

Cumulative dissertation submitted in partial fulfillment of  
the requirements for the doctoral degree  
(Dr. rer. nat.)

**Protein-Coding Gene Repertoires —  
Annotation, Characterization, and  
Variability in Holometabola**

by  
**Jeanne Wilbrandt**  
from  
Berlin

Bonn, June 2018

Faculty of Mathematics and Natural Sciences,  
Rheinische Friedrich-Wilhelms-Universität, Bonn



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

Angefertigt am Zoologischen Forschungsmuseum Alexander Koenig, Bonn.



Finanziert durch die DFG Projekte MI 649/16-1 und NI 1387/3-1.

Erstgutachter:	Prof. Dr. Bernhard Misof
Zweitgutachter:	Prof. Dr. Oliver Niehuis
Fachnahes Kommissionsmitglied:	Prof. Dr. Dietmar Quandt
Fachfremdes Kommissionsmitglied:	Prof. Dr. Gabriele König

Tag der Promotion: 11. Oktober 2018

Erscheinungsjahr: 2018



---

## Summary

---

**T**HE RAPIDLY FILLING TREASURE TROVE of sequenced genomes from diverse organisms allowed comparative genomics to shift from analyzing individual genes towards investigating entire genomes and their components. The comparison of gene and genome structures across species has the potential to reveal major trends of genome evolution. However, this data deluge also poses unprecedented problems, mostly related to computational requirements like handling, quality assurance, and efficient analysis (MUIR *et al.*, 2016). Standardization is lagging behind, especially concerning the tools and terminology to describe not one but all genes of a species. Consequently, previously unattainable questions, such as “Which evolutionary mechanisms and processes shape the composition of protein-coding gene repertoires?”, are only now open to in-depth exploration.

In this work, I focus on the protein-coding gene repertoires of insects. Insects, and especially Holometabola (species that undergo pupal development and metamorphosis), are tremendously species-rich and diverse considering their phylogenetic age. To understand the basis of their evolutionary success, it is necessary to investigate the dynamics and evolvability of their genomes. My thesis thus revolves around the extent of gene repertoire changes to eventually address the central questions: are major evolutionary transitions within insects (e.g., radiations) correlated to significant changes of the repertoire? How large is the variation of gene structures within a repertoire? Are there classes of genes that are prone to change? Large-scale analyses of gene repertoires to tackle these

themes are not trivial in that they require the application of clear definitions and suitable tools, two prerequisites that were previously not fully appreciated. Thus, to approach my focal questions, it was necessary to develop new methods of data acquisition and evaluation.

**T**HIS THESIS IS STRUCTURED in six parts as follows. Firstly, I provide a primer for the reader, i.e., background information on genomes and protein-coding repertoires (I). Having thus the field of comparative genomics introduced, a few central terms are explained to prepare the ground for the following review of the state of the art (including the deficits in relation to my research that existed until now). The specific research questions addressed in my thesis are outlined (section I.3.1).

**I**N THE FIRST RESULTS PART (II), I discuss the problem of lacking tools for repertoire-wide gene structure analysis and present such a tool. This lack and an observed inconsistency in the use of definitions called for a standard at the descriptive level with an undeviating use of standardized terms, and for software that infers the required data to structurally describe and analyze protein-coding gene repertoires (e.g., lengths and counts of exons) under these strict definitions. I developed a new open-source command-line tool: COGNATE, the Comparative Gene Annotation Characterizer. As part of this project, the standard definitions of gene and genome structures were revised and provided as a working draft suggestion for further reference. The developed tool has recently been published in *BMC Genomics*.

**I**N THE SECOND RESULTS PART (III), I employ the tool COGNATE to address the question whether or not automatically generated gene predictions are suitable to analyze general gene structure parameters. Attending to this question was necessary, because annotation algorithms are not free of errors and it has been argued that only human review of the predictions ensures sufficient reliability. If this was true without exception, explorations of gene repertoires would be stalled until all predicted genes had been manually curated. Thus, after outlining the prospects and limits of both automated and manual annotation, I analyzed the effect of manual curation on predicted structural properties of protein-coding genes by comparing annotated gene sets

from seven insect species sequenced by the i5k initiative. The properties of automatically generated gene models and their manually curated replacements do not differ extensively, and major correlative trends regarding gene structures can be recovered from both sets. From these results I conclude that gene models yielded from unsupervised annotation procedures are a suitable data basis to characterize structural gene features of a whole repertoire. This manuscript has been submitted for publication to *BMC Genomics* and is currently under review.

**H**AVING BOTH suitable tool and data at hand, the third results part (IV) turns to the exemplary application of COGNATE and down-stream analysis methods as means to structurally characterize gene repertoires of Hymenoptera (sawflies, bees, wasps, and ants). This work is part of a large-scale project that examines two newly sequenced genomes of non-apocritan “symphytans” with regard to the major transition from phytophagous to parasitoid hymenopteran life styles from various angles. One of these angles are protein-coding gene structure variations within and between the repertoires. This analysis is a necessary first step in comparative genomics to range in measured magnitudes of parameters and to build expectations on hymenopteran protein-coding gene structures. The two focal species possess small genomes and gene repertoires compared to the range of hymenopteran genome and gene repertoire sizes, but a strikingly high GC content of more than 41 % is found in both species; this is in stark contrast to the honeybee. Although gene counts vary considerably among the analyzed Hymenoptera (twelve species in total), their genomes harbor a very similar amount of protein-coding sequences. Gene length and composition complexity (reflected by the number of exons) appear to be slightly decreased in derived wasp-waisted Hymenoptera, with the exception of the bumblebee. The findings are discussed and provide a basis for a more comprehensive comparison in the following part. My results will be included in a publication in *Current Biology* following this year.

**T**HE ANALYSIS APPROACH established by me is used in the last results part (V) to address my original question of variability in protein-coding gene repertoires in a characterization of the repertoires of a larger, unique species sample. The sample covers a wide range of divergence times within

Hymenoptera, including the two previously analyzed “symphytans” and 16 other species, and allows the comparison to seven other insect and one millipede outgroup species. Previous research suggested universal patterns of conservation within a gene repertoire in relation to others by which it could be partitioned: the highly conserved core gene set, the moderately conserved shell genes with a patchy distribution across taxa, and the cloud that contains genes shared by very few directly related or no other species. However, the characteristics regarding structure and function of the genes classified within partitions have previously not been examined, although this will help to answer questions of how (a balance of) conservation (in the core) and variation (in the cloud, the source of novelty) in gene repertoires is established and maintained. Further research is required, but my results show that considerable differences exist in gene structural parameters between conserved core-genes and the lineage-specific cloud gene set.

**T**HE LAST PART (VI) comprises a general discussion of my findings in context of the state of the art, a general conclusion, and an outlook.



---

# Contents

---

<b>Summary</b>	<b>i</b>
<b>Contents</b>	<b>xi</b>
<b>I GENERAL INTRODUCTION</b>	<b>1</b>
<b>1 Background – What is comparative genomics?</b>	<b>3</b>
1.1 Genome deciphering and interpretation . . . . .	4
1.2 Comparisons in biology . . . . .	7
1.3 A small history of comparative genomics . . . . .	11
<b>2 State of the art</b>	<b>15</b>
2.1 Genome components and genome size variation . . . . .	15
2.2 Protein-coding gene repertoire evolution and dynamics . . . . .	18
2.2.1 Gene structure dynamics and correlates . . . . .	19
2.2.2 Gene family dynamics . . . . .	21
2.2.3 Gene and gene family turnover . . . . .	23
<b>3 Thesis focus and aim</b>	<b>27</b>
3.1 Research questions . . . . .	28
3.1.1 Hymenopteran gene repertoire structures? . . . . .	28
3.1.2 Adequacy of automatically annotated gene models? . . . . .	29
3.1.3 A tool to characterize gene structure? . . . . .	29
3.2 Research map . . . . .	30

<b>Bibliography I</b>	<b>31</b>
<b>II THE TOOL COGNATE</b>	<b>49</b>
<b>1 Abstract</b>	<b>51</b>
<b>2 Introduction</b>	<b>53</b>
2.1 A lack of standards in gene structure characterization . . . . .	53
2.2 Why another tool? . . . . .	55
<b>3 Methods and implementation</b>	<b>57</b>
3.1 Input and running . . . . .	57
3.2 Output . . . . .	60
<b>4 Results and discussion</b>	<b>63</b>
4.1 Applicability of COGNATE . . . . .	63
4.2 Standardization problems . . . . .	64
4.3 Standardization suggestions . . . . .	69
<b>5 Conclusion</b>	<b>71</b>
<b>6 Additional publication information</b>	<b>73</b>
6.1 Availability and requirements . . . . .	73
6.2 Acknowledgements . . . . .	74
6.3 Funding . . . . .	74
6.4 Authors' contributions . . . . .	74
<b>Bibliography II</b>	<b>75</b>
<b>III AUTOMATICALLY GENERATED VS MANUALLY CURATED MODELS</b>	<b>79</b>
<b>1 Abstract</b>	<b>81</b>
<b>2 Introduction</b>	<b>83</b>
2.1 Prevalence of automated annotation and its problems . . . . .	84
2.2 Manual curation as corrective . . . . .	85
2.3 Implication of 'gold standard' manual curation . . . . .	86

2.4	Benchmarking manual versus automated annotation . . . . .	86
2.5	Automatic vs manual . . . . .	87
<b>3</b>	<b>Results</b>	<b>89</b>
3.1	Curator experience varies . . . . .	89
3.2	Curator's dilemmas . . . . .	91
3.3	Gene models selected for manual curation are representative . . .	92
3.4	The effect of manual curation . . . . .	94
3.5	Comparison to another automated annotation procedure . . . . .	98
3.6	Correlative trends of gene counts and coverages . . . . .	99
3.7	Correlative trends of gene composition . . . . .	102
<b>4</b>	<b>Discussion</b>	<b>103</b>
4.1	Reflections on manual curation . . . . .	104
4.2	The influence of manual curation on gene structure . . . . .	105
4.3	The influence of tool choice . . . . .	106
4.4	Elucidating trends using automatically predicted gene models . .	108
4.5	Conclusion . . . . .	109
4.6	Future directions . . . . .	109
<b>5</b>	<b>Material and Methods</b>	<b>111</b>
5.1	Data sample . . . . .	111
5.2	Set preparation . . . . .	111
5.3	Non-canonical start positions in MAKER predictions . . . . .	112
5.4	Curator experience analysis . . . . .	112
5.5	Structural parameter and correlative trend analysis . . . . .	112
5.6	BRAKER-vs-MAKER analysis . . . . .	113
<b>6</b>	<b>Additional publication information</b>	<b>115</b>
6.1	Availability of data and materials . . . . .	115
6.2	Funding . . . . .	116
6.3	Authors' contributions . . . . .	116
6.4	Acknowledgements . . . . .	116
	<b>Bibliography III</b>	<b>117</b>

