda	dbEST	370	27	
Rhagoletis pomonella	Euarthropoda, Hexapoda	dbEST	160	7	
Drosophila arizonae[#]	Euarthropoda, Hexapoda	dbEST	770	88	55
DROSOPHILA ANANASSAE[#]	Euarthropoda, Hexapoda	UCSC	29,704	673	113
DROSOPHILA ERECTA[#]	Euarthropoda, Hexapoda	UCSC	17,531	673	117
DROSOPHILA MELANOGASTER[#, §]	Euarthropoda, Hexapoda	InParanoid	13,854	752	129
Meloidogyne hapla[#]	Nematoda	dbEST	7,802	252	92
CAENORHABDITIS ELEGANS[#, §]	Nematoda	InParanoid	20,084	749	127
CAENORHABDITIS REMANEI[#]	Nematoda	InParanoid	25,595	719	126
CAENORHABDITIS BRIGGSAE[#, §]	Nematoda	InParanoid	19,334	711	126
Haemonchus contortus[#]	Nematoda	dbEST	5,842	262	98
Ascaris suum[#]	Nematoda	dbEST	9,165	197	84
Xiphinema index[#]	Nematoda	dbEST	4,824	228	89
Trichinella spiralis[#]	Nematoda	dbEST	8,843	373	111
CAPITELLA CAPITATA[#, §]	Annelida	JGI	32,415	724	122
HELOBDELLA ROBUSTA[#]	Annelida	JGI	23,432	730	126
Lumbricus rubellus[#]	Annelida	dbEST	10,386	196	94
LOTTIA GIGANTEA[#, §]	Mollusca	JGI	23,851	672	120
Crassostrea gigas[#]	Mollusca	dbEST	14,857	339	102
Argopecten irradians[#]	Mollusca	dbEST	3,610	95	59

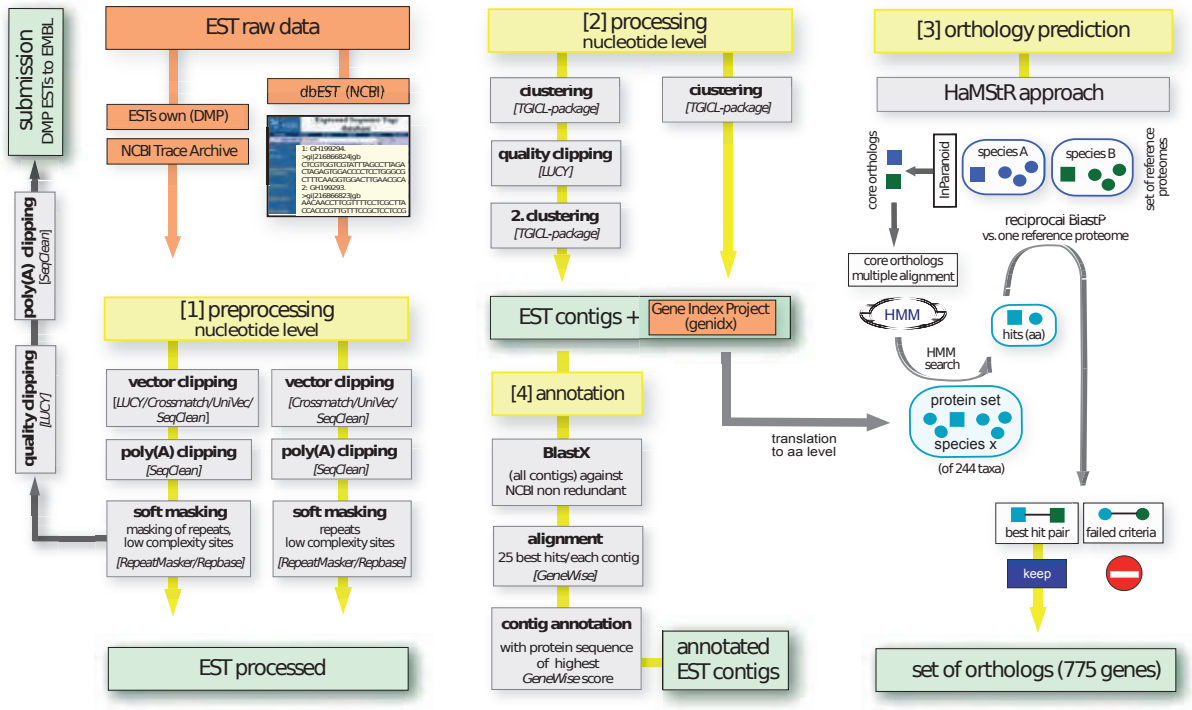

Figure A.2.: Processing of EST data, orthology assignment and annotation. EST raw data of own and published EST projects were mined and processed in four major steps: preprocessing, processing, orthology prediction and annotation. (1) In the preprocessing, own EST sequences were screened for vectors and poly(A) tails using LUCY (Chou and Holmes, 2001). All sequences including published ESTs were screened for contamination by comparison against UniVec (`http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html`) with Crossmatch (Green, 1996) and SeqClean (Pertea, 2005-2006), which screens for poly(A) tails as well. Sequences < 100 nucleotides were discarded. Repetitive elements in the remaining ESTs were soft masked with RepeatMasker (Smit et al., 1996-2004) using Repbase (Jurka et al., 2005). (2) ESTs were clustered using the TGICL (Pertea et al., 2003). ESTs of the SPP 1174 were quality clipped with LUCY and clustered again to obtain and keep longer sequences for the contig assembly. Contigs were translated into amino acid level and (3) integrated in HaMStR (Ebersberger et al., 2009). A set of of reference proteomes was compiled from InParanoid with *Daphnia pulex*, *Tribolium castaneum*, *Bombyx mori*, *Apis melifera* (mandatory) and *Aedes aegypti*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Capitella* sp., *Lottia gigantea*, *Homo sapiens*, *Tetradon nigroviridis* and *Xenopus tropicalis* as 'primer' taxa. Thereby, at leat one taxon was represented in the core ortholog set from non-mandatory taxa. Multiple alignments of the core ortholog set were used to build profile Hidden Markov Models (hmmer-package, `http://hmmer.janelia.org/`) to search in each EST set of 244 taxa for hits. A reciprocal BlastP (Altschul et al., 1997) decided about the survival of a hit. For the re-blast step, the proteome of the presumably evolutionary closest primer taxon for each considered species was chosen. This approach upended in a set of 775 orthologs. (4) EST contigs were annotated using a BlastX search against NCBI's non-redundant protein database. The protein sequences of the 25 best hits for each contig were aligned with GeneWise (Birney et al., 2004). The contig was annotated according to the protein sequence with the highest GeneWise score. Single EST reads were submitted to EMBL (Tab. A.1).

Table A.3.: Genes selected by HaMStR and used in phylogenetic analyses. Gene ID – numerical internal identifier corresponding to gene number (partition number) of the raw data matrix; Protein ID – FlyBase-ID from Ensembl Archive February 2007 (Arch. 02/07) for *Drosophila melanogaster* (`http://feb2007.archive.ensembl.org/` respectively AEE-ID from Inparanoid4 v6.1 for *Aedes aegypti* `http://inparanoid6.sbc.su.se`); Gene / Description – Description of genes as determined from the Ensemble Archive / Flybase for *D. melanogaster* (Dmel), from InParanoid v6.1 for *A. aegypti* (Aaeg) or from HomoloGene for *Homo sapiens* (Hsap), `http://www.ncbi.nlm.nih.gov/homologene`; other studies – genes shared with other studies: ph – Philippe et al. (2009); de – Delsuc et al. (2008); du – Dunn et al. (2008); ba – Baurain et al. (2007); genes are assigned to the gene name in squared brackets used in previous studies; Rib. Protein – ribosomal protein coding gene (x); Pot. rel. info. content – information content calculated by MARE; No. of taxa in data set – number of taxa in the raw data set; present in data subset – this indicates the presence of respective genes in the optimal selected data subset (SOS); No. of taxa in data subset – number of taxa in the optimal data subset (SOS).

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
12061	FBpp0078222	ADP-ribosylation factor 1				0.92	107	x	88
11924	FBpp0081153	Tubulin alpha-1 chain				0.92	149	x	102
11899	FBpp0078664	26S proteasome non-ATPase regulatory subunit 14				0.91	70	x	61
11735	FBpp0076890	26S protease regulatory subunit 8		ph, de, ba [nsf1-G]		0.90	81	x	76
11806	FBpp0081524	Beta-2 tubulin				0.90	127	x	96
11491	FBpp0083502	AP-2	clathrin coat assembly protein ap17			0.90	51	x	46
11394	FBpp0073292	Rpt3	26S protease regulatory subunit 6b *Proteasome 26S subunit ATPase 4*	de, ba [nsf1-L]		0.89	70	x	66
12024	FBpp0083645	AP-50, isoform A	clathrin coat associated protein ap-50			0.89	56	x	54
11846	FBpp0082140	Vacuolar ATP synthase subunit B		ph, de, ba [vatb]		0.89	74	x	69
11362	FBpp0084434	Histone H2A				0.88	80	x	70
11958	FBpp0078984	smt3				0.87	74	x	62
11637	FBpp0086701	40S ribosomal protein S23		ph, de, ba [rps23]	x	0.86	142	x	95
11460	FBpp0083906	26S protease regulatory subunit 4		ph, de, ba [nsf1-M]		0.86	74	x	69
11547	FBpp0085265	Elongation factor 2		ph, de, ba [ef2-EF2]		0.86	71	x	65
12071	FBpp0088250	ATP synthase beta chain, mitochondrial precursor				0.86	119	x	95
11624	FBpp0081592	AP-47	clathrin coat assembly protein ap-1			0.85	58	x	55
11511	FBpp0076145	CG6767-PB, isoform B	ribose-phosphate pyrophosphokinase 1			0.85	61	x	59
11609	FBpp0083843	Tat-binding protein-1	26S protease regulatory subunit 6a	ph, de, ba [nsf1-K]		0.85	78	x	71
11484	FBpp0088174	CG1970-PA	NADH-ubiquinone oxidoreductase fe-s protein 2 (ndufs2)			0.84	77	x	67
11902	FBpp0087084	GTP-binding protein 128up				0.84	58	x	55
11442	FBpp0077792	Splicing factor U2af 38 kDa subunit		de [u2snrnp]		0.83	58	x	53
11552	FBpp0071808	60S ribosomal protein L23		ph, du, de, ba [rpl23a]	x	0.83	127	x	91
11760	FBpp0074520	Cdc42 homolog	rac GTPase			0.82	68	x	60
11645	FBpp0073446	Heat shock 70 kDa protein cognate 3 precursor		ph, de, ba [hsp70-E]		0.82	64	x	58
11868	FBpp0079999	Vacuolar ATP synthase catalytic subunit A isoform 2		ph, de, ba [vata]		0.82	51	x	48
11762	FBpp0078847	CG9140-PA	NADH-ubiquinone oxidoreductase flavoprotein 1 (ndufv1)			0.81	71	x	63

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11377	FBpp0082724	SF2	arginine/serine-rich splicing factor			0.80	47	x	46
11759	FBpp0081401	CG8351-PA	chaperonin	ph, de, ba [cct-N]		0.78	64	x	60
11635	FBpp0080639	40S ribosomal protein S26		ph, de, ba [rps26]	x	0.78	124	x	93
11983	FBpp0082535	Tropomyosin-2				0.77	129	x	98
11617	FBpp0079992	CG5525-PA	chaperonin *T-complex protein 1 subunit delta*	ph, de, ba [cct-D]		0.77	73	x	67
11639	FBpp0085586	40S ribosomal protein S18		ph, du, de, ba [rps18]	x	0.77	136	x	97
11366	FBpp0077571	Enolase				0.77	94	x	78
11634	FBpp0083684	T-complex protein 1 subunit alpha	chaperonin	ph, de, ba [cct-A]		0.76	64	x	62
11393	FBpp0075700	Eukaryotic translation initiation factor 2 beta subunit		ph, de, ba [if2b]		0.76	70	x	60
11379	FBpp0071226	CG7033-PB, isoform B	chaperonin	ph, de, ba [cct-B]		0.76	66	x	63
11893	FBpp0072197	26S proteasome non-ATPase regulatory subunit 7				0.76	62	x	58
12097	FBpp0085919	Polyadenylate-binding protein				0.76	60	x	54
11514	FBpp0074180	40S ribosomal protein S5a		ph, ba [rps5]	x	0.75	148	x	108
11750	FBpp0073328	GTP-binding nuclear protein Ran				0.75	88	x	77
11681	FBpp0082459	CG3731-PB, isoform B	mitochondrial processing peptidase beta subunit	du [rpl27]		0.75	88	x	79
11874	FBpp0079187	Guanine nucleotide-binding protein beta subunit-like protein				0.75	133	x	104
11962	FBpp0080495	Vacuolar ATP synthase subunit H				0.75	66	x	61
11965	FBpp0077142	60S ribosomal protein L27a		ph, de, ba [rpl27]	x	0.75	136	x	95
11450	FBpp0086603	Proteasome p44.5 subunit, isoform B				0.75	76	x	70
11917	FBpp0082464	VhaPPA1-1	vacuolar ATP synthase proteolipid subunit			0.74	71	x	66
11411	FBpp0082516	Heat shock 70 kDa protein cognate 4				0.74	109	x	92
11660	FBpp0078024	26S proteasome non-ATPase regulatory subunit 4				0.74	62	x	61
11385	FBpp0088565	Eukaryotic initiation factor 3 p66 subunit				0.74	72	x	66
11642	FBpp0077419	Phosphoglycerate kinase				0.74	79	x	72
11695	FBpp0077741	lesswright, isoform A	ubiquitin-conjugating enzyme E2 i			0.74	65	x	60
11587	FBpp0073626	40S ribosomal protein S15Aa		ph, de, ba [rps22a]	x	0.73	119	x	91
11829	FBpp0071794	ATP synthase alpha chain, mitochondrial precursor				0.73	103	x	92
12012	FBpp0074825	Catalase				0.73	63	x	61
12121	FBpp0086269	Ribosomal protein S15, isoform B		ph, du, de, ba [rps15]	x	0.73	133	x	97
11479	FBpp0081234	Probable small nuclear ribonucleoprotein Sm D2		du [small nuclear ribonucleo-protein polypeptide D2]		0.72	57	x	50
11798	FBpp0085483	Vacuolar ATP synthase 16 kDa proteolipid subunit				0.72	98	x	81
11848	FBpp0086468	Vacuolar ATP synthase subunit D 1				0.72	75	x	62
11627	FBpp0080691	Probable 26S proteasome non-ATPase regulatory subunit 3				0.72	65	x	62
11454	FBpp0073847	Adenosylhomocysteinase		ph, de [Sadhchydrolase-E1]		0.72	94	x	81
12054	FBpp0077637	CG5001-PA	DNA-J/hsp40			0.72	65	x	60
11911	FBpp0078134	60S acidic ribosomal protein P0		ph, de, ba [rpp0]	x	0.72	148	x	108
12019	FBpp0076393	Isocitrate dehydrogenase, isoform F				0.71	79	x	66
11375	FBpp0077716	60S acidic ribosomal protein P1		du, de, ba [rla2-B]	x	0.71	134	x	89
11855	FBpp0082571	Surfeit locus protein 4 homolog				0.71	60	x	57
11583	FBpp0082788	T-complex protein 1 subunit gamma		ph, de, ba [cct-G]		0.71	57	x	55
11563	FBpp0081581	Calreticulin precursor				0.70	100	x	87
11429	FBpp0086381	CG8446-PA	*lipoyltransferase 1*			0.69	57	x	55

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
12033	FBpp0084585	CG5590-PA	short-chain dehydrogenase			0.69	72	x	65
11437	FBpp0078532	CG9769-PA	eukaryotic translation initiation factor 3f eif3f			0.69	73	x	62
11577	FBpp0082062	Proteasome subunit alpha type 2		ph, de, ba [psma-D]		0.69	64	x	54
11534	FBpp0071451	Proteasome subunit alpha type 4				0.69	76	x	64
11591	FBpp0072968	CG32276-PB, isoform B	*stress-associated endoplasmic reticulum protein family member 2*			0.68	104	x	76
11603	FBpp0077740	Signal peptide protease				0.68	58	x	57
11910	FBpp0072801	60S ribosomal protein L8		ph, du, de, ba [rpl2]	x	0.67	134	x	101
11802	Fbpp0071279	Oligosaccharyltransferase 48kD subunit	Dolichyl-diphosphooligosaccharide protein glycosyltransferase			0.67	68	x	64
11555	FBpp0080395	CaBP1	protein disulfide-isomerase A6 precursor			0.67	72	x	64
12120	FBpp0081780	Arginine methyltransferase 1				0.67	67	x	62
11814	Fbpp0076960	CG1532-PA	lactoylglutathione lyase			0.66	61	x	56
11567	FBpp0086066	Proteasome subunit alpha type 5		ph, de, ba [psma-A]		0.66	66	x	61
11451	FBpp0076152	40S ribosomal protein S9		ph, ba [rps9]	x	0.66	118	x	86
11710	FBpp0080724	Ribosomal protein L30, isoform A		ph, du, de, ba [rpl30]	x	0.66	117	x	88
11580	FBpp0110423	ribosomal protein L5		ph, de, ba [rpl5]	x	0.65	133	x	105
11976	FBpp0085489	Succinate dehydrogenase [ubiquinone] iron-sulfur protein, mitochondrial precursor				0.65	73	x	64
12029	FBpp0082985	CG7998-PA	malate dehydrogenase			0.65	86	x	75
11849	FBpp0072312	60S ribosomal protein L19		ph, du, de, ba [rpl19a]	x	0.65	134	x	93
12047	FBpp0084901	CG7834-PB, isoform B	electron transfer flavoprotein beta-subunit			0.65	73	x	68
11350	FBpp0076859	Uev1A, isoform B	ubiquitin-conjugating enzyme			0.64	62	x	58
11386	FBpp0079640	CG5362-PA	malate dehydrogenase			0.64	90	x	78
12112	FBpp0072250	Inorganic pyrophosphatase				0.63	71	x	59
11428	FBpp0073989	Proteasome subunit alpha type 7-1				0.63	72	x	65
11711	FBpp0075766	60S ribosomal protein L10a-2		ph, de, ba [rpl1]	x	0.63	133	x	103
11499	FBpp0070430	CG8636-PA	eukaryotic translation initiation factor *eukaryotic translation initiation factor 3 subunit 4*	du [eukaryotic translation initiation factor 3, subunit 4 delta]		0.63	81	x	71
11652	FBpp0085889	Eip55E	cystathionine beta-lyase			0.63	72	x	63
11378	FBpp0079472	yippee interacting protein 2				0.63	79	x	71
11919	FBpp0083371	40S ribosomal protein S20		ph, du, de, ba [rps20]	x	0.63	135	x	95
11772	FBpp0087186	walrus, isoform B	electron transport oxidoreductase			0.62	68	x	61
11928	FBpp0070047	60S ribosomal protein L10		ph, de, ba [grc5]	x	0.62	155	x	109
11932	FBpp0070871	Lethal (1), isoform A	citrate synthase			0.62	70	x	65
11380	FBpp0084617	60S ribosomal protein L4		ph, de, ba [rpl4B]	x	0.62	134	x	107
11820	FBpp0076804	Thioredoxin-like				0.61	75	x	69
11707	FBpp0088441	40S ribosomal protein S7		ph [rps7]	x	0.61	134	x	103
11663	FBpp0088522	Ubiquitin conjugating enzyme 10				0.61	69	x	62
11912	FBpp0074599	Clathrin light chain				0.61	79	x	68
12046	FBpp0088505	Annexin-B9				0.61	76	x	64
11992	FBpp0087164	Erp60, isoform B	protein disulfide isomerase			0.61	101	x	87
11754	FBpp0071766	40S ribosomal protein S16		ph, du, de, ba [rps16]	x	0.60	138	x	101
11984	FBpp0100039	Voltage-dependent anion-selective channel				0.60	104	x	84
11773	FBpp0075382	Proteasome 2 subunit		ph, de, ba [psmb-K]		0.60	92	x	81