<b>IV</b>	<b>HYMENOPTERAN REPERTOIRE FEATURES</b>	<b>123</b>
<b>1</b>	<b>Introduction</b>	<b>125</b>
<b>2</b>	<b>Methods</b>	<b>129</b>
2.1	Species . . . . .	129
2.2	Analysis . . . . .	130
<b>3</b>	<b>Results</b>	<b>133</b>
3.1	Assembly features . . . . .	133
3.2	Component sizes . . . . .	134
3.3	Structural features . . . . .	134
3.3.1	Transcripts . . . . .	134
3.3.2	Exons and CDSs . . . . .	137
3.3.3	Introns . . . . .	139
3.4	Distributions and correlations of GC content . . . . .	139
<b>4</b>	<b>Discussion</b>	<b>147</b>
4.1	Component sizes . . . . .	147
4.2	Tentative inference of gene structure evolution trends . . . . .	148
<b>5</b>	<b>Conclusion</b>	<b>151</b>
	<b>Bibliography IV</b>	<b>153</b>
<b>V</b>	<b>CONSERVATION CLASSES OF THE REPERTOIRES</b>	<b>155</b>
<b>1</b>	<b>Summary</b>	<b>157</b>
<b>2</b>	<b>Introduction</b>	<b>159</b>
<b>3</b>	<b>Methods</b>	<b>163</b>
3.1	Species sample . . . . .	163
3.2	Sequencing of genomes and transcriptomes . . . . .	165
3.3	Assembly and repeat masking . . . . .	167
3.4	MySQL database setup . . . . .	168
3.5	Protein-coding gene annotation and structural characterization . . . . .	168
3.6	Assessment of gene space coverage . . . . .	170

3.7	Protein domain annotation and analysis . . . . .	170
3.8	Phylogenetic tree . . . . .	172
3.9	Orthology prediction and partitioning of repertoires . . . . .	172
3.10	Plotting . . . . .	174
<b>4</b>	<b>Results</b>	<b>175</b>
4.1	The complete repertoires . . . . .	176
4.1.1	Gene counts and BUSCO assessment . . . . .	176
4.1.2	Overall gene structure parameter distributions . . . . .	176
4.2	Universal landscapes of ortholog groups . . . . .	176
4.3	Comparing core, shell, and cloud . . . . .	182
4.3.1	General gene structure medians . . . . .	182
4.3.2	Species-specific gene structure distributions . . . . .	183
4.3.3	Protein domain counts and arrangement diversity . . . . .	186
<b>5</b>	<b>Discussion</b>	<b>197</b>
5.1	Universal patterns of conservation . . . . .	198
5.2	Characteristics of conservation classes . . . . .	198
5.2.1	Gene structures . . . . .	199
5.2.2	Protein domains . . . . .	200
5.3	Future directions . . . . .	201
<b>6</b>	<b>Conclusion</b>	<b>203</b>
	<b>Bibliography V</b>	<b>205</b>
<b>VI</b>	<b>GENERAL DISCUSSION &amp; CONCLUSION</b>	<b>213</b>
<b>1</b>	<b>General discussion</b>	<b>215</b>
1.1	Prerequisites . . . . .	215
1.2	Gene repertoires of Hymenoptera . . . . .	216
1.2.1	Gene structure variation along the phylogenetic tree . . . . .	216
1.2.2	Gene structure variation within repertoire partitions . . . . .	218
1.3	Future prospects . . . . .	220
1.3.1	Methodological aspects . . . . .	220
1.3.2	Repertoire dynamics: the role of gene turnover . . . . .	220
1.3.3	A window to the past leading to the future . . . . .	222

<b>2</b>	<b>General conclusion</b>	<b>223</b>
	<b>Bibliography VI</b>	<b>225</b>
<b>VII</b>	<b>APPENDICES</b>	<b>231</b>
<b>A</b>	<b>Posters</b>	<b>233</b>
<b>B</b>	<b>Appendix to part II</b>	<b>239</b>
B.1	Supplementary tables . . . . .	239
B.1.1	Definitions . . . . .	239
B.2	Electronic supplements . . . . .	243
B.2.1	Additional file 1: Parameter table . . . . .	243
B.2.2	Additional file 2: Definition table . . . . .	243
B.2.3	Additional file 3: Result table . . . . .	244
B.2.4	Additional file 4: The COGNATE package . . . . .	244
	<b>Supplementary bibliography B</b>	<b>245</b>
<b>C</b>	<b>Appendix to part III</b>	<b>247</b>
C.1	Supplementary Notes . . . . .	247
C.1.1	Species Set . . . . .	247
C.1.2	Annotation . . . . .	248
C.1.3	Extended results . . . . .	251
C.1.4	Extended material and methods . . . . .	253
C.2	Electronic supplements . . . . .	255
C.2.1	Additional file 1: Data sources and used files . . . . .	255
C.2.2	Additional file 2: Curators . . . . .	256
C.2.3	Additional file 3: Non-canonical start codons . . . . .	256
C.2.4	Additional file 4: Automated vs manually curated gene models . . . . .	257
C.2.5	Additional file 5: BRAKER vs MAKER . . . . .	257
C.2.6	Additional file 6: BRAKER-only single-exon gene models . . . . .	257
C.2.7	Additional file 7: COGNATE results . . . . .	258
	<b>Supplementary bibliography C</b>	<b>259</b>

<b>D</b>	<b>Appendix to part IV</b>	<b>265</b>
D.1	Electronic supplements . . . . .	265
D.1.1	Additional file 1: Species and COGNATE median data . .	265
D.1.2	Additional file 2: COGNATE results . . . . .	266
D.1.3	Additional file 3: Density plots . . . . .	266
<b>E</b>	<b>Appendix to part V</b>	<b>267</b>
E.1	Supplementary Notes . . . . .	267
E.1.1	Assembly calls and library combinations . . . . .	267
E.2	Electronic supplements . . . . .	271
E.2.1	Additional file 1: MySQL database . . . . .	271
E.2.2	Additional file 2: MySQL commands . . . . .	271
E.2.3	Additional file 3: Perl and R scripts . . . . .	272
E.2.4	Additional file 4: COGNATE results . . . . .	272
	<b>Supplementary bibliography E</b>	<b>273</b>
	<b>List of Figures</b>	<b>275</b>
	<b>List of Tables</b>	<b>277</b>
	<b>List of Abbreviations</b>	<b>279</b>
	<b>Danksagung</b>	<b>281</b>
	<b>Erklärung</b>	<b>283</b>





---

# **GENERAL INTRODUCTION**

---



---

## Background – What is comparative genomics?

---

**H**OW DOES BIODIVERSITY EVOLVE? How and why do species and their genomes change during evolution? Which factors drive or restrain these changes at the genomic level? These are the fundamental questions guiding the quest to understand biological diversity. Research needs to address these questions focusing on individual aspects like genome components, *e.g.*, repertoires of non-coding, transposable, and protein-coding elements. In my thesis, I focus on the protein-coding gene repertoires, analyzed in a comparative genomics context. Here, I will shortly introduce the field of comparative genomics, its terminology and history. In the subsequent two chapters, the state of the art regarding several aspects of repertoire composition and dynamics is reviewed, and the questions of my thesis, based on these premises, are presented.

## 1.1 Genome deciphering and interpretation

The study of genomes was sparked by the discovery that chromosomes carry the hereditary information of organisms (independently by Boveri, Sutton, and DeVries in 1902 and 1903; WILSON, 1925). Further momentum was gained when the structure of deoxyribonucleic acid (DNA), the chromosomal molecules, was revealed (WATSON and CRICK, 1953). Since these early years, many terms have been coined to designate genomic units. Some of these terms have been moulded according to the advances of genetics and genomics. Unfortunately, some are also frequently misused. To avoid confusion, the most important terms shall be introduced here.

The designation ‘genome’ was introduced by WINKLER (1920, p. 165) and is nowadays – after some modulation (*e.g.*, MAHNER and KARY, 1997) – widely understood as all genetic material (DNA) of an organism. The term ‘gene’, invented by Wilhelm JOHANNSEN (1909) to designate units of heredity found within chromosomes<sup>I.1</sup>, has been similarly challenged and adapted (FALK, 1986; SNYDER and GERSTEIN, 2003). However, the definition of ‘gene’ is even today somewhat malleable, as are the definitions of the genic subunits ‘exon’ (“regions which will be expressed”) and ‘intron’ (“intragenic regions”); the latter two were introduced by GILBERT (1978) to account for the mosaic nature of eukaryotic genes. To be as clear as possible, I will use these terms according to the reviewed definitions as they were formulated and published during my work (WILBRANDT *et al.* (2017), see Table B.1.1).

In order to study genomes, a human- or machine-readable representation of the actual molecules has to be generated. While ‘genome’ strictly speaking refers to biological molecules – *i.e.*, polynucleotides consisting of adenine, guanine, thymine, and cytosine including their physical modifications (for example, methylation) – a ‘genome assembly’ (the result of an assembly process) relates to the representation of these molecules, a sequence of the letters A, G, T, and C. In the context of genomics analyses, these terms are often used interchangeably for

---

<sup>I.1</sup> Johannsen’s gene definition was based on eukaryotic nuclear chromosomes as carrier of hereditary information. The DNA of mitochondria, plastids, prokaryotes, and archaea is not necessarily organized chromosomal structures, and the term ‘linkage group’ might be more appropriate.

convenience. Most of the published assemblies of the last decade were obtained using next generation sequencing (NGS)/‘shotgun’ approaches (reviewed by, *e.g.*, BAKER, 2012). These have in common that reads, fragments of genomic DNA (in the case of genome sequencing) of a certain length, are sequenced and afterwards assembled into contiguous sequences (*i.e.*, contigs). Contigs can then be ordered and linked into scaffolds, which may contain gaps. The less contigs and scaffolds are obtained (assuming the absence of assembly errors), the better – the minimum achievable number naturally equals the chromosome (or equivalent DNA organizational unit) number of the given organism. Repetitive regions, *e.g.*, simple tandem repeats, telomeres, or transposable elements, can obfuscate the assembly process due to their self-matching nature (MCCOY *et al.*, 2014). Techniques (reviewed by PHILIPS *et al.*, 2017) that will hopefully allow the error-free sequencing or assembly of whole chromosomes are currently being developed, *e.g.*, by Oxford Nanopore and PacBio. Until these are widely employed, assembly quality (further discussed in section III.2.1) is a confounding factor in down-stream analyses, because the actual size and diversity of genomic component repertoires will be underestimated.

Even if the sequence of nucleotides along each DNA molecule is known, it is yet impossible to directly infer and interpret this information to build a living organisms *in vitro*. The complexity of processes leading from genotype (makeup of heritable traits) to phenotype (result of genotype expression under environmental influences, including behavior) is enormous. Reading and labelling the information encoded in the genomic sequence using a computer or human brain rather than the original cell is the task called ‘annotation’.

*The value of a genome is only as good as its annotation. It is the annotation that bridges the gap from the sequence to the biology of the organism.*

(STEIN, 2001)

Structural and functional annotation of protein-coding genes refers to the delineation of beginnings and ends of gene structures and the identification of the function(s) exerted by the gene’s protein product(s) (MQ ZHANG, 2002) (for a more detailed introduction to functional and structural annotation, see sections III.2 and C.1). Think of this analogy to illustrate (the problems of) gene annotation: if the genome is similar to the magnetic tape of a music cassette,

What do you do  
in  
**Comparative Genomics**


NP Science Communication Conference on 6-8 November 2017 in Berlin

FORSCHUNGS  
MUSEUM  
KOENIG

Mitglied der  
Leibniz  
Leibniz-Gemeinschaft

Characterizing gene repertoires or  
**DISCOVER your MUSIC**

Jeanne Wilbrandt<sup>1</sup>, Bernhard Misof<sup>1</sup>, Oliver Niehuis<sup>2</sup>



*Imagine ...*

- you, the notoriously curious  
and nerdy scientist,  
find a cassette.  
It's from your favorite band,  
**the Genome!**  
Your cassette player  
is sorely missing now...  
will you hear this fabulous music  
ever again?  
Have a go: how can you  
**identify the songs**  
of their 'Living Mystery' album?  
You start by laying out  
what you know:  
a song has start and end,  
but may include pauses.  
What else can you think of?  
Continue to build a helping machine  
and search for **signals** in the tape salad.  
Fail! You are puzzled,  
lost in **translation**.  
But what you get is  
- after all -  
discernible noise.  
Suddenly, stuck between  
two bags of coffee,  
you find **another** cassette!  
Another album of the Genome,  
one you never heard before.  
Finally  
a chance to dig  
into the rumor:  
did they **develop** a new style?  
Will the **comparison** help you to  
tune in on the music?

*Our reality ...*

— we strive to identify **genes** (regions on the band) in the **genome** (cassette band)  
that code for **proteins** (songs) and describe their **properties** (length, composition)  
to **compare** these to those of other **species** (albums).  
By this, we aim to illuminate the **evolution** of gene repertoires (music styles)  
and to understand the **diversity** of living organisms (music).

© jeanne.wilbrandt@leibniz-fmk.de  
1. Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany  
2. Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert Ludwig University of Freiburg, Hauptstr. 1, 79104 Freiburg, Germany  
our tool COGNATE can be found at doi:10.1186/s12864-017-3670-6

© 2017 Leibniz  
Leibniz-Gemeinschaft

Bundesministerium  
für Bildung  
und Forschung

Ministerium für Innovationen,  
Wissenschaft und Forschung  
des Landes Nordrhein-Westfalen

Figure I.1 – The cassette metaphor. (Continued on next page)

**Figure I.1 – The cassette metaphor.**

(Continued)

This poster was presented at the N<sup>2</sup> Science Communication Conference, where the topics of PhD studies were to be presented in a manner understandable to a laymen audience. The analogy of trying to read out the information encoded on a magnetic tape of a cassette without the original device used to explain the intricacies of protein-coding gene prediction in genomic sequences found broad appreciation.

the protein-coding genes can be imagined as songs encoded on it. We want to read the tape so that we can listen to the music, but cannot use a cassette player (the cell). To locate and decode the stretches of tape carrying relevant information, knowledge not only of the physical properties of the tape, but also of the properties of the songs (*e.g.*, start signal) is required (see also I.1).

Eukaryotic genomes comprise on average more than 20,000 protein-coding genes (ELLIOTT and GREGORY, 2015b). These are too many to be located individually by hand in reasonable time, thus the task of annotation demands automation. Since the 1980s, software is constantly being developed and improving (reviewed by, *e.g.*, BRENT and GUIGÓ, 2004, see section C.1). Automated annotation can make use of multiple lines of evidence, like the formalized knowledge of gene properties (*e.g.*, the ‘Kozak rules’, KOZAK, 1991) and sequence similarity, but can be severely hampered by incomplete assembly. In many cases, the review by human experts (referred to as ‘manual curation’) can remedy such problems (MISRA *et al.*, 2002). Thus it is often advocated as the only way to obtain reliable annotations (GUIGÓ *et al.*, 2006). The advantages and problems of automated and manual annotation are further introduced and discussed in chapter III.2.

## 1.2 Comparisons in biology

The nature of evolution directly implies a rationale for comparisons in biology. Since all organisms are related, we expect to find commonalities, stemming from a common ancestor. However, the similarity of two compared characters does by itself not imply the descent from a common precursor. Similarity alone is neither a sufficient nor a necessary criterion of evolutionary relatedness (EISEN, 1998). Nonetheless, it can be hypothesized that a relationship (homology)

exists, *i.e.*, the two characters are homologous, if the degree of similarity is greater than expected by chance among unrelated characters (KUZNIAR *et al.*, 2008). However, convergent evolution can result in similarities that are not a consequence of relatedness, termed homoplasies (EISEN, 1998). The comparison of genomes in context of phylogenetic relationships showed that the phylogenetic relationships of genes may be more complicated than those of the species they belong to. Three cases of homology need to be distinguished: (1) the relationship of two genetic elements originating from vertical descent, *i.e.*, a speciation event, is called ‘orthology’ (FITCH, 1970); (2) the relationship by a duplication event within the genome of an organism (within a species) is termed ‘paralogy’ (FITCH, 1970); (3) horizontal gene transfer (HGT; *i.e.*, transmission of genes between two organisms) leads to ‘xenology’ (GRAY and FITCH, 1983). Subunits of genes may be subject to individual evolutionary trajectories (*e.g.*, due to gene fission and fusion), thus the mentioned concepts are relevant at all levels of genomic comparison (KOONIN, 2005).

Sequence similarity is frequently used to predict both homology relationships among genes and genomic sequences as well as functional roles. Such inferences are often based on the ‘ortholog conjecture’ a.k.a. ‘ortholog functional conservation hypothesis’ (STUDER and ROBINSON-RECHAVI, 2009; PD THOMAS *et al.*, 2012). The conjecture states that orthologs are (functionally) more similar to each other than paralogs (of the same age) (GABALDÓN and KOONIN, 2013). The ortholog conjecture was challenged (*e.g.*, by claims of higher functional similarity among paralogs than among orthologs, NEHRT *et al.*, 2011), but has been vindicated by further evidence and bias control (*e.g.*, ALTENHOFF and DESSIMOZ, 2012; FORSLUND *et al.*, 2011; HENRICSON *et al.*, 2010; ROGOZIN and ROGOZIN, 2014; PD THOMAS *et al.*, 2012).

The high value of studying genomes for the life sciences is uncontested: only by using genome data is it possible to trace gene loss events (*i.e.*, confirm the absence of a gene), to identify orthologs and paralogs without ambiguity (*e.g.*, using three-way best reciprocal hits, ALTENHOFF and DESSIMOZ, 2012), and to investigate genome organization and their evolution by rearrangements and duplications (reviewed by, *e.g.*, KOONIN, 2009). Indeed these are the main questions tackled by early comparative genome studies (ELLEGREN, 2008). The comparison of genomes furthermore allows to trace both micro- and macro-evolutionary paths (*e.g.*, BRANSTETTER *et al.*, 2018) and helps to spotlight those



species with special features in contrast to others that reward being individually studied and questioned in detail.

Year	Organism	Reference
1976	MS2 (first bacteriophage genome)	FIERS <i>et al.</i> (1976)
1978	SV40 (first viral genome)	FIERS <i>et al.</i> (1978)
1981	human mitochondrion (first eukaryotic organellar genome)	ANDERSON <i>et al.</i> (1981)
1986	<i>Nicotiana tabacum</i> chloroplast (first chloroplast genome)	SHINOZAKI <i>et al.</i> (1986)
1992	<i>Saccharomyces cerevisiae</i> chromosome III (first chromosome)	OLIVER <i>et al.</i> (1992)
1995	<i>Haemophilus influenzae</i> (first bacterial genome, first free-living organism), <i>Mycoplasma genitalium</i>	FLEISCHMANN <i>et al.</i> (1995) and FRASER <i>et al.</i> (1995)
1996	<i>Saccharomyces cerevisiae</i> (first eukaryote genome)	GOFFEAU <i>et al.</i> (1996)
1997	first genomes of archaea	KLENK <i>et al.</i> (1997) and DR SMITH <i>et al.</i> (1997)
1997	<i>Escheria coli</i> (first bacterial model-organism)	BLATTNER <i>et al.</i> (1997)
1998	<i>Caenorhabditis elegans</i> (first multicellular organism)	THE C. ELEGANS SEQUENCING CONSORTIUM (1998)
1999	<i>Homo sapiens</i> chromosome 22 (first human chromosome)	DUNHAM <i>et al.</i> (1999)
2000	<i>Drosophila melanogaster</i> (first insect genome)	ADAMS (2000)
2000	<i>Arabidopsis thaliana</i> (first plant genome)	THE ARABIDOPSIS GENOME INITIATIVE (2000)
2001	<i>Homo sapiens</i> draft genomes	LANDER <i>et al.</i> (2001) and VENTER <i>et al.</i> (2001)
2002	<i>Tetraodon rubripes</i> draft genome (first fish genome)	APARICIO <i>et al.</i> (2002)
2002	<i>Mus musculus</i> (first whole-genome comparative analysis of mammals)	THE MOUSE GENOME SEQUENCING CONSORTIUM (2002)

**Table I.1** – Early milestones of genome sequencing (after PEVSNER, 2009, 527 ff.)

### 1.3 A small history of comparative genomics

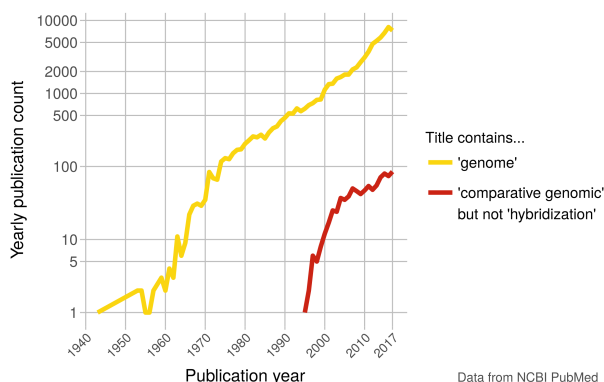
The first completely sequenced genomes were those of bacteriophages and viruses (FIERS *et al.*, 1976, 1978), published in the late 1970s. The following 25 years were marked by an accelerating succession of milestones (see Table I.1), for example, the first sequencing of an eukaryote (OLIVER *et al.*, 1992), a metazoan (THE C. ELEGANS SEQUENCING CONSORTIUM, 1998), and, finally, of a human (VENTER *et al.*, 2001). As soon as two sequenced genomes were available, these were compared. In early comparisons of viral and prokaryotic genomes, analyses were largely confined to protein-coding gene content and order (*e.g.*, GOLDBACH, 1987). Given the central role of protein-coding genes in carrying information translated into cellular building blocks and the long-matured foundations of analyses targeting them, this focus remained strong when new genomes were added (*e.g.*, RUBIN *et al.*, 2000).