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11847	FBpp0088242	40S ribosomal protein S3a		ph, de, ba [rps1]	x	0.59	139	x	103
11649	FBpp0076848	CG4769-PA	cytochrome C1			0.58	93	x	82
12040	FBpp0084306	Ribosomal protein L27			x	0.58	129	x	87
12035	FBpp0073344	Glutamine synthetase 2, cytoplasmic				0.57	90	x	79
12115	FBpp0087972	cathD	cathepsin d			0.57	96	x	81
12093	FBpp0082645	NADH:ubiquinone reductase 23kD subunit precursor				0.57	70	x	66
12081	FBpp0084762	Elongation factor 1-gamma				0.56	137	x	109
11619	FBpp0086103	60S ribosomal protein L18a		ph, du, de, ba [rpl20]	x	0.56	137	x	97
11869	FBpp0077580	Rieske iron-sulfur protein, isoform B	ubiquinol-cytochrome c reductase iron-sulfur subunit	du [Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide I]		0.56	97	x	80
11391	FBpp0075618	40S ribosomal protein S4		ph, ba [rps4]	x	0.55	144	x	105
11584	FBpp0081488	Proteasome subunit beta type 3		ph, du, de, ba [psmb-I]		0.54	81	x	72
11383	FBpp0086973	Nascent polypeptide-associated complex alpha subunit				0.53	94	x	81
11778	FBpp0087608	60S ribosomal protein L31		ph, de, ba [rpl31]	x	0.53	122	x	89
11844	FBpp0099686	40S ribosomal protein S8		ph, du, ba [rps8]	x	0.53	152	x	104
11618	FBpp0085166	Ribosomal protein L6, isoform B		ph, de, ba [rpl6]	x	0.47	145	x	106
11841	FBpp0110173	hydrogen-transporting ATP synthase, G-subunit, putative				0.46	110	x	78
12122	FBpp0078354	60S ribosomal protein L13A			x	0.45	136	x	98
12123	FBpp0083376	Ribosomal protein S30, isoform B		du [Ubiquitin-like FUBI and ribo-ribosomal protein S30 precursor]	x	0.44	136	x	94
12074	FBpp0072084	CG3195-PA, isoform A	60S ribosomal protein L12	ph, du, de, ba [rpl12b]	x	0.44	136	x	100
11793	FBpp0076602	Ribosomal protein L18		ph, du, de, ba [rpl18]	x	0.42	124	x	95
11417	FBpp0087352	Ras-related protein Rab-3				0.88	34		
11816	Fbpp0079447	Pka-C1: cAMP-dependent protein kinase catalytic subunit	cAMP-dependent protein kinase catalytic subunit			0.87	44		
11387	FBpp0075260	diablo				0.86	36		
12009	FBpp0088695	CG2944-PF, isoform F	*splA/ryanodine receptor domain and SOCS box containing 4*			0.84	30		
11405	FBpp0077302	Protein mothers against dpp				0.83	29		
12073	FBpp0074756	reptin				0.83	46		
11755	FBpp0083248	CG10889-PA	*zinc finger CCCH-type containing 12B*			0.82	21		
11783	FBpp0079615	Transcription initiation factor IIB				0.81	46		
11860	FBpp0070208	SNF1A/AMP-activated protein kinase, isoform B				0.81	35		
11803	FBpp0079634	CG5343-PA	orf protein			0.81	47		
12118	FBpp0099616	cAMP-dependent protein kinase type I regulatory subunit				0.80	52		
11398	FBpp0088599	Potassium voltage-gated channel protein Shaker	voltage-gated potassium channel			0.80	19		
11954	FBpp0079565	Putative ATP-dependent RNA helicase me31b				0.80	43		
11509	FBpp0087094	Small nuclear ribonucleoprotein SM D3				0.79	43		
12087	FBpp0082743	COP9 signalosome complex subunit 5				0.79	46		
11865	FBpp0070361	Unc-76, isoform B				0.79	31		
11680	FBpp0088583	CG11266-PG, isoform G	splicing factor			0.78	49		
11989	FBpp0083135	CG5451-PA	WD-repeat protein			0.78	37		
11797	AAEL007662-PA	casein kinase				0.78	46		
12084	FBpp0079951	Ef1-like factor				0.78	28		
11850	FBpp0099884	UGP, isoform A				0.77	49		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11496	FBpp0078469	Katanin 60		de [nsf1-N]		0.77	34		
11687	FBpp0084528	CG5934-PA				0.76	36		
12070	AAEL005833-PA	cytosolic purine 5-nucleotidase				0.76	36		
11956	FBpp0086942	Guanine nucleotide-binding protein G(q) subunit alpha				0.76	36		
11616	FBpp0086375	Lissencephaly-1 homolog				0.76	46		
11673	FBpp0083973	Syntaxin-1A				0.75	38		
11991	FBpp0083588	CG6439-PA	isocitrate dehydrogenase			0.75	56		
12060	FBpp0074486	6-phosphofructo-2-kinase, isoform I				0.75	41		
11384	FBpp0081448	CG11990-PA	cdc73 domain protein			0.75	30		
11542	FBpp0080659	Sterol carrier protein X-related thiolase				0.74	54		
11763	FBpp0072052	Guanine nucleotide-binding protein G(s), alpha subunit				0.74	27		
11700	FBpp0079629	RluA-1, isoform C				0.74	26		
11589	FBpp0083573	Probable ATP-dependent RNA helicase pitchoune				0.74	38		
11940	FBpp0085430	CG10465-PA	*potassium channel tetramerisation domain containing 10*			0.74	33		
12065	FBpp0110163	CAMP-dependent protein kinase catalytic subunit				0.74	50		
12007	FBpp0074691	tricornered				0.74	28		
11864	FBpp0073387	DNA-directed RNA polymerase II largest subunit		du [DNA directed RNA polymerase II polypeptide C]		0.74	15		
11921	FBpp0085737	Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial precursor				0.73	35		
11406	FBpp0083112	endophilin A, isoform B				0.73	30		
11726	FBpp0088499	Protein ariadne-1				0.73	42		
11640	FBpp0071553	CG4279-PA	Sm protein G putative			0.73	42		
11564	FBpp0085131	CG31005-PA	trans-prenyltransferase			0.73	32		
11508	FBpp0080261	Suppressor of hairless protein				0.73	18		
11788	FBpp0110435	synaptosomal associated protein				0.73	38		
11672	FBpp0071424	Inosine-5'-monophosphate dehydrogenase				0.73	46		
11610	FBpp0070859	Spliceosomal protein on the X				0.73	28		
11980	FBpp0085902	GTP-binding-protein				0.72	46		
11523	FBpp0078624	CG14641-PA	RNA binding motif protein			0.72	37		
11990	FBpp0081290	ADP-ribosylation factor-like protein 8				0.72	49		
11768	FBpp0074278	CG6842-PA	skd/vacuolar sorting			0.72	43		
11934	FBpp0070250	CG32810-PB	*potassium channel tetramerisation domain containing 5*			0.72	34		
11578	FBpp0071600	Rae1				0.72	46		
11436	FBpp0084036	atlastin, isoform B				0.72	41		
11821	FBpp0070883	Serine/threonine-protein phosphatase PP-V		de [stcproptase2a-c]		0.72	47		
11490	FBpp0081483	Aryl hydrocarbon receptor nuclear translocator homolog				0.72	20		
11796	FBpp0081704	pontin				0.71	34		
11549	FBpp0076078	Ard1, isoform A				0.70	51		
12038	FBpp0082507	CG4203-PA	*KIAA0892*			0.70	26		
12088	FBpp0079676	Stress-activated protein kinase JNK				0.70	26		
11718	FBpp0086599	CG32105-PB	*LIM homeobox transcription factor 1, alpha*			0.70	20		
11785	FBpp0080801	Tyrosine-protein phosphatase Lar precursor				0.70	17		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11496	FBpp0078469	Katanin 60		de [nsf1-N]		0.77	34		
11687	FBpp0084528	CG5934-PA				0.76	36		
12070	AAEL005833-PA	cytosolic purine 5-nucleotidase				0.76	36		
11956	FBpp0086942	Guanine nucleotide-binding protein G(q) subunit alpha				0.76	36		
11616	FBpp0086375	Lissencephaly-1 homolog				0.76	46		
11673	FBpp0083973	Syntaxin-1A				0.75	38		
11991	FBpp0083588	CG6439-PA	isocitrate dehydrogenase			0.75	56		
12060	FBpp0074486	6-phosphofructo-2-kinase, isoform I				0.75	41		
11384	FBpp0081448	CG11990-PA	cdc73 domain protein			0.75	30		
11542	FBpp0080659	Sterol carrier protein X-related thiolase				0.74	54		
11763	FBpp0072052	Guanine nucleotide-binding protein G(s), alpha subunit				0.74	27		
11700	FBpp0079629	RluA-1, isoform C				0.74	26		
11589	FBpp0083573	Probable ATP-dependent RNA helicase pitchoune				0.74	38		
11940	FBpp0085430	CG10465-PA	*potassium channel tetramerisation domain containing 10*			0.74	33		
12065	FBpp0110163	CAMP-dependent protein kinase catalytic subunit				0.74	50		
12007	FBpp0074691	tricornered				0.74	28		
11864	FBpp0073387	DNA-directed RNA polymerase II largest subunit		du [DNA directed RNA polymerase II polypeptide C]		0.74	15		
11921	FBpp0085737	Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial precursor				0.73	35		
11406	FBpp0083112	endophilin A, isoform B				0.73	30		
11726	FBpp0088499	Protein ariadne-1				0.73	42		
11640	FBpp0071553	CG4279-PA	Sm protein G putative			0.73	42		
11564	FBpp0085131	CG31005-PA	trans-prenyltransferase			0.73	32		
11508	FBpp0080261	Suppressor of hairless protein				0.73	18		
11788	FBpp0110435	synaptosomal associated protein				0.73	38		
11672	FBpp0071424	Inosine-5'-monophosphate dehydrogenase				0.73	46		
11610	FBpp0070859	Spliceosomal protein on the X				0.73	28		
11980	FBpp0085902	GTP-binding-protein				0.72	46		
11523	FBpp0078624	CG14641-PA	RNA binding motif protein			0.72	37		
11990	FBpp0081290	ADP-ribosylation factor-like protein 8				0.72	49		
11768	FBpp0074278	CG6842-PA	skd/vacuolar sorting			0.72	43		
11934	FBpp0070250	CG32810-PB	*potassium channel tetramerisation domain containing 5*			0.72	34		
11578	FBpp0071600	Rae1				0.72	46		
11436	FBpp0084036	atlastin, isoform B				0.72	41		
11821	FBpp0070883	Serine/threonine-protein phosphatase PP-V		de [stcproptase2a-c]		0.72	47		
11490	FBpp0081483	Aryl hydrocarbon receptor nuclear translocator homolog				0.72	20		
11796	FBpp0081704	pontin				0.71	34		
11549	FBpp0076078	Ard1, isoform A				0.70	51		
12038	FBpp0082507	CG4203-PA	*KIAA0892*			0.70	26		
12088	FBpp0079676	Stress-activated protein kinase JNK				0.70	26		
11718	FBpp0086599	CG32105-PB	*LIM homeobox transcription factor 1, alpha*			0.70	20		
11785	FBpp0080801	Tyrosine-protein phosphatase Lar precursor				0.70	17		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11507	FBpp0081840	CG17184-PB, isoform B	*ADP-ribosylation factor interacting protein 2*			0.65	27		
11588	FBpp0087472	CG12140-PA	electron transfer flavoprotein-ubiquinone oxidoreductase			0.65	34		
11728	FBpp0099695	Dystrobrevin-like, isoform A				0.65	19		
11716	FBpp0071992	no extended memory, isoform B				0.65	31		
12018	FBpp0083611	Pyruvate kinase				0.65	77		
11830	FBpp0087865	Rs1				0.65	29		
12076	FBpp0078099	CG7145-PD, isoform D	pyrroline-5-carboxylate dehydrogenase			0.65	57		
11556	FBpp0079843	CG14939-PA	*cyclin Y*			0.65	31		
11553	FBpp0071285	Puff-specific protein Bx42				0.65	43		
11955	FBpp0110314	conserved hypothetical protein				0.65	36		
11667	FBpp0077214	CG17593-PA	*coiled-coil domain containing 47*			0.65	49		
12069	FBpp0071046	Protein bys				0.65	40		
11709	FBpp0074022	CG9911-PA, isoform A	endoplasmic reticulum resident protein (ERp44) putative			0.65	44		
11389	FBpp0081988	Putative inner dynein arm light chain	axonemal inner arm dynein light chain			0.65	24		
12051	FBpp0072144	Probable eukaryotic translation initiation factor 6		ph, de, ba [if6]		0.65	60		
11525	FBpp0086098	eIF3-S9, isoform B				0.64	76		
11432	FBpp0079642	CG33303-PA	ribophorin			0.64	68		
11712	FBpp0070249	CG14782-PA	*pleckstrin homology domain containing, family F (with FYVE domain) member 2*			0.64	31		
11905	FBpp0078433	DNA-directed RNA polymerases I, II, and III 14.4 kDa polypeptide				0.64	47		
11598	FBpp0075202	CG5284-PA, isoform A	chloride channel protein 3			0.64	26		
11853	FBpp0081617	CG8500-PA	MRAS2 putative			0.64	19		
11890	FBpp0086340	mrj, isoform D				0.64	49		
11801	FBpp0072419	Tudor-SN	ebna2 binding protein P100			0.64	55		
11455	FBpp0082728	belphegor				0.64	34		
12057	FBpp0076921	lethal (1) G0269				0.64	25		
11574	FBpp0077676	Clipper				0.64	30		
11971	FBpp0070873	Transmembrane GTPase Marf				0.64	39		
11891	FBpp0081958	CG18347-PA	mitochondrial glutamate carrier protein			0.64	32		
11786	FBpp0084191	CG11859-PA	serine/threonine-protein kinase rio2 (rio kinase 2)			0.64	35		
11629	FBpp0079617	CHIP				0.63	46		
11632	FBpp0078997	nop5				0.63	49		
11608	FBpp0078606	ATP-dependent RNA helicase abstrakt				0.63	31		
11351	FBpp0070651	cap binding protein 80, isoform A				0.63	28		
11975	FBpp0080282	crinkled, isoform A				0.63	17		
11349	FBpp0083972	4EHP	eukaryotic translation initiation factor 4e type			0.63	56		
11529	FBpp0076244	Probable signal recognition particle 68 kDa protein	srp68			0.63	50		
11355	FBpp0081374	belle	DEAD box ATP-dependent RNA helicase			0.63	41		
12053	FBpp0072788	CG9018-PB, isoform B	*regulation of nuclear pre-mRNA domain containing 1B*			0.63	38		
11368	FBpp0075729	RhoGAP68F				0.63	36		
11831	FBpp0078191	CG6838-PB, isoform B	*ADP-ribosylation factor GTPase activating protein 2*			0.63	55		
11613	FBpp0086590	CG12797-PA	WD-repeat protein			0.63	42		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11799	FBpp0078371	MLF1-adaptor molecule				0.63	30		
11883	FBpp0089034	Armadillo segment polarity protein				0.63	21		
11880	FBpp0071407	Mannosyl-oligosaccharide alpha-1,2-mannosidase isoform 2				0.63	30		
11771	FBpp0078161	Tenascin major				0.63	16		
11843	FBpp0078887	CG9523-PA	*FIC domain containing*			0.63	28		
11531	FBpp0085140	CG31004-PB, isoform B	*sushi domain containing 2*			0.63	23		
11571	AAEL012316-PA	arsenical pump-driving ATPase				0.63	31		
12031	AAEL014285-PA	growth hormone inducible transmembrane protein		du [growth hormone inducible transmembrane protein]		0.63	59		
11501	FBpp0074151	Probable small nuclear ribonucleoprotein G		du [small nuclear ribonucleoprotein polypeptide G]		0.63	61		
11520	FBpp0073134	Fumarylacetoacetase				0.62	52		
12042	FBpp0082569	CG6194-PA	*ATG4 autophagy related 4 homolog D*			0.62	29		
11647	FBpp0078811	Tetraspanin 26A				0.62	34		
11964	FBpp0089047	Voltage-dependent calcium channel type D alpha-1 subunit				0.62	13		
11739	FBpp0087699	Receptor mediated endocytosis 8				0.62	17		
12049	FBpp0100147	conserved membrane protein at 44E, isoform A				0.62	29		
11742	FBpp0070418	CG16903-PA	cyclin I			0.62	38		
11822	FBpp0078893	CG9547-PA	acyl-CoA dehydrogenase			0.62	51		
11424	FBpp0079870	escl, isoform A				0.62	40		
11413	FBpp0086223	Flap endonuclease 1				0.62	39		
11705	FBpp0075485	Protein frizzled precursor				0.62	25		
11737	FBpp0087722	Dynamitin				0.62	53		
11701	AAEL010002-PB	5-formyltetrahydrofolate cyclo-ligase				0.62	53		
11939	FBpp0076523	Protein henna				0.62	59		
11576	FBpp0070104	Beta-amyloid-like protein precursor				0.62	37		
11527	FBpp0084894	CG31033-PB, isoform B	*ATG16 autophagy related 16-like 1*			0.62	18		
12102	FBpp0081633	CG9461-PA	F-box only protein			0.62	23		
12092	FBpp0088153	Eph receptor tyrosine kinase, isoform D				0.62	18		
11631	FBpp0070368	6-phosphogluconate dehydrogenase, decarboxylating				0.62	56		
12075	FBpp0084499	CG6051-PA	lateral signaling target protein			0.62	29		
11903	FBpp0072723	CG1140-PA, isoform A	succinyl-coa: 3-ketoacid-coenzyme a transferase			0.61	43		
11896	FBpp0086289	HMG Coenzyme A synthase, isoform A				0.61	42		
11922	FBpp0079976	PICK1, isoform B				0.61	30		
11996	FBpp0078360	sec23, isoform B				0.61	24		
11625	FBpp0088955	Protein tumorous imaginal discs, mitochondrial precursor				0.61	56		
12066	FBpp0070793	CG3016-PA	ubiquitin-specific protease			0.61	23		
11913	FBpp0083976	Rox8, isoform F				0.61	39		
12062	FBpp0088862	Hypothetical protein CG7816				0.61	36		
11416	FBpp0088910	CG1732-PB, isoform B	sodium/chloride dependent neurotransmitter transporter			0.61	30		
11530	FBpp0070469	Hypothetical protein CG32795 in chromosome 1				0.61	44		
11565	FBpp0077998	CG7338-PA	ribosome biogenesis protein tsr1			0.61	50		
11881	FBpp0082624	CG4525-PA	*tetratricopeptide repeat domain 26*			0.61	21		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
12103	FBpp0084774	CG1458-PA	*CDGSH iron sulfur domain 2*			0.61	68		
12014	FBpp0075942	CG7628-PA	phosphate transporter			0.61	29		
11607	FBpp0071259	CG12135-PA	*CWC15 spliceosome-associated protein homolog*			0.61	51		
11930	FBpp0074330	CG6179-PA	*nitric oxide synthase interacting protein*			0.61	44		
12016	FBpp0070751	RhoGAP5A, isoform A				0.61	24		
11630	FBpp0079643	CG5366-PA	cullin-associated NEDD8-dissociated protein 1			0.60	20		
12013	FBpp0087648	RNA-binding protein 8A				0.60	56		
11666	FBpp0072660	Hsp90 co-chaperone Cdc37				0.60	56		
11483	FBpp0077886	Lipoic acid synthase, isoform B	lipoic acid synthetase			0.60	48		
11390	FBpp0075731	Neurexin-4 precursor				0.60	17		
11929	FBpp0074822	Aut1				0.60	40		
11396	FBpp0070064	Molybdenum cofactor synthesis protein cinnamon	molybdopterin biosynthesis protein			0.60	29		
11719	FBpp0081331	CG10153-PA	*trafficking protein particle complex 5*			0.60	35		
11978	FBpp0087366	CG11777-PA	cyclophilin-10			0.60	30		
11727	FBpp0071303	CG3004-PA	vegetatible incompatibility protein HET-E-1 putative			0.60	39		
11600	FBpp0087340	CG7686-PA	*LTV1 homolog*			0.60	53		
12107	FBpp0085500	CG3358-PB, isoform B	*TatD DNase domain containing 1*			0.60	33		
11372	FBpp0087938	Nup44A, isoform A				0.60	40		
11792	FBpp0071262	CG17446-PA	cpg binding protein			0.60	27		
11840	FBpp0078381	CG2185-PA	calcineurin b subunit			0.60	61		
11926	AAEL002852-PA	conserved hypothetical protein				0.60	16		
11357	FBpp0074246	CG8142-PA	replication factor C 37-kDa subunit putative			0.60	42		
12011	FBpp0070162	CG11642-PC, isoform C	translocation associated membrane protein			0.60	66		
11447	FBpp0086703	CG8394-PA	amino acid transporter			0.60	21		
11512	FBpp0074937	NUCB1				0.60	51		
11671	FBpp0071392	CG32687-PA	internalin A putative			0.60	46		
11606	FBpp0077047	lethal (1) G0196, isoform E				0.60	21		
12094	FBpp0079675	CG5676-PA				0.59	51		
11518	FBpp0073557	CG4332-PA	*CLPTM1-like*			0.59	35		
11925	FBpp0071138	Probable phenylalanyl-tRNA synthetase alpha chain				0.59	37		
12079	FBpp0078891	CG9543-PA	*coatomer protein complex, subunit epsilon*			0.59	58		
11654	FBpp0077263	Probable tyrosyl-DNA phosphodiesterase				0.59	33		
11407	FBpp0076124	Ubiquitin-conjugating enzyme E2-22 kDa	ubiquitin-conjugating enzyme E2-25kDa			0.59	50		
11643	FBpp0071688	Protein ariadne-2				0.59	33		
11731	FBpp0085222	lethal (3) s1921				0.59	47		
12056	FBpp0081800	Sorbitol dehydrogenase-2				0.59	69		
11767	FBpp0086875	F-box/SPRY-domain protein 1				0.59	21		
11944	FBpp0071269	CG12121-PA	lung seven transmembrane receptor			0.59	26		
11561	FBpp0072481	CG13887-PB, isoform B	B-cell receptor-associated protein bap	du [B-cell receptor-associated protein 31]		0.59	70		
11997	FBpp0079914	Threonyl-tRNA synthetase, isoform C		ba [trs]		0.59	27		
11834	FBpp0077129	CG15433-PA	elongator component putative			0.59	35		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11538	FBpp0087714	CG8080-PA	*chromosome 5 open reading frame 33*			0.59	32		
11478	FBpp0074792	CG6812-PA	sideroflexin 123			0.59	33		
12000	FBpp0073354	CG1749-PA	ubiquitin-activating enzyme E1			0.59	46		
11694	FBpp0070933	Serine/threonine-protein kinase				0.59	16		
11817	FBpp0110272	multiple C2 domain and transmembrane region protein				0.59	20		
11570	FBpp0071229	CG7039-PA	ARL3 putative			0.59	42		
11692	FBpp0079812	Replication factor C 38kD subunit				0.59	42		
11761	FBpp0088881	supercoiling factor, isoform B				0.59	49		
11524	FBpp0086373	Cysteinyl-tRNA synthetase				0.59	38		
12110	FBpp0099935	CG11919-PA, isoform A	peroxisome assembly factor-2 (peroxisomal-type ATPase 1)			0.59	26		
11884	FBpp0071478	CDK5RAP3-like protein				0.59	44		
11427	FBpp0075042	rogdi, isoform A				0.59	29		
11960	FBpp0087073	CG8841-PC, isoform C	*chromosome 17 open reading frame 28*			0.59	18		
11488	FBpp0079946	Probable ribosome production factor 1	U3 small nucleolar ribonucleoprotein protein imp4			0.59	44		
12080	FBpp0082525	CG4338-PA	*chromosome 16 open reading frame 42*			0.59	37		
11358	FBpp0078721	thickveins, isoform D				0.59	31		
11677	FBpp0071189	CG12125-PA	*family with sequence similarity 73, member B*			0.59	24		
12109	FBpp0075866	CG11660-PA, isoform A	serine/threonine-protein kinase rio1 (rio kinase 1)			0.58	33		
12050	FBpp0081087	CG2656-PA	*GPN-loop GTPase 3*			0.58	41		
11356	FBpp0080407	CG5861-PA	*transmembrane protein 147*			0.58	47		
11651	FBpp0075139	CG4933-PA	o-sialoglycoprotein endopeptidase			0.58	29		
11752	FBpp0088329	Calcium-dependent secretion activator				0.58	13		
11878	FBpp0071669	GlcT-1				0.58	28		
11657	FBpp0070443	40S ribosomal protein S12, mitochondrial precursor			x	0.58	38		
11871	FBpp0074990	UDP-sugar transporter UST74c				0.58	29		
11920	FBpp0074662	Rpn1				0.58	30		
11953	FBpp0084464	BM-40-SPARC				0.58	68		
11863	FBpp0083098	Mekk1, isoform B				0.58	17		
11614	FBpp0099673	Tousled-like kinase, isoform D				0.58	24		
11438	FBpp0078382	MTA1-like, isoform B				0.58	17		
11381	FBpp0081659	lethal (3) IX-14				0.58	20		
11686	FBpp0072142	Protein within the bgcn gene intron				0.57	38		
11875	FBpp0084118	CG5805-PA	mitochondrial glutamate carrier putative			0.57	17		
11360	FBpp0088059	Dpld (Protein dappled)				0.57	17		
12090	FBpp0075734	CG6910-PA	myoinositol oxygenase			0.57	53		
11734	AAEL010797-PA	RNA polymerase II holoenzyme component				0.57	31		
11477	FBpp0078633	CG3756-PA	DNA-directed RNA polymerase			0.57	44		
12048	FBpp0086107	Anaphase-promoting complex subunit 10				0.57	36		
11987	FBpp0081601	CG9373-PA	myelinprotein expression factor			0.57	43		
11882	FBpp0073588	CG1622-PA	*PRP38 pre-mRNA processing factor 38*			0.57	33		
11810	FBpp0086757	CG12295-PB: straightjacket	dihydropyridine-sensitive l-type calcium channel			0.57	16		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11668	FBpp0070389	mitochondrial ribosomal protein L14			x	0.57	38		
11858	FBpp0086063	Ngp				0.56	34		
11795	FBpp0084691	CG1646-PC, isoform C	*PRP39 pre-mRNA processing factor 39 homolog*			0.56	43		
11544	FBpp0085838	GDI interacting protein 3, isoform C				0.56	33		
11400	FBpp0071061	Integrin beta-PS precursor	integrin beta subunit			0.56	45		
11550	FBpp0076134	ATP synthase B chain, mitochondrial precursor				0.56	16		
11725	FBpp0076486	pebble, isoform D				0.56	18		
11804	FBpp0087806	CG8635-PA	*zinc finger CCCH-type containing 15*			0.56	45		
11674	FBpp0074121	CG9099-PA	*density-regulated protein*			0.56	54		
11839	FBpp0081520	Probable maleylacetoacetate isomerase 2				0.56	43		
11559	FBpp0110208	calnexin				0.56	59		
11794	FBpp0084559	rapsynoid				0.56	19		
11656	FBpp0075168	Tyrosyl-tRNA synthetase				0.56	45		
11489	FBpp0071445	CG9236-PA	calcium and integrin-binding protein 1			0.56	22		
12111	FBpp0084144	CG11920-PA	U3 small nucleolar ribonucleoprotein protein imp4			0.56	45		
11774	FBpp0085422	O-glycosyltransferase, isoform B				0.56	19		
11683	FBpp0086129	Fat-spondin, isoform B				0.56	52		
11461	FBpp0081481	Protein neuralized				0.55	29		
11835	FBpp0079780	CG6724-PA	WD-repeat protein			0.55	45		
11974	FBpp0083581	CG6015-PA	pre-mRNA splicing factor prp17			0.55	35		
11354	FBpp0081552	CG8286-PA	tetratricopeptide repeat protein putative			0.55	48		
12063	FBpp0078685	Probable GDP-mannose 4,6 dehydratase				0.55	38		
12078	FBpp0087709	Mystery 45A				0.55	34		
11972	FBpp0072382	mrityu, isoform C				0.55	24		
11828	FBpp0082895	CG5840-PB, isoform B	pyrroline-5-carboxylate reductase			0.55	51		
11866	FBpp0083076	Probable 28 kDa Golgi SNARE protein				0.55	37		
11382	FBpp0082888	Sur-8, isoform A				0.55	33		
11704	FBpp0081370	CG8036-PD, isoform D	transketolase I			0.55	41		
11487	FBpp0083131	Prp18				0.55	38		
11691	FBpp0071063	Glutamate--cysteine ligase				0.55	30		
11458	FBpp0074562	CG32528-PA	parvin			0.54	48		
11861	FBpp0074026	Katanin 80, isoform B				0.54	20		
11586	FBpp0074835	CG6841-PA	pre-mRNA splicing factor			0.54	25		
11959	FBpp0080906	La protein homolog				0.54	55		
11404	FBpp0080622	CG10333-PA	DEAD box ATP-dependent RNA helicase			0.54	20		
11467	FBpp0072979	CG11537-PB, isoform B	*hippocampus abundant transcript 1*			0.54	20		
11364	FBpp0075238	PDCD-5	*programmed cell death 5*	du, de [pace6]		0.54	53		
11370	FBpp0081276	pyd3				0.54	49		
11722	FBpp0080048	Coatomer subunit beta'				0.54	21		
11425	FBpp0076861	Kinesin-like protein at 64D				0.54	27		
11675	FBpp0076332	CG7112-PA	rab6 GTPase activating protein gapcena (rabgap1 protein)			0.54	23		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11895	FBpp0081475	CG18005-PA	red protein (ik factor) (cytokine ik)			0.54	30		
11827	FBpp0081719	Sirt6				0.54	29		
12045	FBpp0079832	CG6509-PB, isoform B	discs large protein			0.54	14		
11615	FBpp0070367	CG3835-PA, isoform A	D-lactate dehydrognease 2			0.54	31		
11937	FBpp0079577	Ubiquitin thioesterase otubain-like protein				0.54	41		
11546	FBpp0072404	mitochondrial ribosomal protein L17		ph, du, de, ba [rpl17]	x	0.54	38		
12099	FBpp0085155	Coatomer protein, isoform B				0.54	28		
11915	FBpp0073739	MRNA-capping-enzyme				0.54	46		
11579	FBpp0077551	CG31938-PA	*exosome component 3*			0.54	31		
11776	FBpp0072455	Probable UDP-glucose 4-epimerase				0.54	47		
11446	FBpp0084351	CG6095-PB, isoform B	exocyst complex-subunit protein 84kDa-subunit putative			0.54	31		
11434	FBpp0074582	CG14232-PA	*acyl-Coenzyme A binding domain containing 3*			0.54	29		
11809	FBpp0085181	CG1800-PA: partner of drosha	double-stranded binding protein putative			0.53	23		
11521	FBpp0071095	CG10932-PA	acetyl-coa acetyltransferase mitochondrial			0.53	65		
12106	FBpp0075693	Probable phosphomannomutase				0.53	47		
12098	FBpp0086667	CG8531-PA	*DnaJ (Hsp40) homolog, subfamily C, member 11*			0.53	35		
11646	FBpp0075111	COP, isoform B				0.53	62		
11426	FBpp0083921	CG5991-PC, isoform C				0.53	37		
11764	FBpp0073806	CG14407-PA	glutaredoxin			0.53	68		
12067	FBpp0088040	CG11107-PA	ATP-dependent RNA helicase			0.53	22		
11528	FBpp0080117	CG16865-PA	*chromosome X open reading frame 56*			0.53	34		
11769	FBpp0073649	CG11134-PA	*APAF1 interacting protein*			0.53	48		
11590	FBpp0085763	Exostosin-3				0.53	24		
12083	FBpp0075947	Multidrug-Resistance like Protein 1, isoform B				0.53	26		
11894	FBpp0071256	C12.2				0.53	27		
11409	FBpp0074844	CG3961-PB, isoform B	long-chain-fatty-acid coa ligase			0.53	35		
11441	FBpp0073635	CG11178-PB, isoform B	*AVL9 homolog*			0.53	25		
11456	FBpp0072830	misshapen, isoform E				0.53	31		
11444	FBpp0071232	AP-1, isoform E				0.53	17		
11423	FBpp0078070	CG9391-PA, isoform A	myo inositol monophosphatase			0.53	52		
11526	FBpp0074564	CG12703-PA	peroxisomal membrane protein 70 abcd3			0.52	23		
11708	FBpp0081576	eclair				0.52	65		
12043	FBpp0074736	CG8798-PA, isoform A	ATP-dependent Lon protease putative			0.52	27		
11745	FBpp0090943	CG33505-PA	WD-repeat protein			0.52	36		
12096	FBpp0087342	CG12343-PA	*SYF2 homolog, RNA splicing factor*			0.52	51		
12026	FBpp0084813	CG1907-PA	*solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11*			0.52	49		
12116	FBpp0078694	mitochondrial ribosomal protein L24			x	0.52	50		
12114	FBpp0075560	CG10711-PA	conserved hypothetical protein			0.52	46		
11596	FBpp0076073	nudE				0.52	38		
11698	FBpp0078992	Gas41				0.52	31		
11904	AAEL010402-PA	DEAD box ATP-dependent RNA helicase				0.52	20		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11970	FBpp0083436	Exocyst complex component 6				0.52	36		
11756	FBpp0086641	Lamin-C				0.52	37		
11724	FBpp0081283	CG10903-PA	*Williams Beuren syndrome chromosome region 22*			0.52	42		
11721	AAEL004763-PA	conserved hypothetical protein				0.52	26		
11592	FBpp0070806	Lethal (1), isoform A				0.52	21		
11601	FBpp0081834	CG5214-PA	dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase			0.52	45		
11431	FBpp0074226	CG5703-PA	NADH-ubiquinone oxidoreductase 24 kDa subunit	du [NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa]		0.52	71		
11811	FBpp0079495	CG5885-PA	translocon-associated protein gamma subunit			0.52	84		
11775	FBpp0074517	Glucose-6-phosphate 1-dehydrogenase				0.52	42		
11506	FBpp0082642	CG4225-PA	ABC transporter *Mitochondrial ABC transporter 3*			0.51	17		
11826	FBpp0110402	eukaryotic translation initiation factor 3, theta subunit				0.51	24		
11422	FBpp0085952	Dgp-1, isoform A				0.51	33		
11782	FBpp0073828	CG6227-PA	DEAD box ATP-dependent RNA helicase			0.51	17		
11685	FBpp0071217	Polycomb protein l(1)G0020				0.51	30		
11723	FBpp0081799	CG6465-PA	aminoacylase putative			0.51	52		
11376	FBpp0075938	NEDD8-activating enzyme E1 regulatory subunit	app binding protein			0.51	39		
11699	FBpp0075034	CG7728-PA	ribosome biogenesis protein			0.51	29		
11789	FBpp0084190	CG11858-PA	peptidyl-prolyl cis/trans isomerase, putative	du [protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting 1]		0.51	40		
11463	FBpp0087926	drosha				0.51	15		
11867	FBpp0074734	CG8793-PA	*KIAA1012*			0.50	17		
11697	AAEL013319-PA	conserved hypothetical protein		ph, de [stbproptase2a-b]		0.50	15		
11688	FBpp0075069	CG4169-PA	ubiquinol-cytochrome c reductase complex core protein			0.50	82		
11952	FBpp0076782	Regulator of chromosome condensation				0.50	30		
11740	FBpp0074366	Histidyl-tRNA synthetase, isoform B				0.50	45		
11898	FBpp0087506	6-phosphofructokinase				0.50	30		
11892	FBpp0083899	Bifunctional aminoacyl-tRNA synthetase				0.50	24		
11473	FBpp0084489	DNA polymerase alpha subunit B				0.50	29		
12008	FBpp0078184	Secretory Pathway Calcium atpase, isoform C				0.50	17		
11513	FBpp0072531	CG9119-PA	*chromosome 11 open reading frame 54*			0.50	44		
11669	FBpp0073875	CG9245-PB, isoform B	phosphatidylinositol synthase			0.50	46		
11813	Fbpp0089153	smallminded CG8571-PB, isoform B	peroxisome assembly factor-2 (peroxisomal-type ATPase 1)	de [nsf2-B]		0.50	32		
12023	FBpp0083842	3-hydroxy-3-methylglutaryl-coenzyme A reductase				0.50	35		
11659	FBpp0087353	CG16728-PA	*G protein-coupled receptor kinase interacting ArfGAP 2*			0.50	25		
11886	FBpp0080045	Two A-associated protein of 42kDa				0.50	37		
11852	FBpp0084032	CG6643-PB, isoform B	synaptotagmin putative			0.50	33		
11982	FBpp0072779	CG1317-PB	ssm4 protein			0.50	25		
11457	FBpp0082284	falafel, isoform C				0.50	17		
11412	FBpp0070643	CG3564-PA	copii-coated vesicle membrane protein P24			0.50	72		
11837	FBpp0074510	CG14211-PB	dual-specificity protein phosphatase putative			0.49	30		
11449	FBpp0074285	3-hydroxyacyl-CoA dehydrogenase type-2	hydroxyacyl dehydrogenase			0.49	68		
11653	FBpp0085609	Mediator complex subunit 8				0.49	38		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11941	FBpp0085630	CG11208-PA	2-hydroxyphytanoyl-coa lyase			0.49	42		
11914	FBpp0072495	CG13900-PB, isoform B	spliceosomal protein sap			0.49	23		
11889	FBpp0099977	CG1410-PA, isoform A	GTP-binding protein lepa			0.49	30		
11539	FBpp0074936	CG5589-PA	DEAD box ATP-dependent RNA helicase			0.49	36		
11471	FBpp0080509	Aminopeptidase P				0.49	52		
11730	FBpp0086954	Chromatin remodelling complex ATPase chain Iswi				0.49	20		
11459	FBpp0080319	lethal (2) 35Df				0.49	21		
11909	FBpp0078319	CG2051-PC, isoform C	histone acetyltransferase type b catalytic subunit			0.49	38		
11825	FBpp0070181	CG3704-PA	xpa-binding protein 1 (mbdin)			0.49	34		
12039	FBpp0085873	Late endosomal/lysosomal Mp1-interacting protein homolog				0.49	51		
11936	FBpp0086402	CG8386-PA	*ubiquitin-fold modifier conjugating enzyme 1*			0.49	53		
12037	FBpp0080062	Ski6		du, de [rrp46-B]		0.49	37		
11918	FBpp0087402	Caf1-105				0.49	29		
11636	FBpp0084013	Golgin-84				0.49	25		
11500	FBpp0088794	CG33298-PB, isoform B	phospholipid-transporting ATPase 1 (aminophospholipid flippase 1)			0.48	20		
11566	FBpp0082288	neither inactivation nor afterpotential B				0.48	26		
11784	FBpp0072058	Alpha-catenin-related, isoform B				0.48	21		
12077	FBpp0077208	Exocyst complex component 2				0.48	33		
11363	FBpp0083319	CG5434-PA (Srp72)	*Signal recognition particle 72 kDa protein*			0.48	50		
11465	AAEL000324-PA	tyrosine-protein kinase drl				0.48	21		
12017	FBpp0087867	Mlh1				0.48	26		
11741	FBpp0079735	Vacuolar protein sorting protein 72 homolog				0.48	31		
11621	FBpp0072711	CG12091-PA	protein phosphatase 2c			0.48	48		
11408	FBpp0082066	Interleukin enhancer-binding factor 2 homolog	interleukin enhancer binding factor			0.48	54		
11720	AAEL002870-PA	Dipeptidyl-peptidase 3				0.48	55		
12101	FBpp0076771	CG10467-PA	aldose-1-epimerase			0.48	42		
11676	FBpp0075209	Signal sequence receptor		du [signal sequence receptor, beta precursor]		0.48	92		
11644	FBpp0075535	Ral guanine nucleotide exchange factor 2, isoform A				0.48	21		
11352	FBpp0075513	Hsc70Cb, isoform C				0.48	50		
11714	FBpp0089006	CG32626-PD, isoform D	AMP deaminase			0.48	22		
11833	FBpp0075151	multiprotein bridging factor, isoform B		du [endothelial-differentiation-related factor 1 isoform alpha]		0.48	74		
11515	FBpp0083989	Dis3				0.48	32		
11998	FBpp0078512	CG1126-PA	*Bardet-Biedl syndrome 5*			0.48	17		
11435	FBpp0086042	CG6401-PA	glycosyltransferase			0.48	28		
11790	FBpp0072468	CG6905-PA	cell division control protein			0.48	19		
11558	FBpp0071277	Zpr1				0.48	43		
11862	FBpp0077203	CG31957-PA	translation initiation factor 1A putative			0.48	34		
12085	FBpp0086957	CG8632-PB, isoform B	*solute carrier family 30 (zinc transporter), member 9*			0.48	27		
11469	FBpp0083440	Uridine 5'-monophosphate synthase	orotidine-5'-phosphate decarboxylase, putative			0.48	46		
11474	FBpp0070180	CG3703-PA	*RUN domain containing 1*			0.47	23		
11443	FBpp0079162	CG7429-PA	*coiled-coil domain containing 53*			0.47	32		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
12022	FBpp0084411	CG5484-PC, isoform C	*Yip1 interacting factor homolog B*			0.47	54		
11551	FBpp0074964	CG6259-PA	charged multivesicular body protein 5			0.47	55		
11887	FBpp0070637	CG6133-PA	*NOL1/NOP2/Sun domain family, member 2*			0.47	39		
12058	FBpp0074616	FRG1 protein homolog				0.47	42		
11533	FBpp0087458	CG12214-PA, isoform A	tubulin-specific chaperone e			0.47	29		
12104	FBpp0085393	CG7791-PA	mitochondrial intermediate peptidase			0.47	34		
11938	FBpp0083768	CG13827-PA	*peroxisomal biogenesis factor 11 gamma*			0.47	27		
12004	FBpp0071818	Hypothetical UPF0172 protein CG3501				0.47	38		
11439	FBpp0085829	CG15087-PA	*chromosome 11 open reading frame2*			0.47	28		
11361	FBpp0082332	CG3061-PA	DNA-J, putative			0.47	52		
11779	FBpp0070924	COQ7				0.47	46		
11421	FBpp0074715	anti-silencing factor 1				0.47	39		
11626	FBpp0076459	CG7550-PA	2-aminoethanethiol (cysteamine) dioxygenase			0.47	26		
11650	FBpp0080553	Putative conserved oligomeric Golgi complex component 5				0.47	24		
11715	FBpp0086992	CG18177-PB, isoform B				0.47	28		
11981	FBpp0077333	CG3542-PB, isoform B	U1 small nuclear ribonucleoprotein putative			0.46	37		
11961	FBpp0084418	CG6420-PA	WD-repeat protein			0.46	19		
11517	FBpp0073082	CG14997-PB, isoform B	sulfide quinone reductase			0.46	52		
11957	FBpp0089113	Transcription elongation factor SPT5				0.46	15		
11414	FBpp0079258	CG12375-PA	metallo-beta-lactamase putative			0.46	47		
11729	FBpp0079469	CG4537-PA	*cysteine-rich PDZ-binding protein*			0.46	29		
11807	FBpp0081556	Spermidine Synthase				0.46	47		
11464	FBpp0079697	CG6415-PA	aminomethyltransferase			0.46	39		
12025	FBpp0082065	Aos1				0.46	41		
11684	FBpp0083351	CG4159-PA	pseudouridylate synthase			0.46	43		
12036	FBpp0072421	Enhancer of bithorax, isoform C				0.46	14		
11743	FBpp0071597	CG9865-PB, isoform B	*phosphatidylinositol glycan anchor biosynthesis, class M (CG9865)*			0.46	32		
11452	FBpp0083214	Vacuolar ATP synthase subunit G				0.46	102		
11770	FBpp0085121	39S ribosomal protein L32, mitochondrial precursor			x	0.46	42		
11856	FBpp0080638	CG12750-PA	cell cycle control protein cwf22			0.46	28		
11433	FBpp0079468	FK506-binding protein 59				0.46	57		
11623	FBpp0075393	CG6859-PA	perixosomal biogenesis factor			0.46	36		
11493	FBpp0085258	CG1416-PC, isoform C	*AHA1, activator of heat shock 90kDa protein ATPase homolog 1*			0.46	50		
11522	FBpp0072564	CG9153-PB, isoform B	hect E3 ubiquitin ligase			0.46	36		
11415	FBpp0072841	mitochondrial ribosomal protein S35			x	0.46	43		
11374	FBpp0085363	SCAP				0.46	17		
11365	FBpp0077150	Probable DNA replication complex GINS protein PSF2	*GINS complex subunit 2 (Psf2 homolog)*			0.46	33		
11367	FBpp0070302	Myb-interacting protein 130				0.46	20		
11494	FBpp0079203	CG8506-PA	*zinc finger, FYVE domain containing 20*			0.46	28		
12001	FBpp0099494	C-1-tetrahydrofolate synthase, cytoplasmic				0.45	18		
11977	FBpp0075120	CG4098-PA	nudix hydrolase 6			0.45	27		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11766	FBpp0100136	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase				0.45	19		
11947	FBpp0072946	CG11526-PB, isoform B	*family with sequence similarity 40, member A*			0.45	20		
11877	AAEL006769-PA	tryptophanyl-tRNA synthetase				0.45	25		
11670	FBpp0080918	CG2614-PA	*KIAA0859*			0.45	34		
11946	FBpp0079699	CG6443-PA	*chromosome 20 open reading frame 43*			0.44	54		
11453	FBpp0072961	CG14967-PA	*KIAA0100*			0.44	18		
11679	FBpp0077525	tho2				0.44	24		
11943	FBpp0078358	CG12170-PA	3-oxoacyl-[acyl-carrier-protein] synthase			0.44	36		
11664	FBpp0074808	CG3808-PA	RNA m5u methyltransferase			0.44	40		
11900	FBpp0073966	Clathrin heavy chain				0.44	18		
11713	FBpp0073663	iodotyrosine dehalogenase	iodotyrosine dehalogenase			0.44	17		
11562	FBpp0079897	CG6746-PA	ptpla domain protein			0.44	53		
12086	FBpp0071031	Probable mitochondrial import receptor subunit TOM40 homolog				0.44	54		
11748	FBpp0087244	CG30022-PA	beta lactamase domain			0.44	55		
11655	FBpp0084069	tolkin, isoform B				0.44	16		
11732	FBpp0084626	CG4849-PA	116 kDa U5 small nuclear ribonucleoprotein component			0.44	17		
11738	FBpp0081355	CG9630-PA	DEAD box ATP-dependent RNA helicase			0.44	29		
11505	FBpp0086380	CG8443-PA	eukaryotic translation initiation factor 3 subunit (eif-3)			0.43	16		
11678	FBpp0072703	CG13926-PA	*chromosome 11 open reading frame 73*			0.43	37		
11466	FBpp0084779	ligatin				0.43	26		
11906	FBpp0081810	CG6608-PB, isoform B	mitochondrial carrier protein putative			0.43	32		
11605	FBpp0076242	CG5026-PA, isoform A	myotubularin			0.43	28		
11445	FBpp0085923	adipose				0.43	25		
11948	FBpp0071194	Probable U3 small nucleolar RNA-associated protein 11				0.43	42		
11901	FBpp0074131	Integrin alpha-PS2 precursor				0.43	17		
11747	FBpp0073491	CG1824-PA	lipid a export ATP-binding/permease protein msba			0.43	26		
11942	FBpp0099560	Protein retinal degeneration B				0.43	23		
12119	FBpp0077133	CG17840-PA	inositol 5-phosphatase			0.42	26		
11818	FBpp0084478	CG5880-PA	*zinc finger, DHHC-type containing 16*			0.42	26		
11371	FBpp0077357	okra				0.42	23		
11845	FBpp0071543	CG30390-PA	*coiled-coil domain containing 101*			0.42	29		
11397	FBpp0083272	Ire-1	Serine threonine-protein kinase *endoplasmic reticulum to nucleus signaling 2*			0.42	19		
11979	FBpp0083132	gatA				0.42	29		
11510	FBpp0075990	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase	acireductone dioxygenase			0.42	53		
12064	FBpp0083137	CG14290-PB	*brain protein 44-like*			0.42	34		
11472	FBpp0082817	CG16941-PA	spliceosome associated protein			0.42	19		
11532	FBpp0070299	CG14805-PA	PAF acetylhydrolase 45 kDa subunit putative			0.42	42		
11781	FBpp0073083	pavarotti				0.42	24		
12124	FBpp0078448	Probable proteasome subunit beta type 4		ph, du, de, ba [psmb-N]		0.41	92		
12100	AAEL010379-PA	ATP-binding cassette transporter				0.41	16		
12089	FBpp0072366	3-phosphoinositide-dependent protein kinase 1				0.41	34		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
12020	FBpp0075609	CG11267-PA	heat shock protein putative	du [Heat shock 10 kDa protein 1 (chaperonin 10)]		0.41	83		
11599	FBpp0076280	CG5288-PC, isoform C	galactokinase			0.41	41		
11462	FBpp0087323	CG6751-PA	WD-repeat protein			0.41	38		
11492	FBpp0110411	conserved hypothetical protein				0.41	18		
11949	FBpp0086399	CG8397-PA	actin binding protein putative			0.41	60		
11662	FBpp0075280	Homeotic gene regulator				0.41	22		
11485	FBpp0083354	Elongin B		du [elongin B isoform a]		0.41	57		
11582	AAEL004330-PA	conserved hypothetical protein				0.41	17		
11369	FBpp0085431	Transcription-associated protein 1 (Nipped-A)	transformation/transcription domain-associated protein			0.41	15		
11986	FBpp0071155	Neuroglian precursor				0.41	15		
11969	FBpp0070319	CG4199-PA, isoform A	disulfide oxidoreductase			0.41	34		
11486	FBpp0110523	nitrate, fromate, iron dehydrogenase				0.41	34		
11765	AAEL011712-PA	diacylglycerol kinase				0.41	12		
11973	FBpp0075344	CG7650-PA	viral IAP-associated factor putative			0.41	41		
11994	FBpp0072767	CG8993-PA	thioredoxin putative			0.41	61		
11758	FBpp0076708	Transportin, isoform A				0.41	19		
11682	FBpp0085071	Protein tailless				0.41	16		
11638	FBpp0073725	CG1461-PA	tyrosine aminotransferase			0.41	40		
11661	FBpp0070304	CG3573-PA	inositol polyphosphate 5-phosphatase			0.40	27		
11746	FBpp0088517	CG5009-PA	acyl-CoA oxidase			0.40	28		
11933	FBpp0084307	CG4743-PA	mitochondrial carrier protein			0.40	26		
11475	FBpp0073983	Actin-like protein 13E				0.40	32		
12052	FBpp0080894	cdc23				0.40	27		
11689	FBpp0087297	BBS4				0.40	20		
11717	FBpp0075685	Protein angel				0.40	24		
12030	FBpp0072119	CG3735-PA	*chromosome 1 open reading frame 107*			0.40	29		
11999	FBpp0079776	CG6700-PA	leukocyte receptor cluster (lrc) member			0.40	29		
11888	FBpp0085589	CG11788-PA	*defective in sister chromatid cohesion 1 homolog*			0.40	34		
11780	FBpp0080922	Importin beta subunit				0.40	19		
11568	FBpp0074481	CG12203-PA	NADH: ubiquinone dehydrogenase putative	du [NADH dehydrogenase (ubiquinone) Fe-S protein 4]		0.40	72		
11897	FBpp0082657	Mitochondrial import inner membrane translocase subunit TIM16				0.40	43		
11597	FBpp0087591	Protein preli-like				0.40	55		
11540	FBpp0087891	CG8709-PA	lipin			0.40	32		
11648	FBpp0080807	Probable phosphomevalonate kinase				0.39	37		
11950	FBpp0088810	Protein arginine N-methyltransferase capsuleen				0.39	37		
11554	FBpp0071033	Pyruvate dehydrogenase phosphatase (CG12151-PA)				0.39	29		
11468	FBpp0079547	Niemann-Pick Type C-1				0.39	24		
11611	AAEL006321-PA	1-acylglycerol-3-phosphate acyltransferase				0.39	20		
11967	FBpp0078711	CG12512-PA	AMP dependent coa ligase			0.39	43		
12072	FBpp0082148	CG5608-PA	*Vac14 homolog*			0.39	33		
11541	FBpp0079586	CG31715-PA	*myotrophin*			0.39	37		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11968	FBpp0087979	Cytochrome b5				0.38	89		
11395	FBpp0076789	Pole2				0.38	27		
12117	FBpp0081376	Dihydroorotate dehydrogenase, mitochondrial precursor				0.38	34		
11927	FBpp0081157	CG1104-PA, isoform A				0.38	27		
11359	FBpp0083514	DNA polymerase alpha catalytic subunit				0.38	17		
11966	FBpp0082813	CG3534-PA	xylulose kinase			0.38	30		
11876	FBpp0074672	Translocase of outer membrane 20				0.38	67		
11908	FBpp0079488	CG13126-PA	*methyltransferase 11 domain containing 1*			0.38	30		
11497	FBpp0110309	poly a polymerase				0.38	21		
11401	FBpp0083840	CG10365-PA, isoform A	*ChaC, cation transport regulator homolog 1*			0.38	38		
11842	FBpp0073355	Probable signal peptidase complex subunit 2		du [signal peptidase complex subunit 2 homolog]		0.38	67		
11791	FBpp0077447	CG9867-PA	glycosyltransferase			0.38	29		
11872	FBpp0078684	CG8891-PA	inosine triphosphate pyrophosphatase (itpase) (inosine triphosphatase)			0.37	37		
11470	FBpp0084728	Protein kinase C				0.37	21		
12095	FBpp0070301	mitochondrial ribosomal protein L16			x	0.37	41		
12041	FBpp0074227	CG5800-PA	DEAD box ATP-dependent RNA helicase			0.37	34		
11923	FBpp0083244	CG4973-PA	zinc finger protein putative			0.37	35		
11885	FBpp0084711	CG1951-PA	*SCY1-like 2*			0.37	21		
11622	FBpp0083022	CG7146-PA	vacuolar protein sorting 39 homolog			0.37	26		
11498	FBpp0081588	CG9399-PA, isoform A	*brain protein 44*			0.37	60		
11353	FBpp0074004	CG32579-PA	*XK, Kell blood group complex subunit-related family, member 6*			0.37	21		
11870	FBpp0079567	CG31717-PA	*phosphatidic acid phosphatase type 2 domain containing 2*			0.37	39		
11951	FBpp0077965	UPF0315 protein				0.37	43		
11560	FBpp0078275	jagunal, isoform C				0.37	47		
11690	FBpp0077209	Pdsw, isoform B				0.37	71		
11777	FBpp0087085	DNA-directed RNA polymerase III 128 kDa polypeptide				0.36	16		
12091	FBpp0083854	Probable oligoribonuclease				0.36	36		
11593	FBpp0081841	CG17187-PA	*DnaJ (Hsp40) homolog, subfamily C, member 17*			0.36	33		
11851	FBpp0071891	Arginine methyltransferase 7				0.36	34		
11581	FBpp0073235	Putative 6-phosphogluconolactonase	6-phosphogluconolactonase			0.36	47		
11481	FBpp0086877	CG4646-PA	*chromosome 1 open reading frame 123*			0.36	36		
12032	FBpp0083853	twister				0.35	17		
11569	FBpp0081451	Adenosine deaminase				0.35	29		
11703	FBpp0080628	CG15161-PA				0.35	21		
11476	FBpp0071426	CG1826-PA	*BTB (POZ) domain containing 9*			0.35	29		
11749	FBpp0078844	CG9154-PA	*N-6 adenine-specific DNA methyltransferase 2 (putative)*			0.35	34		
11706	FBpp0082314	CG9588-PA	26S proteasome non-ATPase regulatory subunit			0.35	48		
11612	FBpp0073995	CG3560-PA	ubiquinol-cytochrome c reductase complex 14 kd protein			0.35	85		
12034	FBpp0076111	Laminin gamma-1 chain precursor				0.35	18		
11548	FBpp0074104	mitochondrial ribosomal protein L22			x	0.35	44		
11620	FBpp0073585	Vesicular-fusion protein Nsf1				0.34	20		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11557	FBpp0079620	CG6206-PB, isoform B	lysosomal alpha-mannosidase (mannosidase alpha class 2b member 1)			0.34	42		
11543	FBpp0089008	Adenine phosphoribosyltransferase				0.34	49		
11602	AAEL007823-PA	PIWI				0.34	18		
12015	FBpp0085924	CG10914-PA				0.34	26		
11931	FBpp0089163	Cleavage and polyadenylation specificity factor, 160 kDa subunit				0.34	19		
11696	FBpp0082172	Xanthine dehydrogenase				0.34	31		
11399	FBpp0086640	DNA-directed RNA polymerase I largest subunit	DNA-directed RNA polymerase I largest subunit			0.33	20		
12006	FBpp0080203	DNA mismatch repair protein spellchecker 1				0.33	22		
11963	FBpp0086591	SMC2				0.33	23		
11744	FBpp0073979	Graf, isoform A				0.33	25		
11440	FBpp0078583	CG9804-PA	lipoate-protein ligase b			0.33	29		
11575	FBpp0084349	Dak1				0.33	56		
11594	FBpp0086887	Tripeptidyl-peptidase 2				0.33	19		
11832	FBpp0072460	Rhythmically expressed gene 2 protein				0.32	25		
12113	FBpp0075106	Probable ATP-dependent RNA helicase Dbp73D				0.32	32		
11585	FBpp0071193	Hypothetical protein CG1785				0.32	41		
11693	FBpp0083857	Putative succinate dehydrogenase [ubiquinone] cytochrome b small subunit, mitochondrial precursor				0.32	63		
11985	FBpp0076589	Signal recognition particle 19 kDa protein	srp19			0.32	50		
11751	FBpp0081763	CG4511-PA	viral IAP-associated factor putative			0.32	51		
11595	FBpp0075755	lethal (3) neo18		du [NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa precursor]		0.32	73		
11633	FBpp0075399	Probable DNA mismatch repair protein MSH6				0.32	25		
12021	FBpp0072426	thoc7, isoform A				0.32	41		
11410	FBpp0070403	Probable ATP-dependent RNA helicase	ATP-dependent RNA helicase			0.31	20		
12005	FBpp0080475	CG31739-PA	aspartyl-tRNA synthetase			0.31	27		
11658	FBpp0081860	mitochondrial ribosomal protein L40			x	0.31	46		
11420	FBpp0082522	ATP synthase O subunit, mitochondrial precursor		du [Mitochondrial ATP synthase, O subunit precursor]		0.31	111		
11503	FBpp0075148	CG33158-PB	translation elongation factor *longation factor Tu GTP binding domain containing 1 isoform 2*			0.31	28		
11819	FBpp0077251	CG33123-PA	leucyl-tRNA synthetase			0.30	18		
11504	FBpp0085690	CG11242-PA	tubulin-specific chaperone b (tubulin folding cofactor b)			0.30	48		
11373	FBpp0078895	CG9542-PA	*arylformamidase*			0.30	28		
11430	FBpp0086226	Superoxide dismutase [Mn], mitochondrial precursor				0.30	81		
11879	FBpp0070639	CG6379-PA	*FtsJ methyltransferase domain containing 2*			0.30	26		
11702	FBpp0071916	CG11079-PC, isoform C	5-formyltetrahydrofolate cyclo-ligase			0.30	41		
11537	FBpp0083226	CG4686-PA				0.30	51		
11945	FBpp0073762	Probable mitochondrial 28S ribosomal protein S25			x	0.29	45		
12044	FBpp0073196	CG15014-PA	*THUMP domain containing 1*			0.29	42		
11800	FBpp0077399	Transportin-Serine/Arginine rich				0.29	23		
12028	FBpp0077173	CG31961-PA, isoform A	tubulin folding cofactor c			0.28	36		
11873	AAEL011682-PA	nuclear pore complex protein nup93				0.28	17		
11907	FBpp0083650	Probable prefoldin subunit 5				0.27	63		