The term ‘comparative genomics’<sup>1,2</sup> was used in its first instance to refer to gene mapping analyses (O’BRIEN and STANYON, 1999) but was coined in the 1990s to include any comparison of genomic data, when it became obvious that more and larger genome sequencing projects would be feasible, successful, and facilitate comparisons beyond the individual gene level (MS CLARK, 1999; ELGAR *et al.*, 1996). In other words, a broader understanding was adopted to “include any approach where the composition of different genomes is related to each other” (ELLEGRÉN, 2008).

Early comparative genomics research has focused on studies related to human disease (*e.g.*, identifying orthologs of disease genes and their network partners across genomes to gain further insight in intact and disturbed pathways; MS CLARK, 1999; NIERMAN *et al.*, 2000; RUBIN *et al.*, 2000). Since then, with the advent and ever-increasing use of NGS techniques, taxon coverage increased in depth (closely related species) and breadth (species from divergent lineages across clades). This, combined with the inclusion of non-model organisms, allowed to address questions of larger scope, and publications of genome sequencing projects and comparative genomic studies amassed constantly (see

---

<sup>1,2</sup> The term ‘genomics’ was coined in 1987 by Victor McKusick and Frank Ruddle as name for their newly founded journal (LEDERBERG and MCCRAY, 2001)



**Figure I.2** – Publications per year with title containing either 'genome' (yellow) or 'comparative genomics' (red, excluding the term 'hybridization'). Data from NCBI PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>. Last accessed 30 March 2018.)

Figure I.2). Consequently, research in comparative genomics split up according to the studied taxonomic groups, for example, bacteria (reviewed by LAND *et al.*, 2015), plants (reviewed by WENDEL *et al.*, 2016), vertebrates (reviewed by, *e.g.*, HEDGES and KUMAR, 2002; KELLEY *et al.*, 2016) and mammals (esp. hominins, reviewed by MURPHY *et al.*, 2004; ROGOZIN, 2014), and invertebrates, (reviewed by BLAXTER *et al.*, 2012; 15K CONSORTIUM, 2013; WURM, 2015) were often studied independently. The fundamental questions addressed in all these studies are nonetheless similar: which genomic features distinguish clades, and which commonalities can be found (from the genomic characteristics of a specific species down to the minimum gene set required to maintain a cell).

During the course of the last decade's comparative genomics enterprises, central tenets of (neo-) Darwinism have been re-evaluated; they were contextualized with the universal regularities that emerged from genome comparisons and gene expression profiling (surveyed by KOONIN, 2009). The major paradigm shifts regarding the nature of evolution can be summed up as follows: (1) evolution is not mainly driven by positive selection but rather the result of a combination of neutral processes and purifying selection; (2) evolution is less gradual than postulated (fueled by duplications, rearrangements, and HGT); and (3) there is not a single Tree of Life in which the whole course of evolution can be depicted (KOONIN, 2009, and references therein).

Genomics just unfolds its power to improve our notion of evolution (RICHARDS, 2015)<sup>I.3</sup>. In contrast to other comparative disciplines in biology, comparative genomics is not yet mature enough to pinpoint diagnostic genomic characteristics of species so that, comparable to morphological determination keys, species can be told apart by regarding their genomic features alone (if this is possible at all). Only recently, ELLIOTT and GREGORY (2015b) published a study named “What’s in a genome?”, aiming at structural and compositional features and thereby posing a question not only of great topical importance and central interest but also previously severely neglected at the genomic scale. This study exploited published data of over 520 species from four organism kingdoms for a quantitative and qualitative overview of correlations between a genome’s size and its structural features and components. Many conceivable questions were left open by the authors, prompting the general and specific questions I addressed during the years of pursuing a doctorate. These are outlined in section I.3.1 and addressed in the parts II to V.

---

<sup>I.3</sup> For a review of the history of insect phylogenetics and the integration of genomic data to resolve long-standing problems, see KJER *et al.* (2016).



---

## State of the art

---

### 2.1 Genome components and genome size variation

**T**HE WEALTH OF COMPONENTS that can be distinguished in a genome — protein-coding and non-coding genes, repetitive and transposable elements, and regulatory regions, just to name a few — and the diversity of their direct and mediated interactions is staggering. Research on how much of each component is contained within a genome is ongoing (*e.g.*, ELLIOTT and GREGORY, 2015b), and the question of whether there are amounts typical in a clade is yet unanswered (a question comparable to “how many legs are typical in each clade?” in a morphological context). The contribution of components to genome size can be a relevant factor in the evolution of genomes and species, either contributing building blocks or interfering with them.

Genome size and component sizes are described for a representative sample of Hymenoptera in this thesis (part IV). The relevance of recording these features follows from previous work on genome size and its biological correlates, which is introduced here.

Each of the above components contributes to genome size. Research has focused on genome size variation and the dynamics within the components contributing most to it. Eukaryotic genome size (sometimes given as 'C value') ranges from 2.3 Mbp in the parasite *Encephalitozoon intestinalis* (CORRADI *et al.*, 2010) to ca. 152 Gbp in the monocot flower *Paris japonica* (PELLICER *et al.*, 2010). This tremendous variation, however, does not coincide with organismic complexity in terms of, for example, cell type diversity (PETROV *et al.*, 2000; ZACHARIAS *et al.*, 2004)<sup>L1</sup> — constituting the so-called C-value enigma (GREGORY, 2002b; CA THOMAS, 1971).

Genome size does not consistently correlate with phenotypic traits. In vertebrates, links to biological correlates have been observed, *e.g.*, egg diameter or habitat stability in fishes (HARDIE and HEBERT, 2004; EM SMITH and GREGORY, 2009), or other phenotypic traits (PETROV, 2001). Contrastingly, no such correlative patterns present themselves in 18 species of New Zealand triplefin fishes (HICKEY and CLEMENTS, 2005), 115 spider species (GREGORY and SHORTHOUSE, 2003), and 22 aspidopods (LEFÉBURE *et al.*, 2017). Smaller genomes seem to be correlated to faster developmental rates (*e.g.*, in aphids, FINSTON *et al.*, 1995; and birds, GREGORY, 2002a) and potentially also to high metabolic rates (*e.g.*, in intracellular parasites, CAVALIER-SMITH, 2005; *Arabidopsis*, YF YANG *et al.*, 2013; mammals, VINOGRADOV, 1995; birds, Q ZHANG and EDWARDS, 2012; and fish, CHAURASIA *et al.*, 2014). A recent study showed that the comparatively small genome size of birds seems to be directly related to flight-associated body-indices (WRIGHT *et al.*, 2014). Only recently, a study suggested that genome sizes of insects mainly depend on phylogenetic relationships, potentially also developmental mode, while crustacean genome size appeared to be more correlated to habitats and life cycle (ALFSNES *et al.*, 2017).

One component that contributes considerably to genome size and genome size variation is that of transposable elements (CHÉNAIS *et al.*, 2012; GREGORY, 2005b; PETROV, 2001). The proportion of insect genomes made up of transposable elements ranges from as much as 60 % in the migratory locust (*Locusta migratoria*), with a genome size of 6.5 Gbp (H WANG *et al.*, 2014), to

---

<sup>L1</sup> Cell type diversity correlates with alternative splicing intensity (BUSH *et al.*, 2017; L CHEN *et al.*, 2014).



as little as 1 % in the tiny genome (90 Mbp) of the Antarctic midge *Belgica antarctica* (KELLEY *et al.*, 2014). Up to a genome size of 500 Mbp, a positive correlation of genome size and the diversity (in terms of superfamilies) of transposable elements has been documented across eukaryotes (ELLIOTT and GREGORY, 2015a). Genome size increases have been shown to correlate with accumulation of transposable elements in populations of small effective size due to a reduced selection efficacy (LEFÉBURE *et al.*, 2017). ELLIOTT and GREGORY (2015b) documented that an average share of 23 % of animal genome size is taken up by transposable elements, another ca. 27 % by other repetitive elements. Another considerable contribution to genome size stems from introns, as has been shown, *e.g.*, for plants (WENDEL *et al.*, 2002) and insects (X WANG *et al.*, 2014). Note that these studies may suffer from the ubiquitous bias towards sequencing small genomes (as they are less expensive to sequence, deductions are based on comparisons of small genomes) despite the inclusion of really large genomes in the latter study: comparing two extremes to a rather homogeneous baseline can result in a strong correlation.

Although the sizes of the compact genomes of bacteria and archaea mostly vary due to differences in protein-coding gene count (K HAN *et al.*, 2013), the contribution of protein-coding genes to eukaryotic and especially insect genome size variation is not extensively studied. According to ELLIOTT and GREGORY (2015b), averagely 10 % of an animal genome consist of coding exons.

Most studies concerned with the mechanisms influencing genome size have been performed on vertebrates. Thus, the following evidence refers mostly to this clade. The main influencers of genome constriction seem to be large segmental deletions (G ZHANG *et al.*, 2014) and slower insertion rates (NEAFSEY and PALUMBI, 2003; VINOGRADOV, 2004). DNA loss appears to be slower in species with large genomes (*e.g.*, salamanders compared to other vertebrates, PETROV *et al.*, 2000; SUN *et al.*, 2012), but shrinkage is possibly limited by functional constraints in comparatively small genomes (AHNERT *et al.*, 2008; GREGORY, 2005a). Recently, an 'accordion' model has been proposed to explain the relatively constant genomes sizes despite varying accumulation rates of transposable elements found in mammals and birds (KAPUSTA *et al.*, 2017); it suggests that genome expansions due to the activity of transposable elements are counteracted by large segmental deletions. Genome size reductions are accompanied by reduced repeat contents and shorter introns (MALMSTROM *et al.*, 2017; NEAFSEY and PALUMBI, 2003; G ZHANG *et al.*, 2014), but also by losses

of protein-coding genes (HUGHES and FRIEDMAN, 2008; MALMSTROM *et al.*, 2017; G ZHANG *et al.*, 2014) and a lower rate of large-scale insertions (NEAFSEY and PALUMBI, 2003).