Table A.3 continued

Gene ID	Protein ID	Gene / Description Ensemble Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / *HomoloGene (Hsap)*	Other studies [abbr.]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
11519	FBpp0072615	CG9187-PA	partner of sld5			0.27	32		
11388	FBpp0087629	CG1884-PB, isoform B				0.27	16		
11736	FBpp0084051	CG13625-PA	*BUD13 homolog*			0.27	34		
11812	FBpp0100031	Protein male-less	ATP-dependent RNA helicase			0.26	26		
11753	FBpp0079316	CG13397-PA	alpha-n-acetylglucosaminidase			0.26	22		
12055	FBpp0072456	Rev1				0.25	20		
12027	AAEL011963-PA	conserved hypothetical protein				0.25	18		
11836	AAEL009888-PA	WD-repeat protein				0.25	25		
11665	AAEL004081-PA	dj-1 protein				0.25	57		
11995	FBpp0080305	CG15261-PA	ribonuclease UK114 putative			0.24	70		
11516	AAEL005494-PA	conserved hypothetical protein				0.17	9		

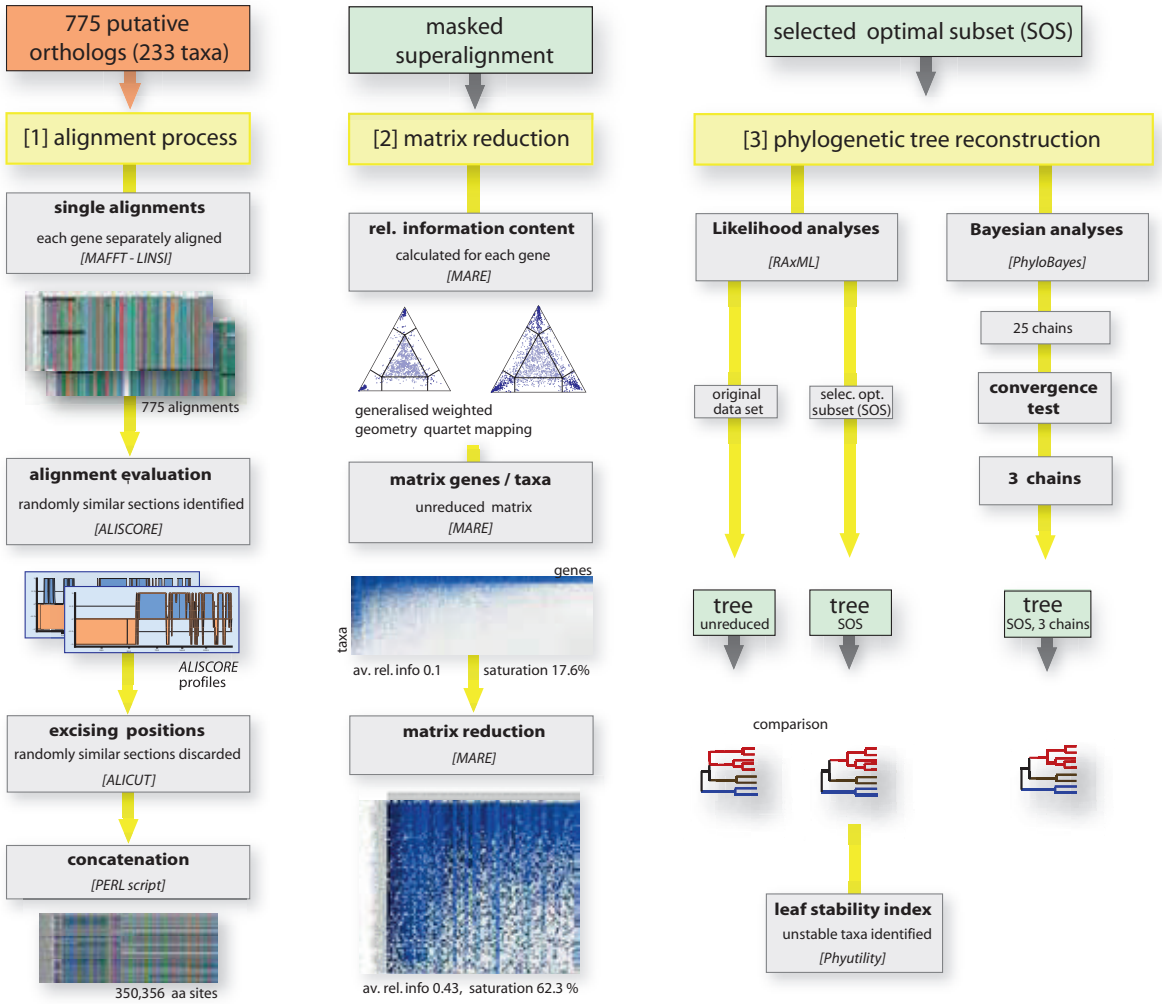

Figure A.3.: (1) Each single gene is separately aligned using MAFFT (Katoh and Toh, 2008). Aliscore (Misof and Misof, 2009) identifies randomly similar sections in each alignment, ALICUT (Kück, 2009) discards by Aliscore negatively scored positions. The genes are concatenated to a masked superalignment. (2) MARE starts with the calculation of the information content (IC) of each gene and taxon in the masked superalignment. The generated matrix taxa *versus* genes is generated with scores of information content. An optimal subset is selected (SOS) by stepwise excluding genes and taxa showing the lowest IC. (3) Phylogenetic trees were calculated using RAxML (Ott et al., 2007) and PhyloBayes (Lartillot et al., 2008). ML trees are reconstructed from the unreduced data set and the SOS. Phyutility (Smith and Dunn, 2008) was used to identify 'unstable' taxa. For the Bayesian approach, 25 chains were computed. After testing for topological incongruences, a majority rule consensus tree was inferred from 3 chains.

Table A.4.: Ribosomal genes used for data set AP_1_ri extracted from the SOS of AP_1. The annotation correspond to Flybase.

Gene ID	gene annotation
11375	60S acidic ribosomal protein P1
11380	60S ribosomal protein L4
11391	40S ribosomal protein S4
11451	40S ribosomal protein S9
11514	40S ribosomal protein S5a
11552	60S ribosomal protein L23
11580	ribosomal protein L5
11587	40S ribosomal protein S15Aa
11618	Ribosomal protein L6, isoform B
11619	60S ribosomal protein L18a
11635	40S ribosomal protein S26
11637	40S ribosomal protein S23
11639	40S ribosomal protein S18
11707	40S ribosomal protein S7
11710	Ribosomal protein L30, isoform A
11711	60S ribosomal protein L10a-2
11754	40S ribosomal protein S16
11778	60S ribosomal protein L31
11793	Ribosomal protein L18
11844	40S ribosomal protein S8
11910	60S ribosomal protein L8
11911	60S acidic ribosomal protein P0
11965	60S ribosomal protein L27a
11847	40S ribosomal protein S3a
11849	60S ribosomal protein L19
11919	40S ribosomal protein S20
11928	60S ribosomal protein L10
12040	Ribosomal protein L27
12074	CG3195-PA, isoform A 60S ribosomal protein L12
12121	CG12121-PA lung seven transmembrane receptor
12122	60S ribosomal protein L13A
12123	Ribosomal protein S30, isoform B

Table A.5.: Included genes with average information content of data set AP_3_oP and its SOS compared with data set AP_1. IC – Information content (gene); SOS – selected optimal subset. Average IC values of genes that are higher than in data set AP_3_oP are printed in red. Genes that are only present in the SOS of AP_3_oP are blue printed.

Gene ID	dataset AP_1 Average IC	dataset AP_3_oP Average IC	present in SOS of AP_3_oP
gene11349g	0.63	0.75	
gene11350g	0.64	0.61	
gene11351g	0.63	0.93	
gene11352g	0.48	0.66	
gene11353g	0.37	0.59	
gene11354g	0.55	0.77	
gene11355g	0.63	0.80	
gene11356g	0.58	0.55	
gene11357g	0.60	0.69	
gene11358g	0.59	0.68	
gene11359g	0.38	0	
gene11360g	0.57	1.00	
gene11361g	0.47	0.70	
gene11362g	0.88	0.89	x
gene11363g	0.48	0.76	
gene11364g	0.54	0.51	
gene11365g	0.46	0.56	
gene11366g	0.77	0.84	x
gene11367g	0.46	0.91	
gene11368g	0.63	0.75	
gene11369g	0.41	< 4 taxa	
gene11370g	0.54	0.76	
gene11371g	0.42	0.86	
gene11372g	0.60	0.75	
gene11373g	0.30	0.42	
gene11374g	0.46	< 4 taxa	
gene11375g	0.71	0.61	x
gene11376g	0.51	0.69	
gene11377g	0.80	0.82	
gene11378g	0.63	0.75	x
gene11379g	0.76	0.85	x
gene11380g	0.62	0.59	x
gene11381g	0.58	0	
gene11382g	0.55	0.91	
gene11383g	0.53	0.60	x
gene11384g	0.75	0.73	
gene11385g	0.74	0.89	x
gene11386g	0.64	0.68	x
gene11387g	0.86	0.75	
gene11388g	0.27	0	
gene11389g	0.65	0.89	
gene11390g	0.60	0	
gene11391g	0.55	0.63	x
gene11392g	0.67	0.75	
gene11393g	0.76	0.78	x
gene11394g	0.89	0.90	x
gene11395g	0.38	0.90	
gene11396g	0.60	0.76	
gene11397g	0.42	0	
gene11398g	0.80	0	
gene11399g	0.33	0	
gene11400g	0.56	0.75	
gene11401g	0.38	0.39	
gene11402g	0.70	0.82	
gene11403g	0.67	0.75	
gene11404g	0.54	0	
gene11405g	0.83	0.97	
gene11406g	0.73	0.93	
gene11407g	0.59	0.68	
gene11408g	0.48	0.66	
gene11409g	0.53	0.82	
gene11410g	0.31	0	
gene11411g	0.74	0.86	x
gene11412g	0.50	0.52	
gene11413g	0.62	0.86	
gene11414g	0.46	0.68	
gene11415g	0.46	0.63	
gene11416g	0.61	0.58	
gene11417g	0.88	0.86	
gene11418g	0.67	0.77	
gene11419g	0.69	0.79	
gene11420g	0.31	0.37	
gene11421g	0.47	0.41	
gene11422g	0.51	0.87	
gene11423g	0.53	0.44	
gene11424g	0.62	0.77	
gene11425g	0.54	0.78	
gene11426g	0.53	0.71	
gene11427g	0.59	0.31	
gene11428g	0.63	0.73	x
gene11429g	0.69	0.77	
gene11430g	0.30	0.39	
gene11431g	0.52	0.58	
gene11432g	0.64	0.78	
gene11433g	0.46	0.56	
gene11434g	0.54	0.93	
gene11435g	0.48	0.86	
gene11436g	0.72	0.79	
gene11437g	0.69	0.74	x
gene11438g	0.58	0	
gene11439g	0.47	0.72	
gene11440g	0.33	0.56	
gene11441g	0.53	0.56	
gene11442g	0.83	0.83	
gene11443g	0.47	0.61	
gene11444g	0.53	0	
gene11445g	0.43	0.78	
gene11446g	0.54	0.70	
gene11447g	0.60	1.00	
gene11449g	0.49	0.60	
gene11450g	0.75	0.85	x
gene11451g	0.66	0.71	x
gene11452g	0.46	0.46	
gene11609g	0.85	0.92	x
gene11610g	0.73	0.89	
gene11611g	0.39	1.00	
gene11612g	0.35	0.29	
gene11613g	0.63	0.65	
gene11614g	0.58	0.73	
gene11615g	0.54	0.84	
gene11616g	0.76	0.87	
gene11617g	0.77	0.85	x
gene11618g	0.47	0.49	x
gene11619g	0.56	0.57	x
gene11620g	0.34	1.00	
gene11621g	0.48	0.66	
gene11622g	0.37	0.94	
gene11623g	0.46	0.65	
gene11624g	0.85	0.93	
gene11625g	0.61	0.79	
gene11626g	0.47	0.44	
gene11627g	0.72	0.81	
gene11628g	0.67	0.74	
gene11629g	0.63	0.80	
gene11630g	0.60	0	
gene11631g	0.62	0.80	
gene11632g	0.63	0.80	
gene11633g	0.32	0.73	
gene11634g	0.76	0.88	x
gene11635g	0.78	0.75	x
gene11636g	0.49	0.65	
gene11637g	0.86	0.83	x
gene11638g	0.41	0.67	
gene11639g	0.77	0.75	x
gene11640g	0.73	0.76	
gene11641g	0.69	0	
gene11642g	0.74	0.77	x
gene11643g	0.59	0.88	
gene11644g	0.48	0	
gene11645g	0.82	0.92	x
gene11646g	0.53	0.51	
gene11647g	0.62	0.80	
gene11648g	0.39	0.49	
gene11649g	0.58	0.66	x
gene11650g	0.47	0.66	
gene11651g	0.58	0.79	
gene11652g	0.63	0.68	x
gene11653g	0.49	0.61	
gene11654g	0.59	0.74	
gene11655g	0.44	0	
gene11656g	0.56	0.88	
gene11657g	0.58	0.65	
gene11658g	0.31	0.36	
gene11659g	0.50	0.86	
gene11660g	0.74	0.81	
gene11661g	0.40	0.90	
gene11662g	0.41	1.00	
gene11663g	0.61	0.70	
gene11664g	0.44	0.64	
gene11665g	0.25	0.31	
gene11666g	0.60	0.70	
gene11667g	0.65	0.75	
gene11668g	0.57	0.54	
gene11669g	0.50	0.62	
gene11670g	0.45	0.66	
gene11671g	0.60	0.76	
gene11672g	0.73	0.90	
gene11673g	0.75	0.85	
gene11674g	0.56	0.67	
gene11675g	0.54	0.86	
gene11676g	0.48	0.54	x
gene11677g	0.59	1.00	
gene11678g	0.43	0.45	
gene11679g	0.44	0.94	
gene11680g	0.78	0.87	
gene11681g	0.75	0.83	x
gene11682g	0.41	0	
gene11683g	0.56	0.66	
gene11684g	0.46	0.71	
gene11685g	0.51	0.87	
gene11686g	0.57	0.60	
gene11687g	0.76	0.82	
gene11688g	0.50	0.59	
gene11689g	0.40	0	
gene11690g	0.37	0.34	
gene11691g	0.55	0.87	
gene11692g	0.59	0.72	
gene11693g	0.32	0.31	
gene11694g	0.59	< 4 taxa	
gene11695g	0.74	0.77	
gene11696g	0.34	0.92	
gene11697g	0.50	0	
gene11698g	0.52	0.68	
gene11699g	0.51	0.66	
gene11700g	0.74	0.84	
gene11701g	0.62	0.72	
gene11702g	0.30	0.44	
gene11703g	0.35	0.60	
gene11704g	0.55	0.78	
gene11705g	0.62	0.54	
gene11706g	0.35	0.46	
gene11707g	0.61	0.59	x
gene11708g	0.52	0.52	
gene11709g	0.65	0.80	
gene11710g	0.66	0.67	x
gene11711g	0.63	0.66	x
gene11868g	0.82	0.91	
gene11869g	0.56	0.65	x
gene11870g	0.37	0.50	
gene11871g	0.58	0.74	
gene11872g	0.37	0.36	
gene11873g	0.28	0.73	
gene11874g	0.75	0.77	x
gene11875g	0.57	< 4 taxa	
gene11876g	0.38	0.46	
gene11877g	0.45	0.75	
gene11878g	0.58	0.84	
gene11879g	0.30	0.88	
gene11880g	0.63	0.80	
gene11881g	0.61	1.00	
gene11882g	0.57	0.91	
gene11883g	0.63	0	
gene11884g	0.59	0.55	
gene11885g	0.37	0.80	
gene11886g	0.50	0.72	
gene11887g	0.47	0.80	
gene11888g	0.40	0.52	
gene11889g	0.49	0.82	
gene11890g	0.64	0.64	
gene11891g	0.64	0.77	
gene11892g	0.50	0.69	
gene11893g	0.76	0.80	
gene11894g	0.53	0.88	
gene11895g	0.54	0.78	
gene11896g	0.61	0.83	
gene11897g	0.40	0.41	
gene11898g	0.50	0.91	
gene11899g	0.91	0.95	x
gene11900g	0.44	0	
gene11901g	0.43	< 4 taxa	
gene11902g	0.84	0.90	x
gene11903g	0.61	0.82	
gene11904g	0.52	0.66	
gene11905g	0.64	0.66	
gene11906g	0.43	0.42	
gene11907g	0.27	0.29	
gene11908g	0.38	0.76	
gene11909g	0.49	0.66	
gene11910g	0.67	0.62	x
gene11911g	0.72	0.71	x
gene11912g	0.61	0.64	
gene11913g	0.61	0.75	
gene11914g	0.49	0.87	
gene11915g	0.54	0.65	
gene11916g	0.65	0.85	
gene11917g	0.74	0.75	x
gene11918g	0.49	0.88	
gene11919g	0.63	0.63	x
gene11920g	0.58	0.85	
gene11921g	0.73	0.90	
gene11922g	0.61	0.81	
gene11923g	0.37	0.67	
gene11924g	0.92	0.92	x
gene11925g	0.59	0.84	
gene11926g	0.60	0.80	
gene11927g	0.38	0.83	
gene11928g	0.62	0.68	x
gene11929g	0.60	0.74	
gene11930g	0.61	0.79	
gene11931g	0.34	0	
gene11932g	0.62	0.73	
gene11933g	0.40	0.66	
gene11934g	0.72	0.84	
gene11935g	0.67	0.89	
gene11936g	0.49	0.52	
gene11937g	0.54	0.79	
gene11938g	0.47	0.58	
gene11939g	0.62	0.80	
gene11940g	0.74	0.83	
gene11941g	0.49	0.80	
gene11942g	0.43	0.67	
gene11943g	0.44	0.74	
gene11944g	0.59	0.84	
gene11945g	0.29	0.32	
gene11946g	0.44	0.59	
gene11947g	0.45	1.00	
gene11948g	0.43	0.62	
gene11949g	0.41	0.50	
gene11950g	0.39	0.74	
gene11951g	0.37	0.41	
gene11952g	0.50	0.63	
gene11953g	0.58	0.61	
gene11954g	0.80	0.93	
gene11955g	0.65	0.63	
gene11956g	0.76	0.91	
gene11957g	0.46	< 4 taxa	
gene11958g	0.87	0.85	x
gene11959g	0.54	0.64	
gene11960g	0.59	0	
gene11961g	0.46	0	
gene11962g	0.75	0.87	
gene11963g	0.33	1.00	
gene11964g	0.62	< 4 taxa	
gene11965g	0.75	0.69	x
gene11966g	0.38	0.58	
gene11967g	0.39	0.69	
gene11968g	0.38	0.37	
gene11969g	0.41	0.77	
gene11970g	0.52	0.76	

Table A.5 continued

Gene ID	dataset AP_1 Average IC	dataset AP_3_oP Average IC	present in SOS of AP_3_oP	Gene ID	dataset AP_1 Average IC	dataset AP_3_oP Average IC	present in SOS of AP_3_oP	Gene ID	dataset AP_1 Average IC	dataset AP_3_oP Average IC	present in SOS of AP_3_oP
gene11453g	0.44	0		gene11712g	0.64	0.92		gene11971g	0.64	0.80	
gene11454g	0.72	0.79	x	gene11713g	0.44	0.93		gene11972g	0.55	0.90	
gene11455g	0.64	0.92		gene11714g	0.48	0.80		gene11973g	0.41	0.53	
gene11456g	0.53	0.93		gene11715g	0.47	0.55		gene11974g	0.55	0.88	
gene11457g	0.50	< 4 taxa		gene11716g	0.65	0.80		gene11975g	0.63	0	
gene11458g	0.54	0.76		gene11717g	0.40	0.49		gene11976g	0.65	0.65	
gene11459g	0.49	1.00		gene11718g	0.70	0.89		gene11977g	0.45	0.77	
gene11460g	0.86	0.91	x	gene11719g	0.60	0.63		gene11978g	0.60	0.76	
gene11461g	0.55	0.76		gene11720g	0.48	0.80		gene11979g	0.42	0.69	
gene11462g	0.41	0.77		gene11721g	0.52	0.72		gene11980g	0.72	0.86	
gene11463g	0.51	< 4 taxa		gene11722g	0.54	1.00		gene11981g	0.46	0.86	
gene11464g	0.46	0.68		gene11723g	0.51	0.73		gene11982g	0.50	0.83	
gene11465g	0.48	0.89		gene11724g	0.52	0.53		gene11983g	0.77	0.79	x
gene11466g	0.43	0.77		gene11725g	0.56	0		gene11984g	0.60	0.56	x
gene11467g	0.54	1.00		gene11726g	0.73	0.89		gene11985g	0.32	0.35	
gene11468g	0.39	0.74		gene11727g	0.60	0.60		gene11986g	0.41	0	
gene11469g	0.48	0.74		gene11728g	0.65	0.60		gene11987g	0.57	0.68	
gene11470g	0.37	1.00		gene11729g	0.46	0.44		gene11988g	0.67	0.83	
gene11471g	0.49	0.70		gene11730g	0.49	0		gene11989g	0.78	0.87	
gene11472g	0.42	0		gene11731g	0.59	0.62		gene11990g	0.72	0.84	
gene11473g	0.50	0.56		gene11732g	0.44	0		gene11991g	0.75	0.79	
gene11474g	0.47	0.47		gene11733g	0.68	0.76		gene11992g	0.61	0.70	x
gene11475g	0.40	0.82		gene11734g	0.57	0.65		gene11993g	0.66	0.71	
gene11476g	0.35	0.91		gene11735g	0.90	0.91	x	gene11994g	0.41	0.43	
gene11477g	0.57	0.72		gene11736g	0.27	0.72		gene11995g	0.24	0.21	
gene11478g	0.59	0.64		gene11737g	0.62	0.64		gene11996g	0.61	1.00	
gene11479g	0.72	0.69		gene11738g	0.44	0.75		gene11997g	0.59	0.85	
gene11480g	0.67	0.86		gene11739g	0.62	< 4 taxa		gene11998g	0.48	0	
gene11481g	0.36	0.38		gene11740g	0.50	0.81		gene11999g	0.40	0.75	
gene11482g	0.67	0.81		gene11741g	0.48	0.71		gene12000g	0.59	0.82	
gene11483g	0.60	0.76		gene11742g	0.62	0.86		gene12001g	0.45	0	
gene11484g	0.84	0.89	x	gene11743g	0.46	0.68		gene12002g	0.68	0.84	
gene11485g	0.41	0.42		gene11744g	0.33	0.84		gene12003g	0.66	0.83	
gene11486g	0.41	0.74		gene11745g	0.52	0.79		gene12004g	0.47	0.49	
gene11487g	0.55	0.79		gene11746g	0.40	0.80		gene12005g	0.31	0.71	
gene11488g	0.59	0.60		gene11747g	0.43	0.83		gene12006g	0.33	1.00	
gene11489g	0.56	0.76		gene11748g	0.44	0.58		gene12007g	0.74	0.77	
gene11490g	0.72	0		gene11749g	0.35	0.35		gene12008g	0.50	< 4 taxa	
gene11491g	0.90	0.89		gene11750g	0.75	0.82	x	gene12009g	0.84	0.91	
gene11492g	0.41	0		gene11751g	0.32	0.38		gene12010g	0.67	0.76	
gene11493g	0.46	0.67		gene11752g	0.58	< 4 taxa		gene12011g	0.60	0.68	
gene11494g	0.46	0.81		gene11753g	0.26	0.67		gene12012g	0.73	0.85	x
gene11495g	0.66	0.73		gene11754g	0.60	0.63	x	gene12013g	0.60	0.66	
gene11496g	0.77	0.93		gene11755g	0.82	0.60		gene12014g	0.61	0.75	
gene11497g	0.38	0.60		gene11756g	0.52	0.62		gene12015g	0.34	0.83	
gene11498g	0.37	0.35		gene11757g	0.66	0.77		gene12016g	0.61	0.94	
gene11499g	0.63	0.63	x	gene11758g	0.41	0		gene12017g	0.48	0.82	
gene11500g	0.48	0		gene11759g	0.78	0.89	x	gene12018g	0.65	0.79	x
gene11501g	0.63	0.65		gene11760g	0.82	0.87	x	gene12019g	0.71	0.85	x
gene11502g	0.66	0.73		gene11761g	0.59	0.67		gene12020g	0.41	0.38	
gene11503g	0.31	0.87		gene11762g	0.81	0.89	x	gene12021g	0.32	0.38	
gene11504g	0.30	0.30		gene11763g	0.74	0.90		gene12022g	0.47	0.61	
gene11505g	0.43	< 4 taxa		gene11764g	0.53	0.53		gene12023g	0.50	0.78	
gene11506g	0.51	0		gene11765g	0.41	0		gene12024g	0.89	0.93	
gene11507g	0.65	0.70		gene11766g	0.45	1.00		gene12025g	0.46	0.55	
gene11508g	0.73	0		gene11767g	0.59	1.00		gene12026g	0.52	0.66	
gene11509g	0.79	0.76		gene11768g	0.72	0.88		gene12027g	0.25	0.80	
gene11510g	0.42	0.47		gene11769g	0.53	0.57		gene12028g	0.28	0.42	
gene11511g	0.85	0.93	x	gene11770g	0.46	0.48		gene12029g	0.65	0.70	x
gene11512g	0.60	0.70		gene11771g	0.63	< 4 taxa		gene12030g	0.40	0.71	
gene11513g	0.50	0.64		gene11772g	0.62	0.76	x	gene12031g	0.63	0.61	
gene11514g	0.75	0.72	x	gene11773g	0.60	0.61	x	gene12032g	0.35	0	
gene11515g	0.48	0.87		gene11774g	0.56	0		gene12033g	0.69	0.69	
gene11516g	0.17	< 4 taxa		gene11775g	0.52	0.79		gene12034g	0.35	0	
gene11517g	0.46	0.70		gene11776g	0.54	0.68		gene12035g	0.57	0.60	x
gene11518g	0.59	0.82		gene11777g	0.36	< 4 taxa		gene12036g	0.46	< 4 taxa	
gene11519g	0.27	0.21		gene11778g	0.53	0.57	x	gene12037g	0.49	0.52	
gene11520g	0.62	0.78		gene11779g	0.47	0.52		gene12038g	0.70	0.74	
gene11521g	0.53	0.76		gene11780g	0.40	0		gene12039g	0.49	0.57	
gene11522g	0.46	0.77		gene11781g	0.42	0.89		gene12040g	0.58	0.62	x
gene11523g	0.72	0.78		gene11782g	0.51	< 4 taxa		gene12041g	0.37	0.82	
gene11524g	0.59	0.79		gene11783g	0.81	0.87		gene12042g	0.62	0.73	
gene11525g	0.64	0.83		gene11784g	0.48	0		gene12043g	0.52	0.86	
gene11526g	0.52	1.00		gene11785g	0.70	< 4 taxa		gene12044g	0.29	0.43	
gene11527g	0.62	0		gene11786g	0.64	0.85		gene12045g	0.54	< 4 taxa	
gene11528g	0.53	0.53		gene11787g	0.67	0.85		gene12046g	0.61	0.70	
gene11529g	0.63	0.75		gene11788g	0.73	0.70		gene12047g	0.65	0.75	x
gene11530g	0.61	0.75		gene11789g	0.51	0.42		gene12048g	0.57	0.73	
gene11531g	0.63	0.77		gene11790g	0.48	0		gene12049g	0.62	0.61	
gene11532g	0.42	0.46		gene11791g	0.38	0.83		gene12050g	0.58	0.66	
gene11533g	0.47	0.86		gene11792g	0.60	0.86		gene12051g	0.65	0.74	
gene11534g	0.69	0.78	x	gene11793g	0.42	0.41	x	gene12052g	0.40	0.81	
gene11535g	0.70	0.90		gene11794g	0.56	1.00		gene12053g	0.63	0.73	
gene11536g	0.70	0.77		gene11795g	0.56	0.48		gene12054g	0.72	0.77	
gene11537g	0.30	0.33		gene11796g	0.71	0.87		gene12055g	0.25	0	
gene11538g	0.59	0.83		gene11797g	0.78	0.86		gene12056g	0.59	0.62	
gene11539g	0.49	0.75		gene11798g	0.72	0.67	x	gene12057g	0.64	0.80	
gene11540g	0.40	0.83		gene11799g	0.63	0.81		gene12058g	0.47	0.61	
gene11541g	0.39	0.43		gene11800g	0.29	0.70		gene12059g	0.68	0	
gene11542g	0.74	0.89		gene11801g	0.64	0.66		gene12060g	0.75	0.79	
gene11543g	0.34	0.40		gene11802g	0.67	0.72		gene12061g	0.92	0.93	x
gene11544g	0.56	0.77		gene11803g	0.81	0.86		gene12062g	0.61	0.77	
gene11545g	0.66	0.75		gene11804g	0.56	0.62		gene12063g	0.55	0.84	
gene11546g	0.54	0.52		gene11805g	0.67	0.85		gene12064g	0.42	0.42	
gene11547g	0.86	0.88	x	gene11806g	0.90	0.94	x	gene12065g	0.74	0.75	
gene11548g	0.35	0.43		gene11807g	0.46	0.57		gene12066g	0.61	< 0.04	
gene11549g	0.70	0.77		gene11808g	0.66	0.69		gene12067g	0.53	1.00	
gene11550g	0.56	< 4 taxa		gene11809g	0.53	0.93		gene12068g	0.70	0.81	
gene11551g	0.47	0.55		gene11810g	0.57	0		gene12069g	0.65	0.71	
gene11552g	0.83	0.84	x	gene11811g	0.52	0.56	x	gene12070g	0.76	0.93	
gene11553g	0.65	0.83		gene11812g	0.26	0.97		gene12071g	0.86	0.89	x
gene11554g	0.39	0.84		gene11813g	0.50	0.75		gene12072g	0.39	0.85	
gene11555g	0.67	0.74	x	gene11814g	0.66	0.70		gene12073g	0.83	0.87	

Table A.5 continued

Gene ID	dataset AP_1 Average IC	dataset AP_3_oP Average IC	present in SOS of AP_3_oP
gene11556g	0.65	0.89	
gene11557g	0.34	0.80	
gene11558g	0.48	0.68	
gene11559g	0.56	0.78	
gene11560g	0.37	0.40	
gene11561g	0.59	0.54	
gene11562g	0.44	0.47	
gene11563g	0.70	0.74	x
gene11564g	0.73	0.79	
gene11565g	0.61	0.77	
gene11566g	0.48	0.49	
gene11567g	0.66	0.71	
gene11568g	0.40	0.42	
gene11569g	0.35	0.65	
gene11570g	0.59	0.68	
gene11571g	0.63	0.78	
gene11572g	0.70	0.83	
gene11573g	0.68	0.79	
gene11574g	0.64	0.84	
gene11575g	0.33	0.44	
gene11576g	0.62	0.73	
gene11577g	0.69	0.70	
gene11578g	0.72	0.82	
gene11579g	0.54	0.57	
gene11580g	0.65	0.66	x
gene11581g	0.36	0.52	
gene11582g	0.41	0.47	
gene11583g	0.71	0.85	
gene11584g	0.54	0.59	
gene11585g	0.32	0.61	
gene11586g	0.54	0.86	
gene11587g	0.73	0.71	x
gene11588g	0.65	0.85	
gene11589g	0.74	0.84	
gene11590g	0.53	0.67	
gene11591g	0.68	0.65	x
gene11592g	0.52	0.89	
gene11593g	0.36	0.56	
gene11594g	0.33	0	
gene11595g	0.32	0.31	
gene11596g	0.52	0.64	
gene11597g	0.40	0.54	
gene11598g	0.64	0.89	
gene11599g	0.41	0.70	
gene11600g	0.60	0.79	
gene11601g	0.52	0.75	
gene11602g	0.34	0.83	
gene11603g	0.68	0.75	
gene11604g	0.67	0.74	
gene11605g	0.43	1.00	
gene11606g	0.60	0	
gene11607g	0.61	0.72	
gene11608g	0.63	0.91	
gene11815g	0.68	0.84	
gene11816g	0.87	0.91	
gene11817g	0.59	0	
gene11818g	0.42	0.83	
gene11819g	0.30	0	
gene11820g	0.61	0.66	x
gene11821g	0.72	0.81	
gene11822g	0.62	0.82	
gene11823g	0.67	0.74	
gene11824g	0.67	0.89	
gene11825g	0.49	0.66	
gene11826g	0.51	0.60	
gene11827g	0.54	0.55	
gene11828g	0.55	0.64	
gene11829g	0.73	0.86	x
gene11830g	0.65	0.88	
gene11831g	0.63	0.82	
gene11832g	0.32	0.64	
gene11833g	0.48	0.46	
gene11834g	0.59	0.86	
gene11835g	0.55	0.69	
gene11836g	0.25	0.83	
gene11837g	0.49	0.73	
gene11838g	0.68	0.74	
gene11839g	0.56	0.56	
gene11840g	0.60	0.59	
gene11841g	0.46	0.37	
gene11842g	0.38	0.41	
gene11843g	0.63	0.72	
gene11844g	0.53	0.54	x
gene11845g	0.42	0.64	
gene11846g	0.89	0.92	x
gene11847g	0.59	0.66	x
gene11848g	0.72	0.77	x
gene11849g	0.65	0.63	x
gene11850g	0.77	0.80	
gene11851g	0.36	0.83	
gene11852g	0.50	0.81	
gene11853g	0.64	1.00	
gene11854g	0.66	0.87	
gene11855g	0.71	0.76	
gene11856g	0.46	0.81	
gene11857g	0.66	0.79	
gene11858g	0.56	0.86	
gene11859g	0.69	0.77	
gene11860g	0.81	0.93	
gene11861g	0.54	0.60	
gene11862g	0.48	0.61	
gene11863g	0.58	0	
gene11864g	0.74	< 4 taxa	
gene11865g	0.79	0.86	
gene11866g	0.55	0.59	
gene11867g	0.50	0	
gene12074g	0.44	0.51	x
gene12075g	0.62	0.86	
gene12076g	0.65	0.84	
gene12077g	0.48	0.78	
gene12078g	0.55	0.80	
gene12079g	0.59	0.63	
gene12080g	0.59	0.60	
gene12081g	0.56	0.61	x
gene12082g	0.68	0.66	
gene12083g	0.53	0.93	
gene12084g	0.78	0.95	
gene12085g	0.48	0.76	
gene12086g	0.44	0.56	
gene12087g	0.79	0.88	
gene12088g	0.70	0.94	
gene12089g	0.41	0.85	
gene12090g	0.57	0.66	
gene12091g	0.36	0.54	
gene12092g	0.62	0	
gene12093g	0.57	0.62	
gene12094g	0.59	0.64	
gene12095g	0.37	0.41	
gene12096g	0.52	0.66	
gene12097g	0.76	0.86	x
gene12098g	0.53	0.69	
gene12099g	0.54	0.90	
gene12100g	0.41	0.67	
gene12101g	0.48	0.47	
gene12102g	0.62	0.83	
gene12103g	0.61	0.63	
gene12104g	0.47	0.79	
gene12105g	0.67	0.80	
gene12106g	0.53	0.60	
gene12107g	0.60	0.76	
gene12108g	0.68	0	
gene12109g	0.58	0.87	
gene12110g	0.59	0.83	
gene12111g	0.56	0.72	
gene12112g	0.63	0.70	
gene12113g	0.32	0.75	
gene12114g	0.52	0.76	
gene12115g	0.57	0.70	x
gene12116g	0.52	0.66	
gene12117g	0.38	0.73	
gene12118g	0.80	0.89	
gene12119g	0.42	0.90	
gene12120g	0.67	0.79	x
gene12121g	0.73	0.75	x
gene12122g	0.45	0.52	x
gene12123g	0.44	0.41	x
gene12124g	0.41	0.41	