## 2.2 Protein-coding gene repertoire evolution and dynamics

Protein-coding genes are the best studied elements among the genomic components, but with a focus on individual genes or specific repertoires, for example, the odorant receptor gene family (BRAND and RAMÍREZ, 2017), or opsin genes (FEUDA *et al.*, 2016). Consequently, there is only limited knowledge on their repertoire dynamics and evolvability that leads to differences in the abundance, composition, and structural configurations of genes and gene classes. This problem has been well described by GRAUR (2015):

*“The study of gene repertoire evolution is still in its infancy, and we know very little about the effectors of gene repertoire change. In fact, every time a new genome is published, one reads about the idiosyncrasies of that genome and the importance of this or that gene for adaptation. For example, much attention is given to autapomorphies. In particular, there exists a large body of literature dealing with human autapomorphies, or what “make us human” (e.g., MIKKELSEN, 2004; NEWTON, 2007; POLLARD, 2009).*

*Because we do not know much about the evolutionary forces driving gene repertoire, genome sequence publications frequently promise but rarely deliver coherent hypotheses concerning adaptation. For example, JIA et al. (2013) promise in their title that the “Aegilops tauschii genome sequence reveals a gene repertoire for wheat adaptation”. Sadly, the article bearing this title mentions no gene involved in wheat adaptation, nor for that matter what this adaptation consists of. Unfortunately, such dissonance between promises made and promises kept is quite common in the genomic literature.”*

DAN GRAUR, 2015, p. 538

This gap of knowledge elicited my interest in the composition of gene repertoires regarding conserved and lineage-specific genes. Structural characterization had to be the starting point, since so little is known (addressed

in all following parts of this thesis). A short survey of previous research is exposed here as a primer.

The gene count in animal gene repertoires is very similar even in differently sized genomes, averagely around 19,000 genes (ELLIOTT and GREGORY, 2015b). Each repertoire contains a core of genes that are conserved across all members of a clade (*e.g.*, the insect-specific conserved core contains ca. 45 % of all genes, WATERHOUSE, 2015).

Nonetheless, gene repertoires are not static: a large part of each repertoire consists of taxon-restricted genes (*e.g.*, HAHN, MV HAN, *et al.*, 2007; HUANG *et al.*, 2013; RÖDELSPERGER *et al.*, 2013; WATERHOUSE, 2015). For the set of these lineage- or species-specific genes, no orthologs or homologs can be found in the genus or clade. This set is sensitive to the given taxon sample during orthology delineation and might thus shrink with a perfect sample (KHALTURIN *et al.*, 2009). However, this set also inevitably includes truly novel genes (CARVUNIS *et al.*, 2012; TAUTZ and DOMAZET-LOŠO, 2011; WISSELER *et al.*, 2013). It appears that there is undeniable turnover, *i.e.*, (potentially balanced) gain and loss, within the repertoire, merely masked by a relatively constant gene count (HAHN, MV HAN, *et al.*, 2007).

Previous studies of the C-value enigma focused on mechanisms of gaining or losing non-coding DNA since this appears to be the part (mainly) responsible for genome size. This means, however, that the coding part has been assumed to be more or less invariable without controlling for the possibility of more subtle changes in the coding repertoire. The following two sections thus sum up what is generally known on gene structure and gene family composition as well as their correlates and evolution.

### 2.2.1 Gene structure dynamics and correlates

The investigation of structural features of the gene repertoire (as done in part IV and V), prompts that I offer results of previous research related to gene structure here.

Studying gene structure, the arrangement of exons and introns, involves (computational and conceptual) problems of intron and exon definition/recognition, splice site signals, and homologization, among others. These areas have been

objects of research since decades, focusing mainly on model organisms with high-confidence gene annotations and will not be discussed here further. The findings, however, were used in the development of automated annotation algorithms. Thus, automated annotations allow to focus on the characterization of lengths and densities of exons and introns (part IV).

The relationship of gene structure and expression level and breadth (number of tissues and developmental stages) has been explored in model organisms (*e.g.*, CAMIOLO *et al.*, 2009; LI *et al.*, 2007; PINGAULT *et al.*, 2015; RAO *et al.*, 2010), and several models explaining them have been proposed (reviewed by WOODY and SHOEMAKER (2011)). Due to the focus of their studies, the gene repertoire-wide trends of gene structure have not been addressed specifically.

The length of (coding) exons influences not only the resulting polypeptide sequence but also the success of competing interaction factors and thus expression levels (PETERSON *et al.*, 1994). It is thus reasonable to assume that their length is under selective pressure or even constraint. It has been shown that exon length seems to be constrained in vertebrates (to < 300 bp, ELLIOTT and GREGORY, 2015b; GELFMAN *et al.*, 2012) and potentially in insects (when comparing *Drosophila melanogaster* and *Locusta migratoria*, X WANG *et al.*, 2014), but the predominant hypothesis, that this is required for the identification of exon boundaries, does not hold (IT CHEN and CHASIN, 1994). Rather, at least in vertebrates, the strength of splice sites determines the recognizability of exons flanked by long (with strong splice sites) or short (with weak splice sites) introns (GELFMAN *et al.*, 2012). The insertion of introns into existing exons is independent of their length (RYABOV and GRIBSKOV, 2008), while exons seem to be more often inserted into longer introns (> 1000 bp), with shorter introns harboring older exons in mammals (M ROY *et al.*, 2008). Another length correlation has been found in humans: the longer an exon, the shorter is the next upstream intron (MQ ZHANG, 1998).

In drosophilids, it has been shown that intron length correlates with ortholog age and behaves clock-like (YANDELL *et al.*, 2006), but it is open whether this finding holds true in a broader scope. Positive as well as negative correlations of recombination rate and intron size have been found, the former in nematodes, the latter in fly and human (PRACHUMWAT *et al.*, 2004). A negative correlation of intron size, intron count as well as length of intergenic stretches and recombination rate has also been confirmed in Hymenoptera (in

*Nasonia vitripennis*, NIEHUIS *et al.*, 2010; the honeybee seems to be an exception, BEYE *et al.*, 2006), in line with the suggestion that increasing intron length is indirectly selected for by a low recombination rate (COMERON and KREITMAN, 2000). Furthermore, introns of highly expressed *C. elegans* genes are shorter by tendency, potentially driven by selection towards minimal transcription costs (CASTILLO-DAVIS *et al.*, 2002). Plants seem to have a lower intron gain rate than other eukaryotes in recent evolution (SW ROY and PENNY, 2007), whereas nematode genes have a higher intron turn-over rate than those of *Drosophila* or mammals (CUTTER *et al.*, 2009). Most eukaryotic lineage evolution appears to be characterized by intron loss with few spurts of massive intron gain, possibly during major evolutionary steps; inference shows that the last common ancestor of eukaryotes featured intron-rich genes (ROGOZIN *et al.*, 2012). This overall loss-trend has been specifically confirmed in flies (AG CLARK *et al.*, 2007; YANDELL *et al.*, 2006) and mammals (COULOMBE-HUNTINGTON and MAJEWSKI, 2007). Contrastingly, it has also been claimed that conserved genes tend to gain and not lose introns in eukaryotes (CARMEL *et al.*, 2007). Proposed mechanisms of intron gain and loss have been reviewed by BELSHAW and BENSASSON (2006), research on the direct and indirect roles of introns has been summarized by JO and CHOI (2015). A detail rendition of their findings is beyond the scope of this thesis.

### 2.2.2 Gene family dynamics

Gene repertoires consist of gene families. Members of gene families are usually considered to be paralogous, although this is relative to the considered phylogenetic split. For convenience, even a gene without identifiable homologous genes in its genomic environment can be considered its own gene family. What is known about gene family sizes, the relation to biological diversity, and the impact on gene structure distributions in a repertoire?

Over 40 % of all gene families differ in size among twelve drosophilid species, while the gene birth rate is similar to that of yeast and mammals (HAHN, MV HAN, *et al.*, 2007). Roughly 3 % evolved at significantly elevated rates, which indicated non-neutral evolutionary processes and stimulated functional studies (AG CLARK *et al.*, 2007). It became apparent that the overall similar total gene

number masked a strong turnover via individual gene gain and loss (HAHN, MV HAN, *et al.*, 2007).

There is no direct correlation between duplication and the evolution of species diversity — note the disparity between time scales of the considered processes. It has been proposed that the role of duplication is confined to the provision of functional redundancy, thus increasing mutational robustness and reducing the risk of extinction (CROW and WAGNER, 2006). The widely used hypothesis (KONDRASHOV *et al.*, 2002; NEMBAWARE *et al.*, 2002) of increased mutation rate due to relaxed selection pressure on paralogs implies that duplicates arise from genes that are unbiased regarding their evolutionary rates, which is not necessarily valid (DAVIS and PETROV, 2004). In contrast, the main source of duplicates seems to be slowly evolving genes, a strongly biased set (DAVIS and PETROV, 2004; JORDAN *et al.*, 2004). Following a duplication, paralogs are retained (FORCE *et al.*, 1999), but selective pressure appears to be asymmetric, facilitating thus biased gene loss and differing retention rates, at least in fish and yeast (BRUNET *et al.*, 2006; HOLLAND *et al.*, 2017). Several models have been suggested to describe the fates of duplicates (JIANG and ASSIS, 2017; ZHAO *et al.*, 2015) and to explain gene retention after duplication (gene dosage balance, functional buffering, subfunctionalization; reviewed in EDGER and PIRES, 2009).

Paralogous genes seem not to follow the general trend of prevalent intron loss. In *Plasmodium* paralogs, accelerated rates of both intron gain and loss have been found (CASTILLO-DAVIS *et al.*, 2004), whereas other eukaryotic (plant) paralogs seem to predominantly gain introns (but in restricted time frames) (BABENKO *et al.*, 2004; KNOWLES and MCLYSAGHT, 2006).

The observation of widely shared single-copy status (WATERHOUSE *et al.*, 2011) instigates the question of whether there is a functional class producing such genes without paralogs. Potentially, dosage sensitivity plays a role, since dosage-sensitive genes are less random and have characteristics that otherwise help discern orthologs (less copy number variation, lower transposition frequency, etc.; EDGER and PIRES, 2009).

### 2.2.3 Gene and gene family turnover

When considering the differential conservation of genes within gene repertoires (some ancient gene families can be found in all living organisms, while others are either old, but lonely or arose as lineage-specific novelty) as in part V of this thesis, investigations have to be built upon the state of the art. Thus, I will shortly review what is known regarding the gain and loss of genes and gene families.

Gene turnover, the change of a repertoire through time by gain and loss of genes (and thereby also families), has been addressed previously. Interestingly, many species of special interest have similar gene turnover rates (measured as gains and losses per gene per million years [ $\lambda$ , HAHN *et al.*, 2005; alternatively birth and death rate are given separately, *e.g.*, ALMEIDA *et al.*, 2014]): yeasts have a  $\lambda$  of 0.002 (HAHN *et al.*, 2005; LYNCH and CONERY, 2003), slightly lower levels have been found in fruit flies ( $\lambda = 0.0012$ , HAHN, MV HAN, *et al.*, 2007) and mammals ( $\lambda = 0.0016$ , DEMUTH *et al.*, 2006; HAHN, DEMUTH, *et al.*, 2007), and slightly higher levels in plants ( $\lambda = 0.003$ , CARRETERO-PAULET *et al.*, 2015). Also, gene turnover rates vary among lineages and gene families (ALMEIDA *et al.*, 2014).