Table A.6.: Additional taxa included in the SOS of `AP_3_oP` and not present in the SOS of `AP_1`.

species	group			order
Epiperipatus sp. TB-2001	Onychophora			
Loxosceles laeta	Chelicerata	Arachnida		Araneae
Ixodes ricinus	Chelicerata	Arachnida	Acari	Ixodida
Amblyomma cajennense	Chelicerata	Arachnida	Acari	Ixodida
Argas monolakensis	Chelicerata	Arachnida	Acari	Ixodida
Ornithodoros parkeri	Chelicerata	Arachnida	Acari	Ixodida
Dermatophagoides farinae	Chelicerata	Arachnida	Acari	Astigmata
Acarus siro	Chelicerata	Arachnida	Acari	Astigmata
Celuca pugilator	Crustacea	Malacostraca		Decapoda
Eriocheir sinensis	Crustacea	Malacostraca		Decapoda
Litopenaeus setiferus	Crustacea	Malacostraca		Decapoda
Pediculus humanus capitis	Hexapoda	Neoptera		Phthiraptera
Rhodnius prolixus	Hexapoda	Neoptera		Hemiptera
Diaphorina citri	Hexapoda	Neoptera		Hemiptera
Hypothenemus hampei	Hexapoda	Neoptera		Coleoptera
Ips pini	Hexapoda	Neoptera		Coleoptera
Callosobruchus maculatus	Hexapoda	Neoptera		Coleoptera
Papilio dardanus	Hexapoda	Neoptera		Lepidoptera
Trichoplusia ni	Hexapoda	Neoptera		Lepidoptera
Agrotis segetum	Hexapoda	Neoptera		Lepidoptera
Lonomia obliqua	Hexapoda	Neoptera		Lepidoptera
Ctenocephalides felis	Hexapoda	Neoptera		Siphonaptera
Anopheles funestus	Hexapoda	Neoptera		Diptera
Sitodiplosis mosellana	Hexapoda	Neoptera		Diptera

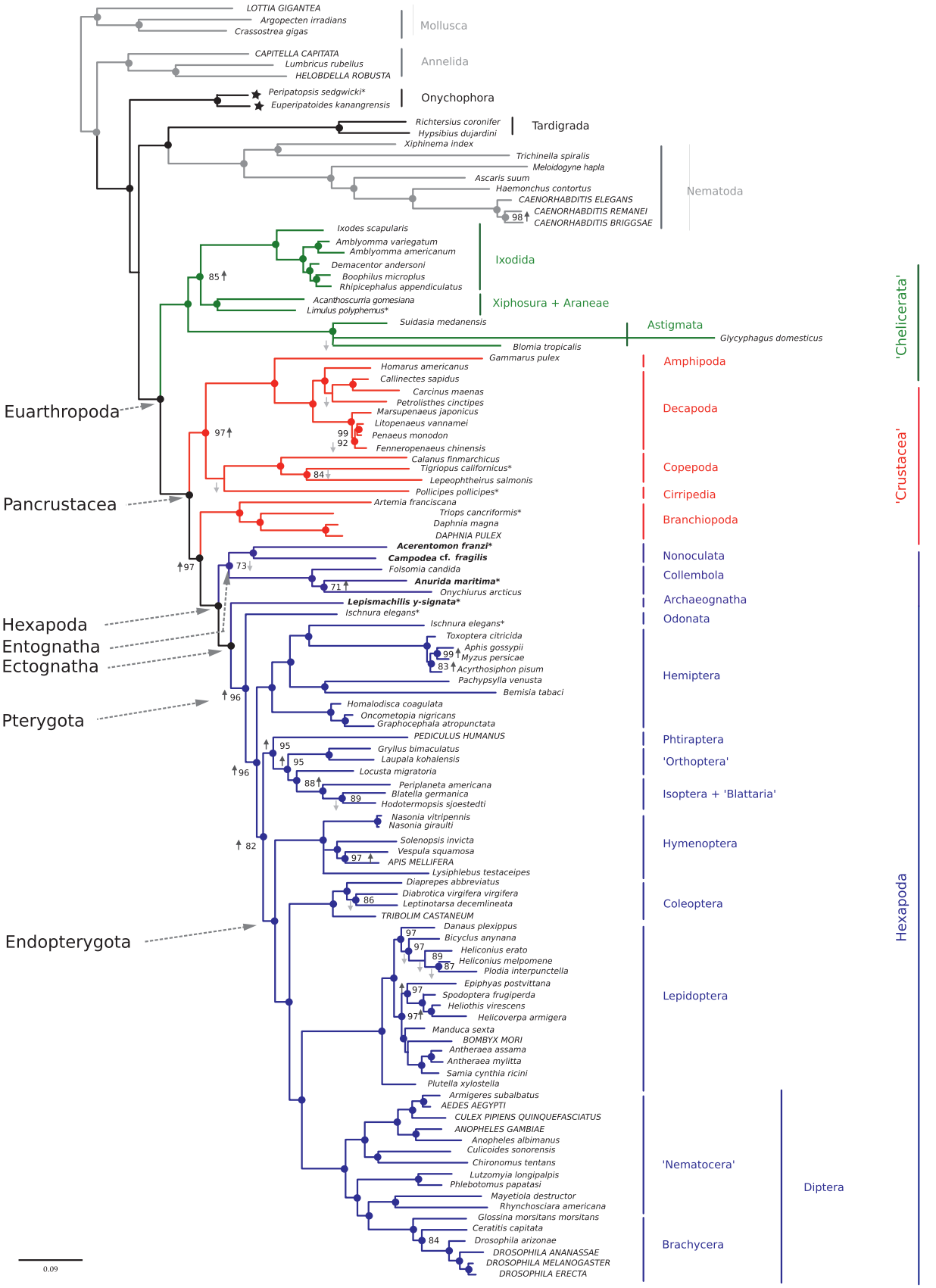

Figure A.4.: Phylogram (majority rule) inferred from the 112-taxon ML analysis based on the SOS of AP_1 while unstable taxa (myriapods, pycnogonids and the mayfly *Baetis* have been excluded. Support values were derived from 1,000 bootstrap replicates. Support values, color code and labeling are specified in Fig. 2.19. Black arrows: increased BS values, gray arrows: decreased BS values compared to SOS of AP_1 (Fig. 2.20). Onychophora appear unstable (leaf stability index < 0.95, see 2.2.7), marked by a star in front of the taxon name.

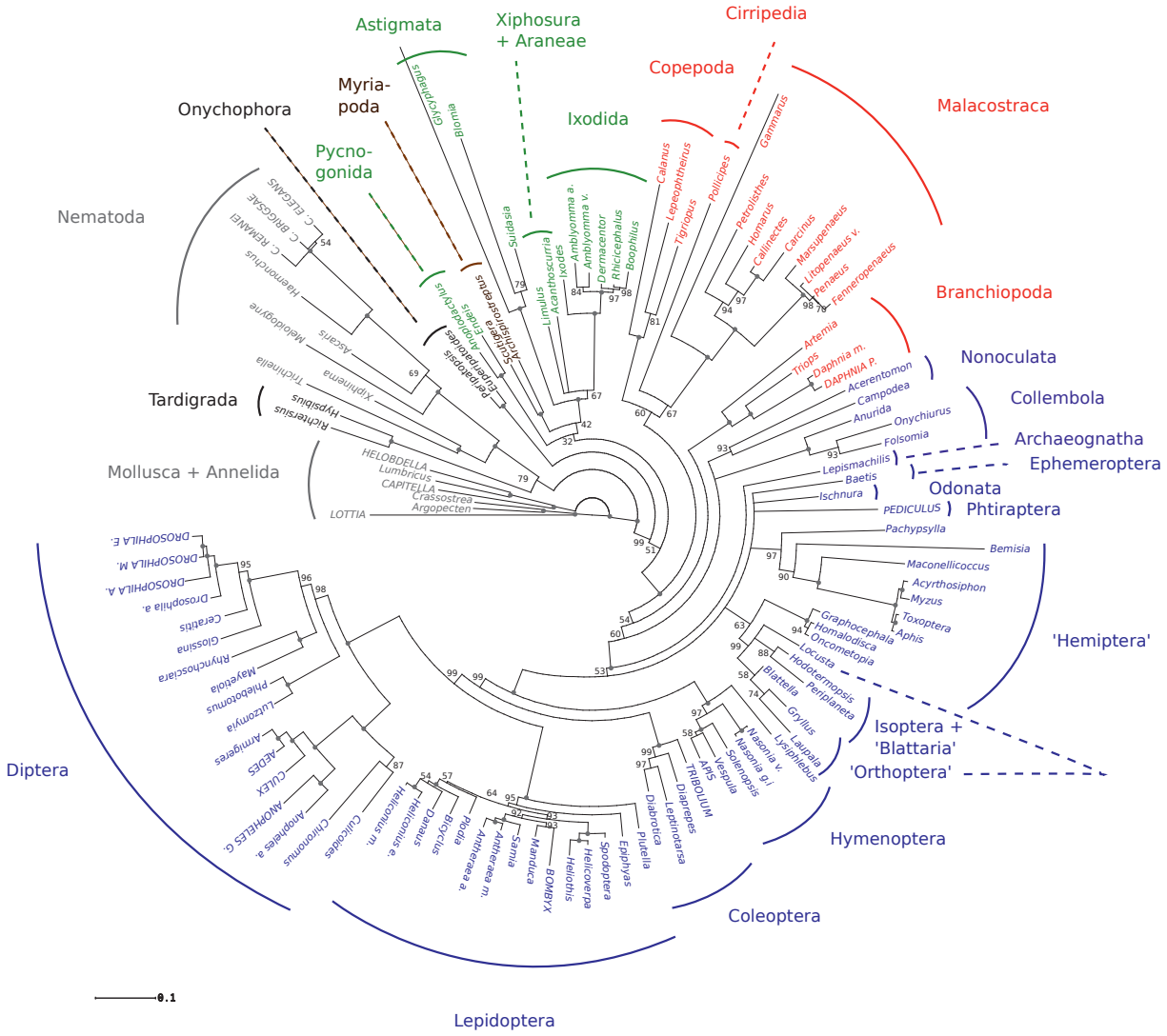

Figure A.5.: ML majority rule tree derived from 32 ribosomal protein coding genes, data set `AP_1_ri`. Support values were derived from 1,000 bootstrap replicates. Support values, color code and labeling are specified in Fig. 2.19.

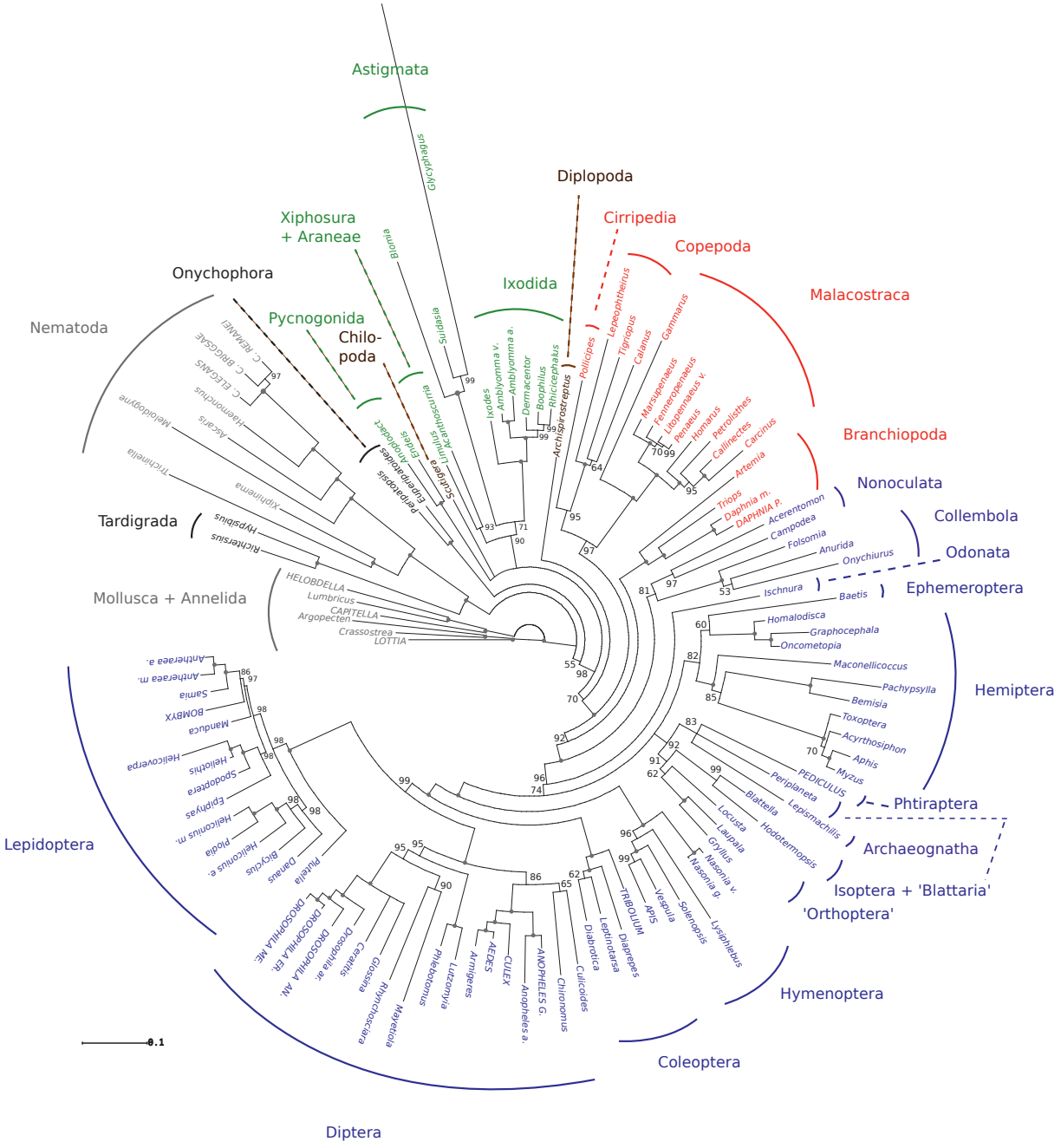

Figure A.6.: ML majority rule tree derived from 97 non-ribosomal protein coding genes, dataset `AP_1_nri`. Support values were derived from 1,000 bootstrap replicates. Support values, color code and labeling are specified in Fig. 2.19.

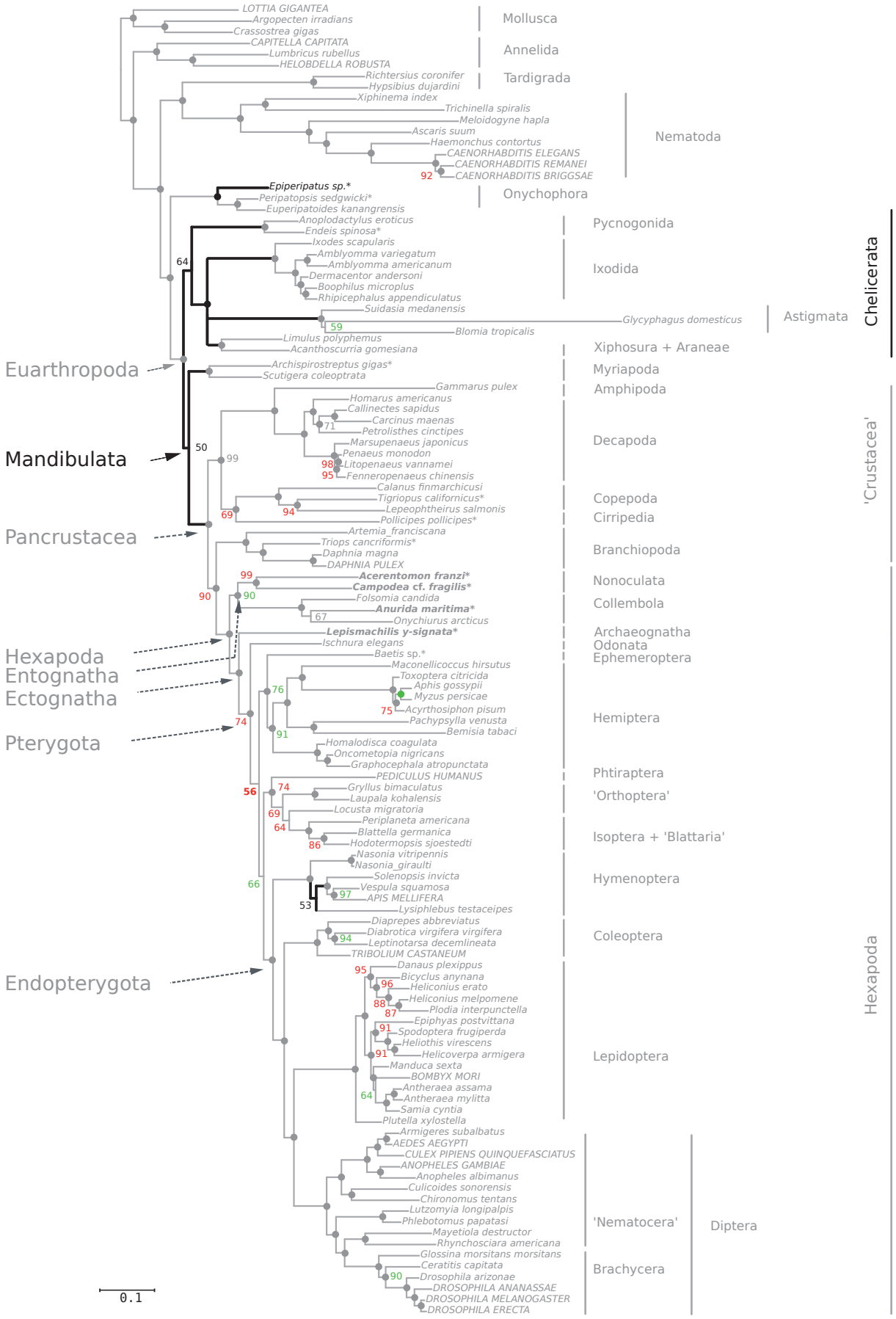

Figure A.7.: Phylogram (majority rule, 1,000 bootstrap replicates) of 118-taxon ML analysis (127 genes) based on the SOS of AP_2. Topological changes compared with Fig. 2.20 are in black; increased BS values: green; decreased BS values: red. Asterisks * indicate EST taxa contributed members of the 'Arthropod DMP Network'; own EST-taxa are in bold-print. Support values and labeling are specified in Fig. 2.19.

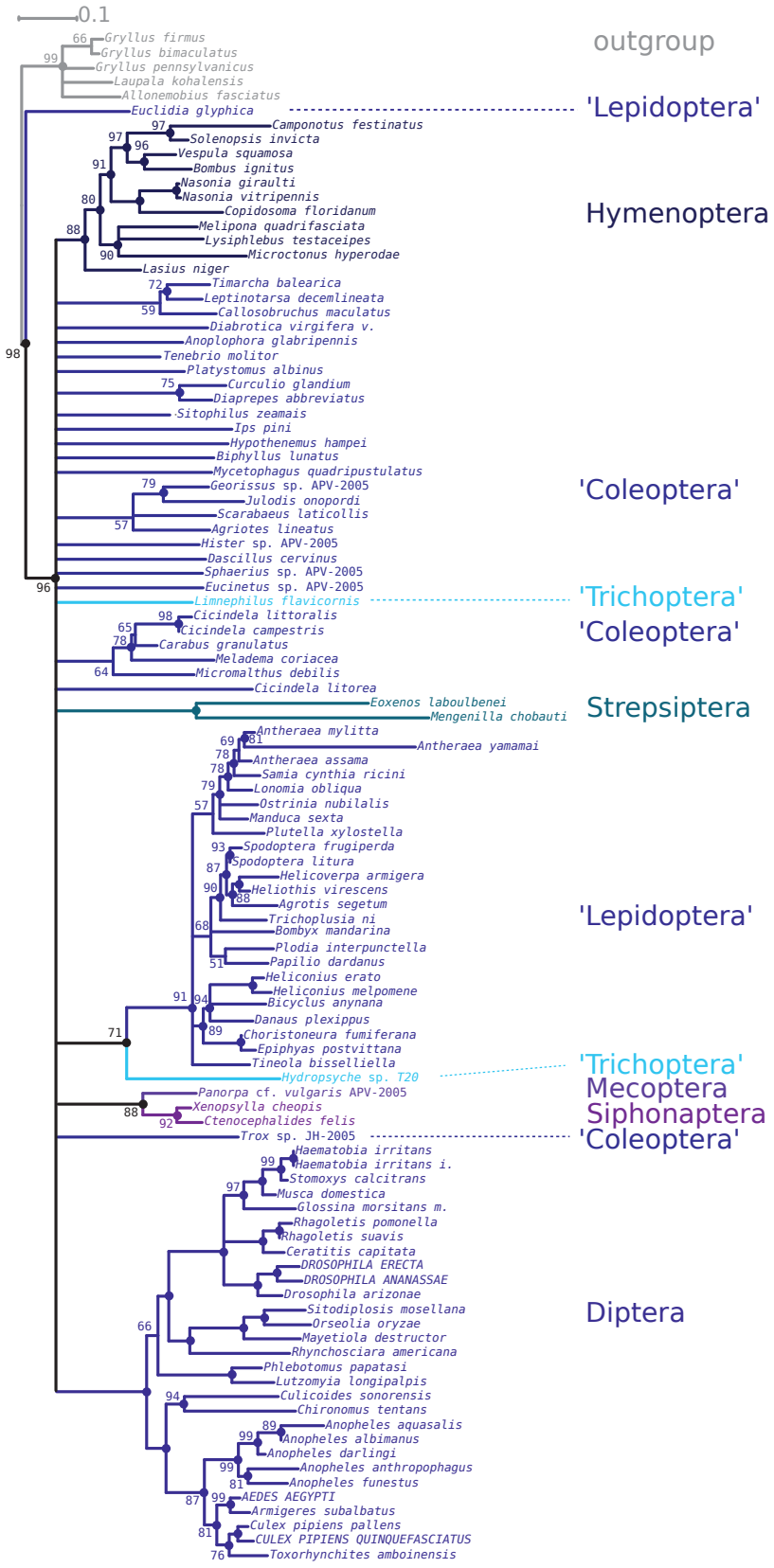

Figure A.8.: Phylogram (majority rule) of 106-taxon ML analysis (775 genes, 100 bootstrap replicates) based on data set `En_oP`. Outgroup species: Ensifera (Orthoptera). Support values are derived from 100 bootstrap replicates. For labeling see Fig. 2.26.