There is evidence that lineage-specific gene family expansions can advance an organism's capacity to adapt by providing a source to diversify structures or regulatory networks (FORÊT and MALESZKA, 2006; HAHN, MV HAN, *et al.*, 2007; KONDRASHOV *et al.*, 2002; VIDAL *et al.*, 2016; Z WANG *et al.*, 2012). For example, the response to pathogens and environmental stress in eukaryotes was found to be connected to lineage-specific expansions (LESPINET *et al.*, 2002). Additionally, the expansion of some (super)families is correlated with the number of cell types in eukaryotic organisms (VOGEL and CHOTHIA, 2006). Finally, an expansion likely played a role in zebrafish immunity (HOWE *et al.*, 2016). Another recent study found that contractions and expansions of transcription factor families correlate with species-specific alterations of organ formation regulation in plants (CARRETERO-PAULET *et al.*, 2015). Hypothetically, specification of morphology (KHALTURIN *et al.*, 2009) as well as life-cycle adaptations (ZHAO *et al.*, 2015) might be driven by taxon-restricted genes. Thus, evidence is accumulating that gene turnover contributes as an important factor to the evolution of diverse organisms.

Despite the potentially significant role of gene turnover in genomic change, studies were limited to a very small taxon sample or failed to sufficiently address confounding factors, for example by accounting for species divergence times in a convincing manner and using a robust phylogeny. A very basic question is also still left open: does the gene turnover rate change over time and between lineages, or is it time-constant as would be expected when a classical birth-and-death model applies (HAHN *et al.*, 2005). Evidence has been reported that supports the rate changes over time, (*e.g.*, CARRETERO-PAULET *et al.*, 2015; HAHN, DEMUTH, *et al.*, 2007; RAPPOPORT and LINIAL, 2015; WISSLER *et al.*, 2013), but genomic factors facilitating and driving gene turnover rate changes still have to be determined. Likely candidate mechanisms are whole genome duplication (*e.g.*, BLOMME *et al.*, 2006) and ectopic (non-homologous) recombination due to the existence and spreading of transposable elements within a genome (*e.g.*, ALBERTIN *et al.*, 2015; S YANG *et al.*, 2008). It is also not clear whether lineage-specific bursts of gene family size are evolutionary neutral (potentially due to a cost in genomic stability, of regulation, or of expression) or rather driven by positive selection (SCHIFFER *et al.*, 2016). Another open point is the extent of contribution of gene turnover and repertoire diversification to the astonishing diversity of insects.

Investigations of gene turnover rates in hexapods are rare, despite the clade's significance as highly diverse group of extant organisms. There are studies that compared gene repertoires of only few species from individual insect orders (*e.g.*, limited to Hymenoptera or Diptera) and others that focused on a limited set of genes. Hymenoptera were studied with a focus on ants (*e.g.*, BONASIO *et al.*, 2010; ROUX *et al.*, 2014). Dipterans (apart from drosophilids, HAHN, MV HAN, *et al.*, 2007) were best covered with regard to midges (GUSEV *et al.*, 2014) and mosquitoes (NEAFSEY *et al.*, 2015). A more comprehensive study was conducted by ZDOBNOV and BORK (2007), covering the holometabolous 'Big Four' (the four most species-rich insect orders: Coleoptera, Diptera, Hymenoptera, Lepidoptera) with twelve species. Most notably, they reported less conservation at the amino acid level among orthologous genes and less orthologs in synteny than studies focused on vertebrates with roughly the same phylogenetic age (ZDOBNOV *et al.*, 2002, 2005). To my knowledge, only one study investigated explicitly the gene turnover rates across Holometabola, namely nine hymenopterans and six dipterans (compared to three outgroup species): RAPPOPORT and LINIAL (2015) found that gains and losses of gene families



were more frequent in hymenopterans than in dipterans. The above findings also vindicate the impression that gene repertoire content is much more volatile in insect genomes than in vertebrate genomes. Until now, however, drawing general conclusions was forestalled by the difficulties of compiling datasets from comparable methods that included accurate divergence time estimates and thus reliable turnover rates and that covered a meaningful sample across insects. Furthermore, none of these studies addressed the mechanistic causes of the surprisingly high changeability of insect gene repertoires and genomes.

Thus, insect gene turnover and its correlates await exploration. Prerequisites are reasonable taxon sample, a dated phylogeny, and reliable gene annotation, as well as the means for primary gene repertoire characterization. It turned out that these basic requirements were not fully available. My thesis focus and aims thus oriented towards resourcing these requirements.



---

## Thesis focus and aim

---

**E**NDOPTERYGOTA, or Holometabola, are insects that undergo metamorphosis during their ontogeny (as opposed to a direct development by growth and moulding in hemimetabolous insects and other arthropods). This group has the highest extant species-richness and diversity among insects, with the highest 'concentration' of diversity in the 'Big Four', the orders Lepidoptera (butterflies and moths), Coleoptera (beetles), Diptera (flies), and Hymenoptera (sawflies, bees, wasps, and ants). Recently, comprehensive and reliable backbone (KJER *et al.*, 2015; MISOF *et al.*, 2014) and crown group phylogenies (*e.g.*, FOSTER *et al.*, 2017; PETERS *et al.*, 2017; SANN *et al.*, 2018; SQ ZHANG *et al.*, 2018) have been published. Given their phylogeny comprising both ancient and recent divergence times and their diversity, holometabolous insects and insects in general appear to be a rewarding group to study genome evolution in all aspects.

The restricted taxon sample of previous repertoire evolution studies leaves open whether they revealed global or lineage-specific patterns. Certainly, not only gene repertoire and family variation may influence genome evolution, but also the structure of genes themselves. Previous comparisons of gene repertoires

and repertoire dynamics in insects were either focused on few genes, gene expression correlates, few species, or depended on data collected with methods of insecure comparability. Elucidating insect gene repertoire composition and variability requires the investigation of presence and structures of single copy genes, gene families, and their distributions among clades (*i.e.*, identification of widespread/taxon-restricted genes). Such a comprehensive approach is unprecedented and thus well suited to tackle a considerable knowledge gap.

The most surprising gap of knowledge when exploring protein-coding gene repertoires, however, is very basic: there were no standards and no standard tools to describe the structure of genes across the whole gene repertoire found in a genome, and no means to compare for example distributions of structural parameters across species. Essentially, there is little known of what to expect when looking at an insect gene repertoire: “What’s in a genome?”, the question posed by ELLIOTT and GREGORY (2015b), guided my explorations.

## 3.1 Research questions

### 3.1.1 Hymenopteran gene repertoire structures?

*What are features, structures, and variability of gene repertoires of hymenopteran genomes?*

The overarching topic – protein-coding gene repertoire evolution – is with this question reduced to a more workable size, focusing only on hymenopteran genomes. Hymenoptera are one of the most diverse animal groups on earth (species-richness given their phylogenetic age), thus it can be expected that their genomes show traces of diversification.

I aim to structurally characterize the repertoires of Hymenoptera and outgroup insects with the goal to elucidate whether a phyletic pattern can be identified, *i.e.*, a combination of features that can be related to a group. This subtopic is addressed in part IV.

Furthermore, I approach the above question by partitioning hymenopteran gene repertoires according to their conservation across the phylogeny and according

to the duplicability of identified gene families. With a characterization of the partitions' gene structures and domain diversity (as proxy for functional diversity), I address the question of variability. This subject is presented in part V.

### 3.1.2 Adequacy of automatically annotated gene models?

*How suitable are automatically inferred models for uncovering taxon-specific gene structural differences in gene repertoires?*

My explorations critically depend on structural gene annotations. Automated annotation procedures still produce false predictions and depend on assembly quality. It has been shown that down-stream analyses can be severely misled. This prompted the question whether the above study could be conducted at all using solely automatically generated gene models. I investigate in part III whether gene models that were generated automatically, *i.e.*, using unsupervised pipelines, are accurate enough to be used to address questions of gene structure evolution.

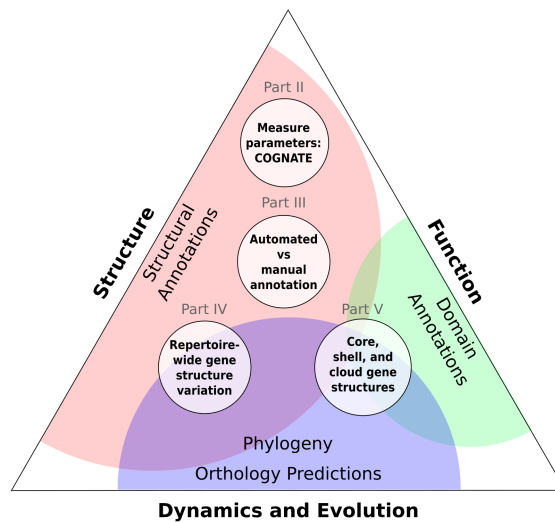
### 3.1.3 A tool to characterize gene structure?

*How can protein-coding genes be structurally described?*

The first prerequisite to describe gene repertoires is a method to obtain the necessary data, preferably in a manner comparable to previously published data. However, there was no standard of data publication when a genome and its repertoire is released. Also, no standard method to acquire this data from own annotations was available. Thus, the first task is to identify sensible parameters and to contrive a workflow to measure these efficiently. Special attention is given to the implementation of clearly outlined definitions. Problems of standardization and possible solutions including my newly developed tool COGNATE are described in part II.

### 3.2 Research map

The research map (Figure I.3) provides a schematic overview of the three main aspects or research areas of protein-coding gene repertoires my work focused on (sides of the triangle) and the used data types / topics (colored circles). The small circles depict the research questions explored in the following parts of this thesis. While the degrees of intersection among the topics are not accurately represented, the figures nicely illustrates the focal points of the presented work.