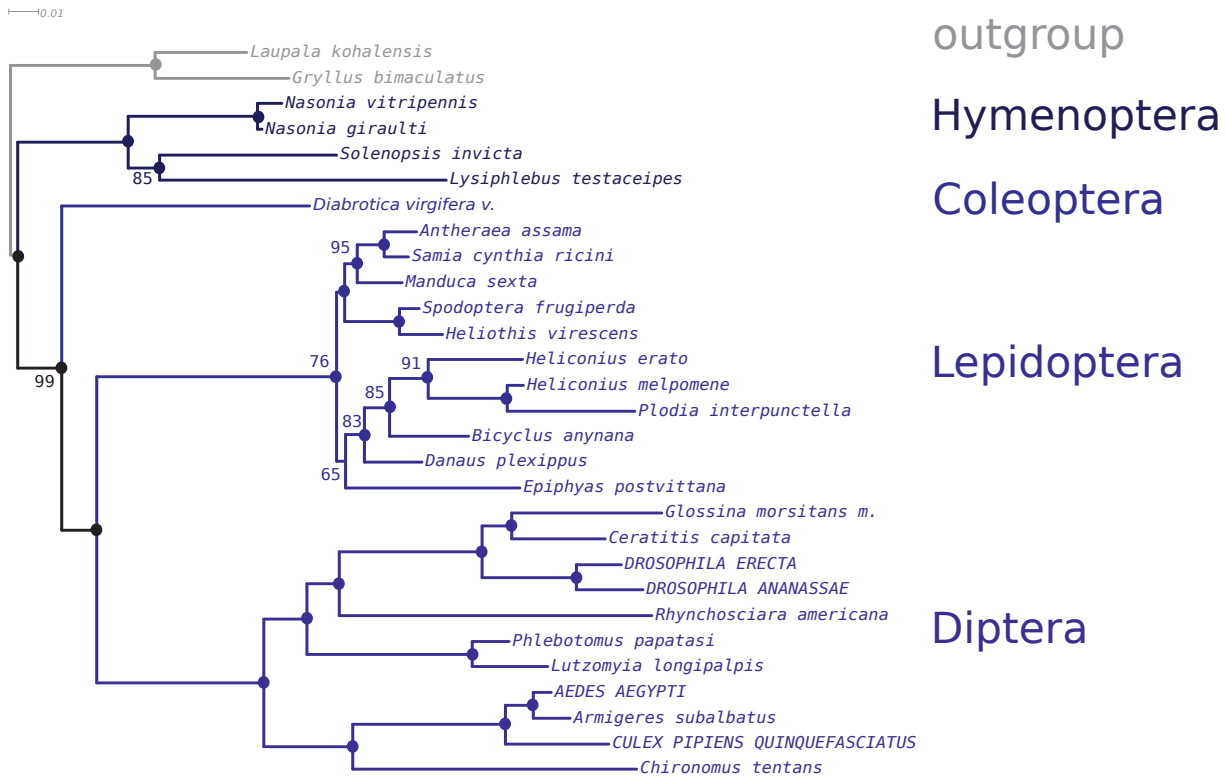

Figure A.9.: Phylogram (majority rule) of 29-taxon ML analysis (63 genes, 1,000 bootstrap replicates) based on the SOS of `En_oP` (Tab. 2.5). Outgroup species: Ensifera (Orthoptera). Support values, color code and labeling are specified in Fig. A.8.

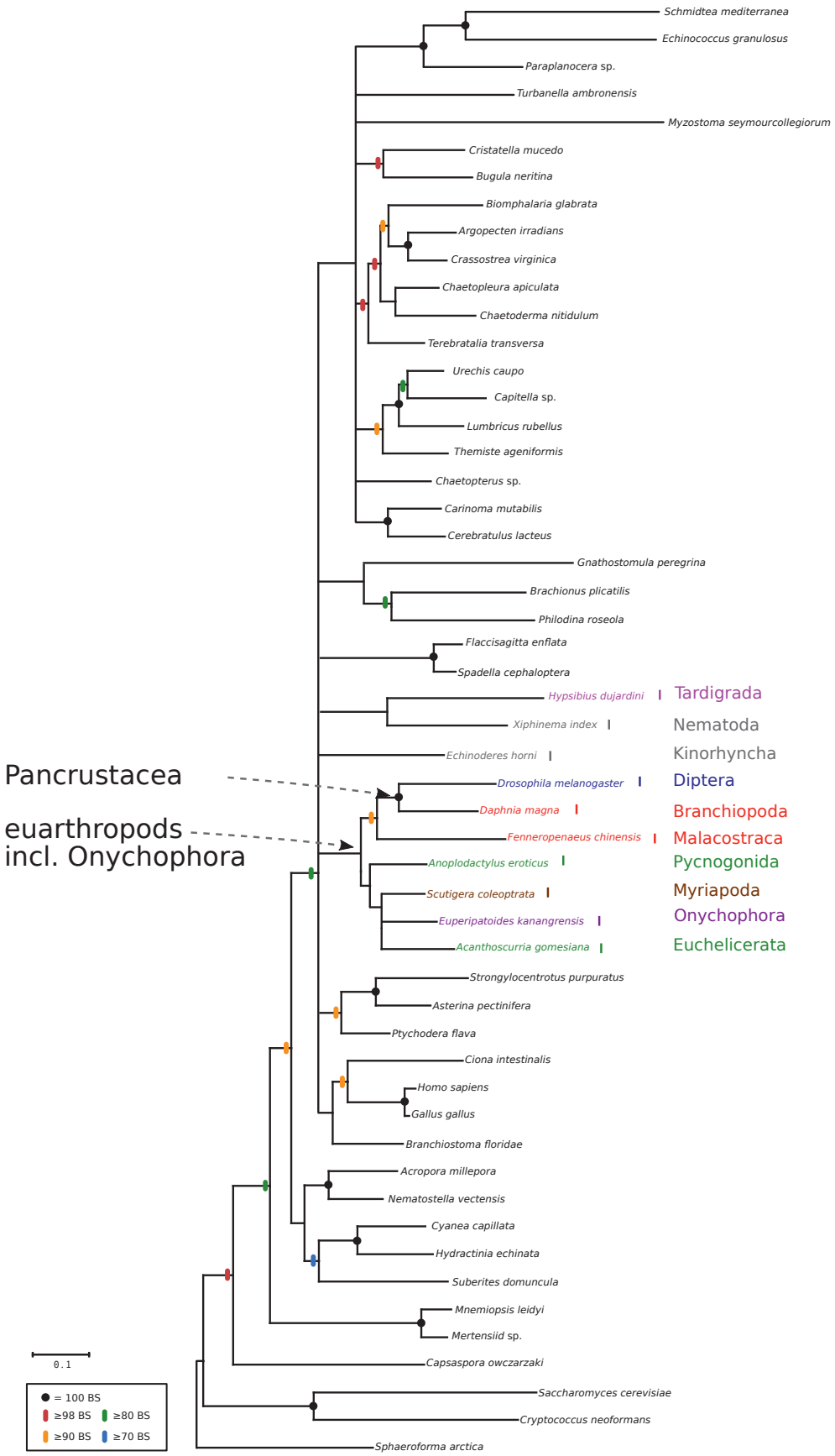

Figure A.10.: Phylogram (majority rule) of a 53-taxon ML analysis (33 genes, 1,000 bootstrap replicates) based on the SOS without taxa weighting of Dunns data set (Dunn et al., 2008). For labeling and color code, see Fig. 2.28.

A.2. Arthropod Phylogeny inferred from nuclear rRNAs genes

Some Figures and Tables have been published in von Reumont et al. (2009).

Table A.7.: Sample locations of taxa used for amplification of the 18S and 28S rRNA gene.

Order	Taxon	Locality	Collection date	Collector	Remarks
Pycnogonida	Colossendeis sp.	ANDEEP I Expedition, Ant XIX-3, Antarctica	29.01.2002	M. Raupach	
Pycnogonida	Nymphon stroemii	Hinlopen Svalbard, Arctica	23.09.2003	d´Dudekem d´Acor	
Notostraca	Triops cancriformis	Marchauen, Austria	2005	E. Eder	
Diplostraca	Daphnia cf. magna	Bonn, Nord-Rhein-Westfalia, Germany	05.10.2005	B. v. Reumont	
	Bosmina sp.	Tegler See, Berlin, Germany	2005	A. Braband	
Ostracoda	Heterocypris incongruens	Hirschweiher, Röttgen, Nord-Rhein-Westfalia, Germany	2005	B. v. Reumont	
	Pontocypris mytiloides	Wilhelmshaven, Niedersachsen, Germany	2007	B. v. Reumont	
Cirripedia	Semibalanus balanoides	Horumersiel, Niedersachsen, Germany	2007	B. v. Reumont	
	Pollicipes pollicipes	Ferrol supermercado, Galicia, Spain	2006	B. v. Reumont	
Branchiura	Argulus cf. foliaceus	Sweden	2007	D. Waloßek	
Mystacocarida	Derocheilocaris typicus	Playa dos ninos, Ferrol, Galicia, Spain	2006	B. v. Reumont	
Copepoda	Tigriopus cf. fulvus	Galicia, Spain	2006	B. v. Reumont	
	Canuella perplexa	Hooksiel, Niedersachsen, Germany	09.05.2006	B. v. Reumont	
Remipedia	Speleonectes tulumensis	Cenote Eden, Puerto Aventuras, Quintana Roo, Mexico	2006	S. Koenemann	
Leptostraca	Nebalia sp.	Ferrol, Galicia, Spain	2006	B. v. Reumont	
Pentastomida	Railltietiella sp.	Asia, host: Hemidactylus cf. frenatus	2007	B. v. Reumont	
Chilopoda	Craterostigmus tasmanianus	Tasmania, Australia			
	Lithobius forficatus	Breitenfurt near Vienna, backyard, Niederösterreich, Austria	28.07.2004	N. Szucsich	
Diplopoda	Polyxenus lagurus	Bonn-Plittersdorf, graveyard, Nord-Rhein-Westfalia, Germany	31.05.2005	B. Huber	
	Monographis sp.	Shanghai, China	2005	Y. Yang	
	Polydesmus complanatus	Breitenfurt near Vienna, Niederösterreich, Austria	November 2006	N. Szucsich	
	Cylindroiulus caeruleocinctus	Breitenfurt near Vienna, urban area, Niederösterreich, Austria	23.10.2004	N. Szucsich	
Pauropoda	Pauropodidae sp.	Panzergraben, Neusiedl am See, Burgenland, Austria	26.04.2006	D. Bartel, N. Szucsich	
Protura	Acerentomon franzi	Lavanttal, Kärnten, Austria	09.10.2005	M. Walzl	
	Baculentulus densus	Shinkoji, Sanada, Udea Nagano, Japan	05.05.2006	R. Machida, M. Fikui	
	Eosentomon sp.	Lavanttal, Kärnten, Austria	09.10.2005	M. Walzl	
	Eosentomon sakura	Zhanjiang Guangdong, China	2002	Y. Luan, Y. Yang	
	Sinentomon erythranum	Suzhou Jiangsu and Hangzhou Zhejiang, China	2002 - 2006	Y. Luan, Y. Yang	
Diplura	Campodea augens	Breitenfurt near Vienna, forest, Niederösterreich, Austria	01.08.2004	N. Szucsich	
	Lepidocampa weberi	Shinoda, Shizuoka, Japan	20.03.2006	K. Sekiya, R. Machida	
	Catajapyx aquilonaris	Leopoldsberg XIX. Bezirk, Vienna, Austria	11.11.2004	M. Hable	
	Parajapyx emeryanus	Shanghai, China	2005	Y. Luan, Y. Yang	
	Octostigma sinensis	Zhanjiang Guangdong, China	2002	Y. Yang	
Collembola	Tetrodontophora bielanensis	Görlitz, Sachsen, Germany	2006	W. Dunger	
	Gomphiocephalus hodgsoni	Victoria Land, Antarcica		F. Frati	
	Billobella aurantiaca	Feniglia, Grosseto, Toscana, Italy	2000	E. Dell'Ampio	
	Anurida maritima	Livorno, Toscana, Italy		R. Dallai	28S
	Anurida maritima	Texel, ferryport, Noord-Nederland, Netherlands	30.08.2006	K. Meusemann	18S
	Podura aquatica	XXII. Bezirk, Vienna, Austria	27.08.2004	M. Sztatecsny, N. Szucsich	28S
	Podura aquatica	T Hoorntje South Texel, Noord-Nederland, Netherlands	30.08.2006	M. Berg	18S
	Cryptopygus antarcticus	Killingbeck Island, Antarcica		A. Carapelli	28S
	Cryptopygus antarcticus	King Georg Islands, Antarcica	2005	M. Raupach	18S
	Isotoma viridis	Rheinbach, Nord-Rhein-Westfalia, Germany	13./14.02.2006	H. Kliebhan	
	Orchesella villosa	Montalbuccio, Toscana, Italy	15.09.2004	E. Dell'Ampio	
	Pogonognathellus flavescens	Breitenfurt near Vienna, Niederösterreich, Austria	01.08.2004	N. Szucsich	
	Megalothorax minimus	Vienna, Austria	27.04.2004, 18.05.2005	N. Szucsich	
	Sminthurus viridis	Breitenfurt near Vienna, Niederösterreich, Austria	05.08.2004	N. Szucsich	
	Allacma fusca	Feniglia, Toscana, Italy	Autumn 2005	P. P. Fanciulli	
	Dicyrtomina saundersi	Siena, Toscana, Italy		P. P. Fanciulli	
Archaeognatha	Machilis hrabei	Leopoldsberg XIX. Bezirk, Vienna, Austria	02.09.2005	N. Szucsich	
	Lepismachilis y-signata	XIII. Bezirk, Vienna, Austria	14.08.2004	N. Szucsich	
	Pedetontus okajimae	Shimoda, Shizuoka, Japan	20.03.2006	R. Machida	
Zygentoma	Lepisma saccharina	VIII. Bezirk, Vienna, Austria	24.10.2004	W. Moser	28S
	Lepisma saccharina	Burscheid, Nord-Rhein-Westfalia, Germany	01.11.2005	J. Dambach	18S
	Ctenolepisma longicaudata	Espririto Santo, Brazil			
Odonata	Brachytron pratense	France			
	Aeshna juncea	France			
	Oxygastra curtisi	France			
	Cordulia aenea	Japan			
	Somatochlora flavomaculata	France			
	Epiophlebia superstes	Japan			
	Progomphus obscurus	USA			
	Sympetrum danae	France			
	Lestes viridis	Germany			
Ephemeroptera	Epeorus sylvicola	Natural History Museum Prague, Czechia	June 2005		
	Siphlonura aestivalis	Natural History Museum Prague, Czechia	August 2005		
Blatteria	Ectobius lapponicus	Hannover Niedersachsen Germany	June 2006	A. Melber	
Auchenorrhyncha	Cercopis vulnerata	Hannover Niedersachsen Germany	June 2006	A. Melber	
Coleoptera	Silpha obscura	Hannover Niedersachsen Germany	June 2006	A. Melber	
Heteroptera	Pyrrhocoris apterus	Hannover, Niedersachsen, Germany	June 2006	A. Melber	
	Rhaphigaster nebulosa	Hannover, Niedersachsen, Germany	November 2006	A. Melber	
	Harocera thoracica	Hannover, Niedersachsen, Germany	April 2006	A. Melber	
Hymenoptera	Nomada sp.	Hannover, Niedersachsen, Germany	April 2006	S. Simon	
	Scolia sp	Tunisia	April 2006	S. Sagasser	
	Tenthredinidae sp.	Hannover, Niedersachsen, Germany	June 2006	A. Melber	
Lepidoptera	Pieris napi	Hannover, Niedersachsen, Germany	July 2006	A. Melber	
Mantophasmatodea	Mantophasma zephyra	breed, South Africa	2005	R. Predel	
	Tyrannophasma gladiator	breed, South Africa	2006	R. Predel	
Dermaptera	Forficula auricularia	Hannover, Niedersachsen, Germany	July 2006	A. Melber	
Mantodea	Hierodula membranacea	breed, Germany	2006		
Phasmatodea	Carausius morosus	breed, India	2004	A. Melber	
	Bacillus rossius	Tunisia	April 2006	S. Sagasser	
Mecoptera	Boreus hyemalis	Soltau, Niedersachsen, Germany	November 2005	A. Melber	
Orthoptera	Anacridium aegypticum	Tunisia	April 2006	S. Sagasser	
	Leptophyes punctatissima	Hannover, Niedersachsen, Germany	June 2006	A. Melber	
	Pholidoptera griseoaptera	Hannover, Niedersachsen, Germany	July 2006	A. Melber	
Plecoptera	Isoperla sp.	Natural History Museum Prague, Czechia	July 2005		
	Nemoura flexuosa	Natural History Museum Prague, Czechia	August 2005		
Siphonaptera	Ctenocephalides felis	breed, Germany	2006	C. Epe	
Trichoptera	Triaenodes sp.	Hannover, Niedersachsen, Germany	September 2006	S. Simon	

Table A.8.: Full taxa list of sampled sequences. * indicates concatenated 18S and 28S rRNA sequences from different species, see Tab. A.10. ** contributed sequences in the present study (author of sequences).

Order	Taxon	Accession number 28S rRNA	length 28S rRNA (bp)	Accession number 18S rRNA	length 18S rRNA (bp)
Arachnida	Amblyomma americanum	AF291874	4005	AF291874	1815
	Dermacentor sp. *	AY859582	3920	L76340	1784
	Chalocheiridius cf. termitophilus	AY859558	3394	AY859559	1773
	Pandinus imperator	AY210830	3777	AY210831	1762
	Siro rubens	AY859602	3762	U36998	1809
	Eremobates sp.	AY859572	3833	AY859573	1767
	Aphonopelma hentzi *	AY210803	3819	DQ639776	1750
	Misumenops asperatus	AY210461	3467	AY210445	1786
	Mastigoproctus giganteus	AY859587	3796	AF005446	1790
	Paraphrynus sp.	AY859594	3785	AF005445	1777
Xiphosura	Limulus polyphemus	AF212167	3772	L81949	1807
Pycnogonida	Callipallene sp.	AY210807	3900	AF005439	1817
	Colossendeis sp.	EU420133 ** (v. Reumont)	3864	EU420135 ** (v. Reumont)	1798
	Anoplodactylus portus	AY859550	3893	AY859551	1809
	Nymphon stroemii	EU420134 ** (v. Reumont)	3818	EU420136 ** (v. Reumont)	1825
Anostraca	Artemia sp. *	AY210805	3628	AJ238061	1809
Notostraca	Triops cancriformis	EU370435 ** (v. Reumont)	3420	EU370422 ** (v. Reumont)	1784
	Triops longicaudatus	AY157606	3458	AF144219	1809
Diplostraca	Daphnia cf. magna	EU370436 ** (v. Reumont)	3823	EU370423 ** (v. Reumont)	2291
	Bosmina sp. *	EU370437 ** (v. Reumont)	3332	Z22731	1875
	Eulimnadia texana	AY859574	3665	AF144211	1813
Ostracoda	Heterocypris incongruens	EU370438 ** (v. Reumont)	3279	EU370424 ** (v. Reumont)	1786
	Pontocypris mytiloides	EU370439 ** (v. Reumont)	3672	EU370425 ** (v. Reumont)	1897
Cirripedia	Semibalanus balanoides	EU370440 ** (v. Reumont)	3274	EU370426 ** (v. Reumont)	1847
	Megabalanus californicus	AY859588	3720	AY520632	1812
	Pollicipes pollicipes	EU370441 ** (v. Reumont)	3549	EU370427 ** (v. Reumont)	1852
Branchiura	Argulus foliaceus	EU370442 ** (v. Reumont)	3512	EU370428 ** (v. Reumont)	1851
Mystacocarida	Derocheilocaris typicus	EU370443 ** (v. Reumont)	3663	EU370429 ** (v. Reumont)	2171
Copepoda	Cyclopidae sp. *	AY210813	3536	AJ746334	1808
	Chondracanthus lophii	DQ180341	3465	L34046	1810
	Tigriopus cf. fulvus	EU370444 ** (v. Reumont)	3532	EU370430 ** (v. Reumont)	1792
	Canuella perplexa	EU370445 ** (v. Reumont)	3462	EU370432 ** (v. Reumont)	1573
	Lepeophtheirus salmonis	DQ180342	3692	AF208263	1799
Remipedia	Speleonectes tulumensis	EU370446 ** (v. Reumont)	3797	EU370431 ** (v. Reumont) / L81936	1302 / 1965
Cephalocarida	Hutchinsoniella macracantha	EF189645	2480	L81935	2018
Leptostraca	Nebalia sp.	EU370447 ** (v. Reumont)	3519	EU370433 ** (v. Reumont)	1789
Anaspidacea	Anaspides tasmaniae	AY859549	3997	L81948	1827
Mysidacea	Heteromysis sp.	AY859578	3400	AY743946	1724
Decapoda	Homarus americanus	AY859581	4351	AY743945	1758
	Penaeus vannamei *	AF124597	5820	DQ079766	1781
Stomatopoda	Squilla empusa	AY210842	3913	L81946	1817
Pentastomida	Raillietiella sp. *	EU370448 ** (v. Reumont) / AY744894	1286 / 1983	EU370434 ** (v. Reumont)	1814
Chilopoda	Craterostigmus tasmanianus	EU376009 ** (Bartel)	4024	EU368617 ** (Meusemann)	1786
	Otostigmus politus	DQ666180	4170	DQ666177	1868
	Scolopendra mutilans	DQ666181	4174	DQ666178	1848
	Scutigera coleoptrata	AY859601	4024	AF000772	1865
	Lithobius forficatus	EF199984	3913	EU368618 ** (Meusemann)	1752
Diplopoda	Polyxenus lagurus	EU376011 ** (Bartel)	3967	EU368619 ** (Meusemann)	1733
	Monographis sp.	EF192437 ** (Bartel / Luan)	3866	AY596371	1744
	Paradoxosomatidae sp.	DQ666182	4288	DQ666179	1797
	Polydesmus complanatus	EU376010 ** (Bartel)	4271	EU368620 ** (Meusemann)	1689
	Cherokia georgiana	AY859562	4225	AY859563	1781
	Orthoporus sp.	AY210828	4124	AY210829	1791
	Cylindroiulus caeruleocinctus	EF199985	4084	EU368621 ** (Meusemann)	1753
Pauropoda	Allopauropus sp.	DQ666185	4406	DQ399857	2227
	Pauropodidae sp.	EU376012 ** (Bartel)	4238	EU368622 ** (Meusemann)	2250
Symphyla	Scutigerella sp.	DQ666184	4471	DQ399856	1902
	Hanseniella sp.	AY210821-22	4539	AY210823	1925
	Symphylella sp.	DQ666183	4558	DQ399855	2057
Protura	Acerentomon franzi	EF199976	4099	EU368597 ** (Meusemann)	1790
	Baculentulus densus *	EU376049	4100	AY037169	1984
	Eosentomon sp.	EU376047 ** (Dell'Ampio)	3654	EU368598 ** (Meusemann)	1860
	Eosentomon sakura	EF192434 ** (Dell'Ampio / Luan)	3789	AY596355	1948
	Sinentomon erythranum	EF192442 ** (Dell'Ampio / Luan)	4043	AY596358	1934
Diplura	Campodeidae sp.	AY859560	3718	AY859561	1866
	Campodea augens	EF199977	4010	EU368599 ** (Meusemann)	1788
	Lepidocampa weberi	EU376050	4061	AY037167	1878
	Catajapyx aquilonaris	EF199978	5016	EU368600 ** (Meusemann)	2154
	Parajapyx emeryanus	EF192440 ** (Dell'Ampio / Luan)	4143	AY037168	2120
	Octostigma sinensis	EF192439 ** (Dell'Ampio / Luan)	4001	AY145134	2138
Collembola	Tetrodontophora bielanensis	EU376051	3868	AY555519	1760
	Gomphiocephalus hodgsoni	EF199969	3893	EU368601 ** (Meusemann)	1746
	Triacanthella sp.	AY859609	3823	AY859610	1758
	Bilobella aurantiaca	AJ251729	3934	EU368602 ** (Meusemann)	1759
	Anurida maritima	AJ251738	3965	EU368603 ** (Meusemann)	1680
	Podura aquatica	EF199970	3899	EU368604 ** (Meusemann)	1696
	Cryptopygus antarcticus	EF199971	3862	EU368605 ** (Meusemann)	1724
	Isotoma viridis	EU376052	3866	AY596361	1748
	Orchesella villosa	EF199972	3867	EU368606 ** (Meusemann)	1739
	Pogonognathellus flavescens	EU376053	3874	EU368607 ** (Meusemann)	1688
	Megalothorax minimus	EF199975	3868	EU368608 ** (Meusemann)	1703
	Sminthurus viridis	EF199973	3912	EU368609 ** (Meusemann)	1695
	Allacma fusca	EU376054	3877	EU368610 ** (Meusemann)	1759
	Dicyrtomina saundersi	EF199974	3871	EU368611 ** (Meusemann)	1739
Archaeognatha	Machilis hrabei	EF199981	3750	EU368612 ** (Meusemann)	1703
	Lepismachilis y-signata	EF199980	3826	EU368613 ** (Meusemann)	1679