**Figure I.3 – Research map.** Research areas, data types, research questions addressed in this thesis.

---

# Bibliography I

---

- ADAMS, MD (2000). **The Genome Sequence of *Drosophila melanogaster***. *Science* 287.5461, pp. 2185–2195. DOI: 10.1126/science.287.5461.2185 (cit. on p. 10).
- AHNERT, SE, TM FINK, and A ZINOVYEV (2008). **How much non-coding DNA do eukaryotes require?** *Journal of Theoretical Biology* 252.4, pp. 587–592. DOI: 10.1016/j.jtbi.2008.02.005 (cit. on p. 17).
- ALBERTIN, CB, O SIMAKOV, T MITROS, ZY WANG, JR PUNGOR, E EDSINGER-GONZALES, S BRENNER, CW RAGSDALE, and DS ROKHSAR (2015). **The octopus genome and the evolution of cephalopod neural and morphological novelties.** *Nature* 524.7564, pp. 220–224. DOI: 10.1038/nature14668 (cit. on p. 24).
- ALFSNES, K, HP LEINAAS, and DO HESSEN (2017). **Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans.** *Ecology and Evolution*, pp. 1–9. DOI: 10.1002/ece3.3163 (cit. on p. 16).
- ALMEIDA, FC, A SÁNCHEZ-GRACIA, JL CAMPOS, and J ROZAS (2014). **Family Size Evolution in *Drosophila* Chemosensory Gene Families: A Comparative Analysis with a Critical Appraisal of Methods.** *Genome Biology and Evolution* 6.7, pp. 1669–1682. DOI: 10.1093/gbe/evu130 (cit. on p. 23).
- ALTENHOFF, AM and C DESSIMOZ (2012). “Inferring Orthology and Paralogy”. *Evolutionary Genomics*. Ed. by M ANISIMOVA. Vol. 855. Methods in Molecular Biology. Humana Press, pp. 259–279 (cit. on p. 8).
- ANDERSON, S, AT BANKIER, BG BARRELL, MHLd BRUIJN, AR COULSON, J DROUIN, IC EPERON, DP NIERLICH, BA ROE, F SANGER, PH SCHREIER, AJH SMITH, R STADEN, and IG YOUNG (1981). **Sequence and organization of the human mitochondrial genome.** *Nature* 290.5806, pp. 457–465. DOI: 10.1038/290457a0 (cit. on p. 10).
- APARICIO, S, J CHAPMAN, E STUPKA, N PUTNAM, Jm CHIA, P DEHAL, A CHRISTOFFELS, S RASH, S HOON, A SMIT, MDS GELPKE, J ROACH, T OH, IY HO, M WONG, C DETTER, F VERHOEF, P PREDKI, A TAY, S LUCAS, P RICHARDSON, SF SMITH, MS CLARK, YJK EDWARDS, N DOGGETT, A ZHARKIKH, SV TAVTIGIAN, D PRUSS, M BARNSTEAD, C EVANS, H BADEN, J POWELL, G GLUSMAN, L ROWEN,

- L HOOD, YH TAN, G ELGAR, T HAWKINS, B VENKATESH, D ROKHSAR, and S BRENNER (2002). **Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes***. *Science* 297.5585, pp. 1301–1310. DOI: 10.1126/science.1072104 (cit. on p. 10).
- BABENKO, VN, IB ROGOZIN, SL MEKHEDOV, and EV KOONIN (2004). **Prevalence of intron gain over intron loss in the evolution of paralogous gene families**. *Nucleic Acids Research* 32.12, pp. 3724–3733. DOI: 10.1093/nar/gkh686 (cit. on p. 22).
- BAKER, M (2012). **De novo genome assembly: what every biologist should know**. *Nature Methods* 9.4, pp. 333–337. DOI: 10.1038/nmeth.1935 (cit. on p. 5).
- BELSHAW, R and D BENSASSON (2006). **The rise and falls of introns**. *Heredity* 96.3, pp. 208–213. DOI: 10.1038/sj.hdy.6800791 (cit. on p. 21).
- BEYE, M, I GATTERMEIER, M HASSELMANN, T GEMPE, M SCHIOETT, JF BAINES, D SCHLIPALIUS, F MOUGEL, C EMORE, O RUEPPELL, A SIRVIÖ, E GUZMÁN-NOVOA, G HUNT, M SOLIGNAC, and RE PAGE (2006). **Exceptionally high levels of recombination across the honey bee genome**. *Genome Research* 16.11, pp. 1339–1344. DOI: 10.1101/gr.5680406 (cit. on p. 21).
- BLATTNER, FR, G PLUNKETT, CA BLOCH, NT PERNA, V BURLAND, M RILEY, J COLLADO-VIDES, JD GLASNER, CK RODE, GF MAYHEW, J GREGOR, NW DAVIS, HA KIRKPATRICK, MA GOEDEN, DJ ROSE, B MAU, and Y SHAO (1997). **The Complete Genome Sequence of *Escherichia coli* K-12**. *Science* 277.5331, pp. 1453–1462. DOI: 10.1126/science.277.5331.1453 (cit. on p. 10).
- BLAXTER, M, S KUMAR, G KAUR, G KOUTSOVOULOS, and B ELSWORTH (2012). **Genomics and transcriptomics across the diversity of the Nematoda**. *Parasite Immunology* 34.2, pp. 108–120. DOI: 10.1111/j.1365-3024.2011.01342.x (cit. on p. 12).
- BLOMME, T, K VANDEPOELE, SD BODT, C SIMILLION, S MAERE, and YVd PEER (2006). **The gain and loss of genes during 600 million years of vertebrate evolution**. *Genome Biology* 7.5, R43. DOI: 10.1186/gb-2006-7-5-r43 (cit. on p. 24).
- BONASIO, R, G ZHANG, C YE, NS MUTTI, X FANG, N QIN, G DONAHUE, P YANG, Q LI, C LI, P ZHANG, Z HUANG, SL BERGER, D REINBERG, J WANG, and J LIEBIG (2010). **Genomic Comparison of the Ants *Camponotus floridanus* and *Harpegnathos saltator***. *Science* 329.5995, pp. 1068–1071. DOI: 10.1126/science.1192428 (cit. on p. 24).
- BRAND, P and SR RAMÍREZ (2017). **The Evolutionary Dynamics of the Odorant Receptor Gene Family in Corbiculate Bees**. *Genome Biology and Evolution* 9.8, pp. 2023–2036. DOI: 10.1093/gbe/evx149 (cit. on p. 18).
- BRANSTETTER, M, AK CHILDERS, D COX-FOSTER, KR HOPPER, KM KAPHEIM, AL TOTH, and KC WORLEY (2018). **Genomes of the Hymenoptera**. *Current Opinion in Insect Science*. DOI: 10.1016/j.cois.2017.11.008 (cit. on p. 8).
- BRENT, MR and R GUIGÓ (2004). **Recent advances in gene structure prediction**. *Current Opinion in Structural Biology* 14.3, pp. 264–272. DOI: 10.1016/j.sbi.2004.05.007 (cit. on p. 7).



- BRUNET, FG, HR CROLLIUS, M PARIS, JM AURY, P GIBERT, O JAILLON, V LAUDET, and M ROBINSON-RECHAVI (2006). **Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes.** *Molecular Biology and Evolution* 23.9, pp. 1808–1816. DOI: 10.1093/molbev/msl049 (cit. on p. 22).
- BUSH, SJ, L CHEN, JM TOVAR-CORONA, and AO URRUTIA (2017). **Alternative splicing and the evolution of phenotypic novelty.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1713, p. 20150474. DOI: 10.1098/rstb.2015.0474 (cit. on p. 16).
- CAMIOLO, S, D RAU, and A PORCEDDU (2009). **Mutational Biases and Selective Forces Shaping the Structure of Arabidopsis Genes.** *PLoS ONE* 4.7. DOI: 10.1371/journal.pone.0006356 (cit. on p. 20).
- CARMEL, L, IB ROGOZIN, YI WOLF, and EV KOONIN (2007). **Evolutionarily conserved genes preferentially accumulate introns.** *Genome Research* 17.7, pp. 1045–1050. DOI: 10.1101/gr.5978207 (cit. on p. 21).
- CARRETERO-PAULET, L, P LIBRADO, TH CHANG, E IBARRA-LACLETTE, L HERRERA-ESTRELLA, J ROZAS, and VA ALBERT (2015). **High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant *Utricularia gibba*.** *Molecular Biology and Evolution* 32.5, pp. 1284–1295. DOI: 10.1093/molbev/msv020 (cit. on pp. 23, 24).
- CARVUNIS, AR, T ROLLAND, I WAPINSKI, MA CALDERWOOD, MA YILDIRIM, N SIMONIS, B CHARLOTEAUX, CA HIDALGO, J BARBETTE, B SANTHANAM, GA BRAR, JS WEISSMAN, A REGEV, N THIERRY-MIEG, ME CUSICK, and M VIDAL (2012). **Proto-genes and de novo gene birth.** *Nature* 487.7407, pp. 370–374. DOI: 10.1038/nature11184 (cit. on p. 19).
- CASTILLO-DAVIS, CI, TBC BEDFORD, and DL HARTL (2004). **Accelerated Rates of Intron Gain/Loss and Protein Evolution in Duplicate Genes in Human and Mouse Malaria Parasites.** *Molecular Biology and Evolution* 21.7, pp. 1422–1427. DOI: 10.1093/molbev/msh143 (cit. on p. 22).
- CASTILLO-DAVIS, CI, SL MEKHEDOV, DL HARTL, EV KOONIN, and FA KONDRASHOV (2002). **Selection for short introns in highly expressed genes.** *Nature Genetics* 31.4, pp. 415–418. DOI: 10.1038/ng940 (cit. on p. 21).
- CAVALIER-SMITH, T (2005). **Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion.** *Annals of Botany* 95.1, pp. 147–175. DOI: 10.1093/aob/mci010 (cit. on p. 16).
- CHAURASIA, A, A TARALLO, L BERNÀ, M YAGI, C AGNISOLA, and G D'ONOFRIO (2014). **Length and GC Content Variability of Introns among Teleostean Genomes in the Light of the Metabolic Rate Hypothesis.** *PLoS ONE* 9.8. DOI: 10.1371/journal.pone.0103889 (cit. on p. 16).
- CHEN, IT and LA CHASIN (1994). **Large exon size does not limit splicing in vivo.** *Molecular and Cellular Biology* 14.3, pp. 2140–2146. DOI: 10.1128/MCB.14.3.2140 (cit. on p. 20).
- CHEN, L, SJ BUSH, JM TOVAR-CORONA, A CASTILLO-MORALES, and AO URRUTIA (2014). **Correcting for Differential Transcript Coverage Reveals a Strong Relation-**

- ship between Alternative Splicing and Organism Complexity. *Molecular Biology and Evolution* 31.6, pp. 1402–1413. DOI: 10.1093/molbev/msu083 (cit. on p. 16).
- CHÉNAIS, B, A CARUSO, S HIARD, and N CASSE (2012). **The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments.** *Gene* 509.1, pp. 7–15. DOI: 10.1016/j.gene.2012.07.042 (cit. on p. 16).
- CLARK, AG *et al.* (2007). **Evolution of genes and genomes on the *Drosophila* phylogeny.** *Nature* 450.7167, pp. 203–218. DOI: 10.1038/nature06341 (cit. on p. 21).
- CLARK, MS (1999). **Comparative genomics: the key to understanding the human genome project.** *BioEssays* 21.2, pp. 121–130. DOI: 10.1002/(SICI)1521-1878(199902)21:2<121::AID-BIES6>3.0.CO;2-O (cit. on p. 11).
- COMERON, JM and M KREITMAN (2000). **The Correlation Between Intron Length and Recombination in *Drosophila*: Dynamic Equilibrium Between Mutational and Selective Forces.** *Genetics* 156.3, pp. 1175–1190 (cit. on p. 21).
- CORRADI, N, JF POMBERT, L FARINELLI, ES DIDIER, and PJ KEELING (2010). **The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*.** *Nature Communications* 1, p. 77. DOI: 10.1038/ncomms1082 (cit. on p. 16).
- COULOMBE-HUNTINGTON, J and J MAJEWSKI (2007). **Characterization of intron loss events in mammals.** *Genome Research* 17.1, pp. 23–32. DOI: 10.1101/gr.5703406 (cit. on p. 21).
- CROW, KD and GP WAGNER (2006). **What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity?** *Molecular Biology and Evolution* 23.5, pp. 887–892. DOI: 10.1093/molbev/msj083 (cit. on p. 22).
- CUTTER, AD, A DEY, and RL MURRAY (2009). **Evolution of the *Caenorhabditis elegans* Genome.** *Molecular Biology and Evolution* 26.6, pp. 1199–1234. DOI: 10.1093/molbev/msp048 (cit. on p. 21).
- DAVIS, JC and DA PETROV (2004). **Preferential Duplication of Conserved Proteins in Eukaryotic Genomes.** *PLoS Biol* 2.3, e55. DOI: 10.1371/journal.pbio.0020055 (cit. on p. 22).
- DEMUTH, JP, TD BIE, JE STAJICH, N CRISTIANINI, and MW HAHN (2006). **The Evolution of Mammalian Gene Families.** *PLoS ONE* 1.1, e85. DOI: 10.1371/journal.pone.0000085 (cit. on p. 23).
- DUNHAM, I *et al.* (1999). **The DNA sequence of human chromosome 22.** *Nature* 402.6761, pp. 489–495. DOI: 10.1038/990031 (cit. on p. 10).
- EDGER, PP and JC PIRES (2009). **Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes.** *Chromosome Research* 17.5, pp. 699–717. DOI: 10.1007/s10577-009-9055-9 (cit. on p. 22).
- EISEN, JA (1998). **Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis.** *Genome Research* 8.3, pp. 163–167. DOI: 10.1101/gr.8.3.163 (cit. on pp. 7, 8).