* indicates concatenated 18S and 28S rRNA sequences from different species; ** contributed sequences in the present study (author of sequences).

Table A.8 continued

Order	Taxon	Accession number 28S rRNA	length 28S rRNA (bp)	Accession number 18S rRNA	length 18S rRNA (bp)
	Pedetontus okajimae	EU376055	3800	EU368614 ** (Meusemann)	1742
Zygentoma	Lepisma saccharina	EU376048 ** (Dell'Ampio)	3506	EU368615 ** (Meusemann)	1703
	Ctenolepisma longicaudata	AY210810	3907	EU368616 ** (Meusemann)	1744
Odonata	Brachytron pratense	EU424323 ** (Letsch)	3738	AF461232	1737
	Aeshna juncea	EU424324 ** (Letsch)	3736	AF461231	1767
	Oxygastra curtisi	EU424325 ** (Letsch)	3736	DQ008194	1787
	Cordulia aenea	EU424326 ** (Letsch)	3795	AF461236	1768
	Somatochlora flavomaculata	EU424327 ** (Letsch)	3795	AF461242	1757
	Epiophlebia superstes	EU424328 ** (Letsch)	3736	AF461247	1835
	Progomphus obscurus	EU424329 ** (Letsch)	3756	AY749909	1843
	Sympetrum danae	EU424330 ** (Letsch)	3756	AF461243	1754
	Leucorrhinia sp.	AY859583	4114	AY859584	1815
	Lestes viridis	EU424331 ** (Letsch)	3747	AJ421949	1867
Ephemeroptera	Callibaetis ferrugineus	AY859557	3887	AF370791	1812
	Epeorus sylvicola *	EU414715 ** (Simon)	3680	AY749837	1808
	Siphlonura aestivalis *	EU414716 ** (Simon)	4151	DQ008181	1784
Phasmatodea	Carausius morosus	EU426878 ** (Simon)	3737	X89488	1899
	Bacillus rossius	EU426879 ** (Simon)	3889	AY121180	1891
Mantophasmatodea	Mantophasma zephyra *	EU414719 ** (Simon)	3383	DQ874153	2018
	Tyrannophasma gladiator	EU426875 ** (Simon)	3878	AY521863	2074
Mantodea	Mantis religiosa	AY859585	3990	AY491153	1734
	Hierodula membranacea *	EU414720 ** (Simon)	3603	AY491194	1734
Blatteria	Gromphadorhina laevigata	AY210819	4015	AY210820	1877
	Ectobius lapponicus	EU426877 ** (Simon)	4006	DQ874125	1808
	Blattella germanica	AF005243	3931	AF005243	1964
Isoptera	Zootermopsis angusticollis	AY859614	4183	AY859615	1873
Dermaptera	Forficula auricularia	EU426876 ** (Simon)	4016	Z97594	1873
Plecoptera	Isoperla sp *	EU414717 ** (Simon)	4299	AF461256	2054
	Nemoura flexuosa *	EU414718 ** (Simon)	3256	AF461257	1763
Heteroptera	Pyrrhocoris apterus *	EU414725 ** (Simon)	3389	AY627318	1829
	Rhaphigaster nebulosa	EU426880 ** (Simon)	3983	X89495	1924
	Harocera thoracica *	EU414726 ** (Simon)	3405	AY252388	1895
Auchenorrhyncha	Cercopis vulnerata *	EU414724 ** (Simon)	3615	AY744798	1856
Hemiptera	Clastoptera obtusa	AF304569	3201	AY744784	1859
	Pectinariophyes reticulata	AF304570	3259	AY744778	1848
Orthoptera	Gomphocerinae sp	AY859546	4187	AY859547	1864
	Anacridium aegypticum *	EU414723 ** (Simon)	3819	AY379759	1833
	Acheta domesticus	AY859544	4092	X95741	1802
	Leptophyes punctatissima *	EU414721 ** (Simon)	3918	AY521867	1897
	Pholidoptera griseoaptera *	EU414722 ** (Simon)	3950	Z97587	1884
Coleoptera	Tenebrio sp *	AY210843	4459	X07801	2083
	Silpha obscura	EU426881 ** (Simon)	2783	AJ810737	1930
Hymenoptera	Myrmecia croslandi	AB052895	3460	AB121786	1766
	Vespula pensylvanica	AY859612	3912	AY859613	1871
	Nomada sp. *	EU414727 ** (Simon)	3386	AY703484	1854
	Scolia sp. *	EU414728 ** (Simon)	3405	EF012932	1851
	Tenthredinidae sp. *	EU414729 ** (Simon)	3472	AF423781	1836
Siphonaptera	Ctenocephalides felis *	EU414732 ** (Simon)	3333	AF423914	1878
Mecoptera	Merope tuber	DQ202351	3736	AF286287	1886
	Boreus hyemalis	EU426882 ** (Simon)	3534	AF423882	1881
Lepidoptera	Pieris napi *	EU414731 ** (Simon)	3743	AF423785	1856
Trichoptera	Oxyethira rossi *	DQ202352	3869	AF423801	1848
	Triaenodes sp. *	EU414730 ** (Simon)	3095	AF286300	1897
Diptera	Acricotopus lucens	AJ586562	3910	AJ586561	1939
	Chironomus tentans	X99212	3973	X99212	1528
	Anopheles albimanus	L78065	4022	L78065	1977
	Aedes albopictus	L22060	4102	X57172	1950
	Drosophila melanogaster	M21017	3900	M21017	1995
	Simulium sanctipauli	AF403805	3733	AF403800	1912
Onyphora	Peripatus sp.	AY210836	3297	AY210837	2476
	Peripatoides novaezealandiae	AF342793	4570	AF342794	2064
Tardigrada	Milnesium sp. *	AY210826	3579	U49909	1844

* indicates concatenated 18S and 28S rRNA sequences from different species; ** contributed sequences in the present study (author of sequences).

Table A.9.: PCR temperature-profiles and conditions of amplyfying gene fragments for basal hexapods and myriapods (18S, 28S). AH: Apterygote hexapods; My: Myriapods; Pau: Pauropodidae sp.; Cr: Crustaceans. °C: temperature in Celsius; X:00: time in minutes; TD: touch down; min: minutes.

Profile	Taxa	Temperature profile	Number of cycles	Gene	Thermocycler	Remarks / Primer specification
1	AH, My, Cr	94°C 3:00 min		18S, 28S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
		94°C 0:35 min	15 cycles			
		60°C 0:30 min, TD -1°C to 45°C				
		72°C 1:30 min				
		94°C 0:35 min	25 cycles			
		50°C 0:30 min				
		72°C 1:30 min				
		72°C 10:00 min				
		4°C				
2	AH, My	95°C 15:00 min		18S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	
		94°C 0:35 min	15 cycles			
		60°C 1:30 min, TD -1°C to 45°C				
		72°C 1:30 min				
		94°C 0:35 min	25 cycles			
		50°C 1:30 min				
		72°C 1:30 min				
		72°C 10:00 min				
		4°C				
3	AH, My	95°C 5:00 min		28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	
		95°C 1:00 min	30 cycles			
		45°C 1:00 min				
		72° 1:00 min				
		72°C 10:00 min				
		4°C				
4	AH, My	95°C 5:00 min		28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	
		95°C 0:30 min	20 cycles			
		45°C 0:30 min				
		72°C 0:45 min				
		95°C 1:00 min	10 cycles			
		56°C 1:00 min				
		72°C 1:00 min				
		72°C 10:00 min				
		4°C				
5	AH, My	95°C 5:00 min		28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	
		95°C 1:00 min	10 cycles			
		56°C 1:00 min				
		72°C 1:00 min				
		95°C 0:30 min	15 cycles			
		56°C 0:30 min				
		72°C 0:45 min				
		72°C 10:00 min				
		4°C				
6	Pau	94°C 4:00 min		28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	M13Rv/M13Fw PCR after picking clones
		94°C 1:00 min	30 cycles			
		55°C 1:00 min				
		72°C 3:15 min				
		72°C 10:00 min				
		4°C				

Table A.10.: List of chimeran species for concatenated 18S and 28S rRNA sequences.

taxon listed as	**28S rRNA**	**18S rRNA**
Dermacentor sp. *	*Dermacentor* sp.	*Dermacentor andersoni*
Aphonopelma hentzi *	*Aphonopelma hentzi*	*Aphonopelma reversum*
Artemia sp. *	*Artemia* sp.	*Artemia franciscana*
Bosmina sp. *	*Bosmina* sp.	*Bosmina longirostris*
Cyclopidae sp. *	Cyclopidae sp.	*Macrocyclops albidus*
Penaeus vannamei *	*Penaeus vannamei*	*Penaeus semisulsatus*
Raillietiella sp. *	*Raillietiella* sp.	*Raillietiella* sp.
Baculentulus densus *	*Baculentulus densus*	*Baculentulus tianmushanensis*
Epeorus sylvicola *	*Epeorus sylvicola*	*Epeorus longimanus*
*Siphlonura aestivalis**	*Siphlonura aestivalis*	*Siphlonura croaticus*
Mantophasma zephyra *	*Mantophasma zephyra*	*Mantophasma* cf. *zephyra*
Hierodula membranacea *	*Hierodula membranacea*	*Hierodula schultzei*
Isoperla sp. *	*Isoperla* sp.	*Isoperla obscura*
Nemoura flexuosa *	*Nemoura flexuosa*	*Nemoura cinerea*
Pyrrhocoris apterus *	*Pyrrhocoris apterus*	*Dysdercus poecilus*
Harocera thoracica *	*Harocera thoracica*	*Polymerus castilleja*
Cercopis vulnerata *	*Cercopis vulnerata*	*Mahanarva costaricensis*
Anacridium aegypticum *	*Anacridium aegypticum*	*Acrida cinerea*
Leptophyes punctatissima *	*Leptophyes punctatissima*	*Microcentrum rhombifolium*
Pholidoptera griseoaptera *	*Pholidoptera griseoaptera*	*Tettigonia viridissima*
Tenebrio sp. *	*Tenebrio* sp.	*Tenebrio molitor*
Nomada sp. *	*Nomada* sp.	*Apis mellifera*
Scolia sp. *	*Scolia* sp.	*Scolia verticalis*
Tenthredinidae sp. *	Tenthredinidae sp.	*Dolerus* sp.
Pieris napi *	*Pieris napi*	*Anthocharis sara*
Ctenocephalides felis *	*Ctenocephalides felis*	*Ctenocephalides canis*
Oxyethira rossi *	*Oxyethira rossi*	*Oxyethira dualis*
Triaenodes sp. *	*Triaenodes* sp.	*Ctenocephalides canis*
Milnesium sp. *	*Milnesium* sp.	*Milnesium tardigradum*

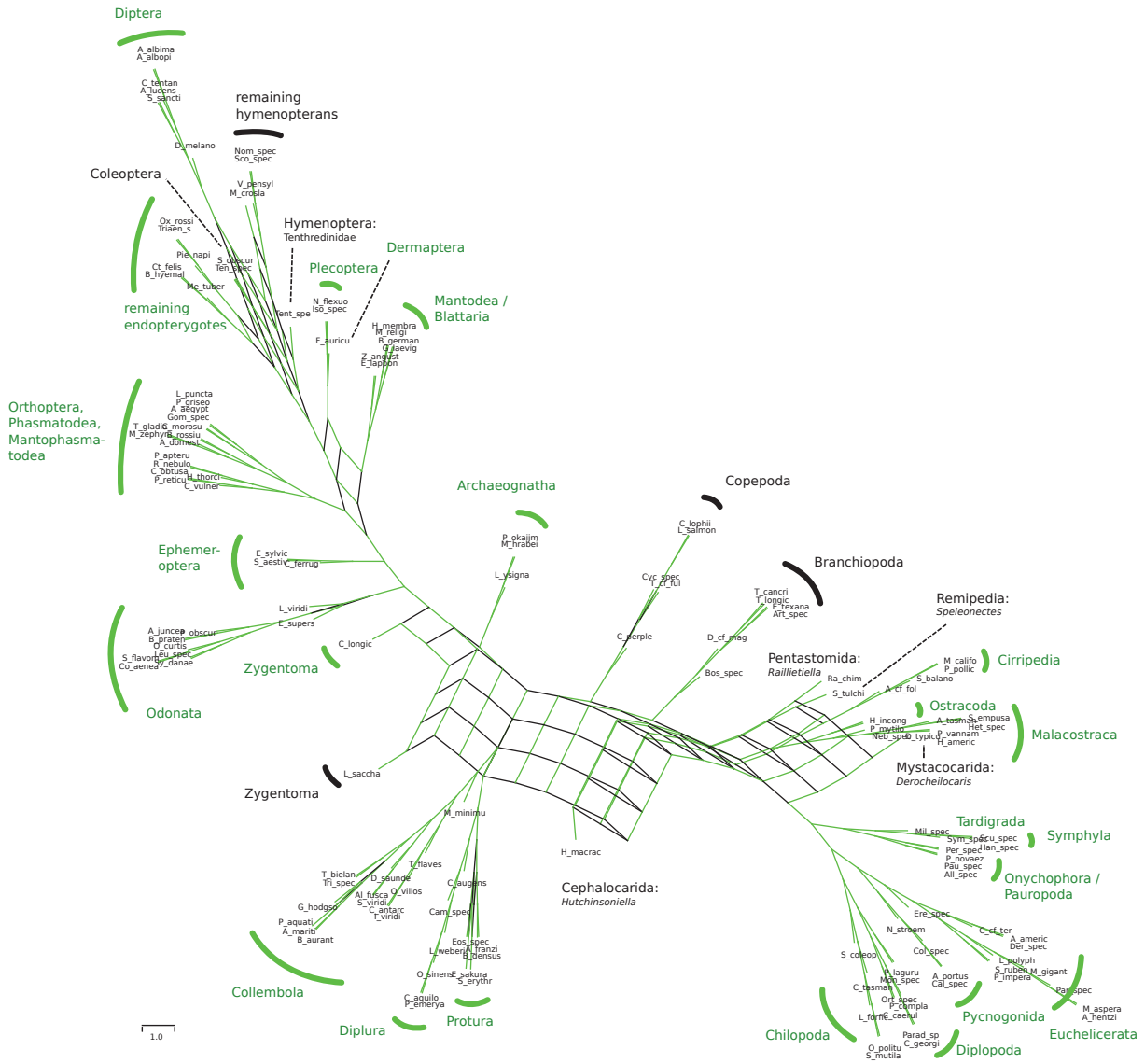

Figure A.11.: Consensus network inferred from both stationary topologies (DNA and RNA/DNA tree). The consensus network was calculated with SplitsTree 4.10 (Huson and Bryant, 2006) (conflict threshold: 0.01). Black lines indicate contradictory relationships that are only present in the DNA tree. Taxa printed in black indicate a different placement compared with the RNA/DNA tree. Green lines indicate a similar topology in the DNA and RNA/DNA tree.

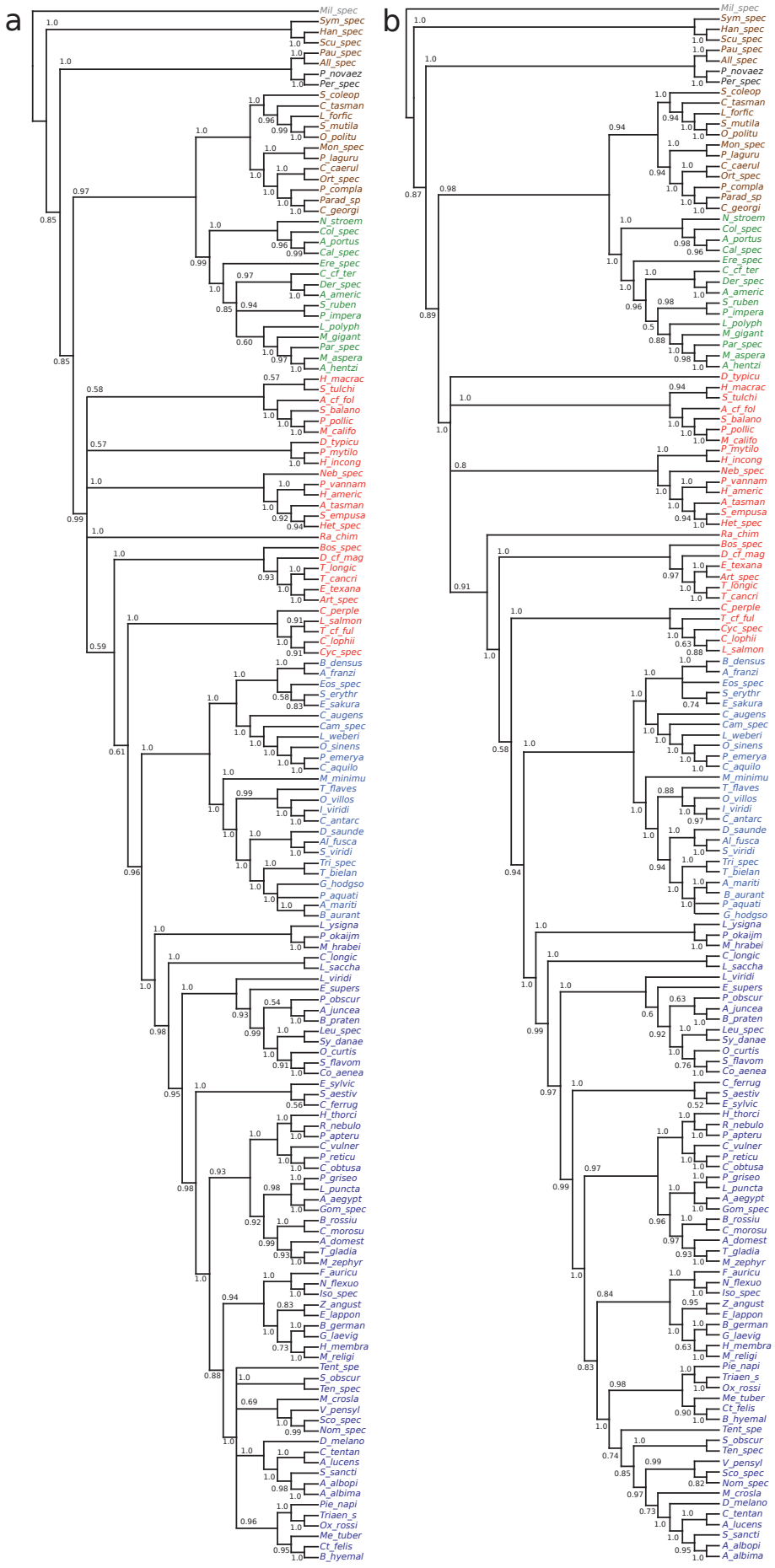

Figure A.12.: Bayesian majority rule consensus trees inferred with *PHASE-2.0* based on the DNA model approach. Color code is specified in Fig. 3.3. **a** DNA model non-stationary (time-heterogeneous) approach; mrc tree deduced from 15,000 sampled trees with posterior probability support values. **b** DNA model stationary (time-homogeneous) approach; nrc tree deduced from 5,000 sampled trees with posterior probability support values.