- ELGAR, G, R SANDFORD, S APARICIO, A MACRAE, B VENKATESH, and S BRENNER (1996). **Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*)**. *Trends in Genetics* 12.4, pp. 145–150. DOI: 10.1016/0168-9525(96)10018-4 (cit. on p. 11).
- ELLEGREN, H (2008). **Comparative genomics and the study of evolution by natural selection**. *Molecular Ecology* 17.21, pp. 4586–4596. DOI: 10.1111/j.1365-294X.2008.03954.x (cit. on pp. 8, 11).
- ELLIOTT, TA and TR GREGORY (2015a). **Do larger genomes contain more diverse transposable elements?** *BMC Evolutionary Biology* 15, p. 69. DOI: 10.1186/s12862-015-0339-8 (cit. on p. 17).
- (2015b). **What's in a genome? The C-value enigma and the evolution of eukaryotic genome content**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678, p. 20140331. DOI: 10.1098/rstb.2014.0331 (cit. on pp. 7, 13, 15, 17, 19, 20, 28).
- FALK, R (1986). **What is a gene?** *Studies in History and Philosophy of Science Part A* 17.2, pp. 133–173. DOI: 10.1016/0039-3681(86)90024-5 (cit. on p. 4).
- FEUDA, R, F MARLÉTAZ, MA BENTLEY, and PW HOLLAND (2016). **Conservation, Duplication, and Divergence of Five Opsin Genes in Insect Evolution**. *Genome Biology and Evolution* 8.3, pp. 579–587. DOI: 10.1093/gbe/evw015 (cit. on p. 18).
- FIERS, W, R CONTRERAS, F DUERINCK, G HAEGEMAN, D ISERENTANT, J MERREGAERT, WM JOU, F MOLEMANS, A RAEYMAEKERS, AVd BERGHE, G VOLCKAERT, and M YSEBAERT (1976). **Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene**. *Nature* 260.5551, pp. 500–507. DOI: 10.1038/260500a0 (cit. on pp. 10, 11).
- FIERS, W, R CONTRERAS, G HAEGEMAN, R ROGIERS, AVd VOORDE, HV HEUVERSWYN, JV HERREWEGHE, G VOLCKAERT, and M YSEBAERT (1978). **Complete nucleotide sequence of SV40 DNA**. *Nature* 273.5658, pp. 113–120. DOI: 10.1038/273113a0 (cit. on pp. 10, 11).
- FINSTON, TL, PDN HEBERT, and RB FOOTIT (1995). **Genome size variation in aphids**. *Insect Biochemistry and Molecular Biology* 25.2, pp. 189–196. DOI: 10.1016/0965-1748(94)00050-R (cit. on p. 16).
- FITCH, WM (1970). **Distinguishing Homologous from Analogous Proteins**. *Systematic Biology* 19.2, pp. 99–113. DOI: 10.2307/2412448 (cit. on p. 8).
- FLEISCHMANN, RD, MD ADAMS, O WHITE, RA CLAYTON, EF KIRKNESS, AR KERLAVAGE, CJ BULT, JF TOMB, BA DOUGHERTY, JM MERRICK, and E AL (1995). **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 269.5223, pp. 496–512. DOI: 10.1126/science.7542800 (cit. on p. 10).
- FORCE, A, M LYNCH, FB PICKETT, A AMORES, YI YAN, and J POSTLETHWAIT (1999). **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 151.4, pp. 1531–1545 (cit. on p. 22).
- FORÊT, S and R MALESZKA (2006). **Function and evolution of a gene family encoding odorant binding-like proteins in a social insect, the honey bee (*Apis mellifera*)**. *Genome Research* 16.11, pp. 1404–1413. DOI: 10.1101/gr.5075706 (cit. on p. 23).

- FORSLUND, K, I PEKKARI, and ELL SONNHAMMER (2011). **Domain architecture conservation in orthologs.** *BMC Bioinformatics* 12.1, p. 326 (cit. on p. 8).
- FOSTER, PG, TMP de OLIVEIRA, ES BERGO, JE CONN, DC SANT'ANA, SS NAGAKI, S NIHEI, CE LAMAS, C GONZÁLEZ, CC MOREIRA, and MAM SALLUM (2017). **Phylogeny of Anophelinae using mitochondrial protein coding genes.** *Royal Society Open Science* 4.11. DOI: 10.1098/rsos.170758 (cit. on p. 27).
- FRASER, CM, JD GOCAYNE, O WHITE, MD ADAMS, RA CLAYTON, RD FLEISCHMANN, CJ BULT, AR KERLAVAGE, G SUTTON, JM KELLEY, JL FRITCHMAN, JF WEIDMAN, KV SMALL, M SANDUSKY, J FUHRMANN, D NGUYEN, TR UTTERBACK, DM SAUDEK, CA PHILLIPS, JM MERRICK, JF TOMB, BA DOUGHERTY, KF BOTT, PC HU, TS LUCIER, SN PETERSON, HO SMITH, CA HUTCHISON, and JC VENTER (1995). **The Minimal Gene Complement of Mycoplasma genitalium.** *Science* 270.5235, pp. 397–404. DOI: 10.1126/science.270.5235.397 (cit. on p. 10).
- GABALDÓN, T and EV KOONIN (2013). **Functional and evolutionary implications of gene orthology.** *Nature Reviews Genetics* 14.5, pp. 360–366. DOI: 10.1038/nrg3456 (cit. on p. 8).
- GELFMAN, S, D BURSTEIN, O PENN, A SAVCHENKO, M AMIT, S SCHWARTZ, T PUPKO, and G AST (2012). **Changes in exon–intron structure during vertebrate evolution affect the splicing pattern of exons.** *Genome Research* 22.1, pp. 35–50. DOI: 10.1101/gr.119834.110 (cit. on p. 20).
- GILBERT, W (1978). **Why genes in pieces?** *Nature* 271.5645, p. 501. DOI: 10.1038/271501a0 (cit. on p. 4).
- GOFFEAU, A, BG BARRELL, H BUSSEY, RW DAVIS, B DUJON, H FELDMANN, F GALIBERT, JD HOHEISEL, C JACQ, M JOHNSTON, EJ LOUIS, HW MEWES, Y MURAKAMI, P PHILIPPSEN, H TETTELIN, and SG OLIVER (1996). **Life with 6000 Genes.** *Science* 274.5287, pp. 546–567. DOI: 10.1126/science.274.5287.546 (cit. on p. 10).
- GOLDBACH, R (1987). **Genome similarities between plant and animal RNA viruses.** *Microbiological sciences* 4.7, pp. 197–202 (cit. on p. 11).
- GRAUR, D (2015). *Molecular and Genome Evolution*. 1st ed. 2090. Sinauer. 500 pp. (cit. on p. 18).
- GRAY, GS and WM FITCH (1983). **Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus.** *Molecular Biology and Evolution* 1.1, pp. 57–66 (cit. on p. 8).
- GREGORY, TR and DP SHORTHOUSE (2003). **Genome Sizes of Spiders.** *Journal of Heredity* 94.4, pp. 285–290. DOI: 10.1093/jhered/esg070 (cit. on p. 16).
- GREGORY, TR (2002a). **A Bird's-Eye View of the C-Value Enigma: Genome Size, Cell Size, and Metabolic Rate in the Class Aves.** *Evolution* 56.1, pp. 121–130. DOI: 10.1111/j.0014-3820.2002.tb00854.x (cit. on p. 16).
- (2002b). **Genome size and developmental complexity.** *Genetica* 115.1, pp. 131–146. DOI: 10.1023/A:1016032400147 (cit. on p. 16).
- (2005a). **Synergy between sequence and size in Large-scale genomics.** *Nature Reviews Genetics* 6.9, pp. 699–708. DOI: 10.1038/nrg1674 (cit. on p. 17).

- (2005b). **The C-value Enigma in Plants and Animals: A Review of Parallels and an Appeal for Partnership.** *Annals of Botany* 95.1, pp. 133–146. DOI: 10.1093/aob/mci009 (cit. on p. 16).
- GUIGÓ, R, P FLICEK, JF ABRIL, A REYMOND, J LAGARDE, F DENOEUDE, S ANTONARAKIS, M ASHBURNER, VB BAJIC, E BIRNEY, R CASTELO, E EYRAS, C UCLA, TR GINGERAS, J HARROW, T HUBBARD, SE LEWIS, and MG REESE (2006). **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biology* 7.1, S2. DOI: 10.1186/gb-2006-7-s1-s2 (cit. on p. 7).
- GUSEV, O, Y SUETSUGU, R CORNETTE, T KAWASHIMA, MD LOGACHEVA, AS KONDRASHOV, AA PENIN, R HATANAKA, S KIKUTA, S SHIMURA, H KANAMORI, Y KATAYOSE, T MATSUMOTO, E SHAGIMARDANOVA, D ALEXEEV, V GOVORUN, J WISECAVER, A MIKHEYEV, R KOYANAGI, M FUJIE, T NISHIYAMA, S SHIGENOBU, TF SHIBATA, V GOLYGINA, M HASEBE, T OKUDA, N SATOH, and T KIKAWADA (2014). **Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge.** *Nature Communications* 5. DOI: 10.1038/ncomms5784 (cit. on p. 24).
- HAHN, MW, TD BIE, JE STAJICH, C NGUYEN, and N CRISTIANINI (2005). **Estimating the tempo and mode of gene family evolution from comparative genomic data.** *Genome Research* 15.8, pp. 1153–1160. DOI: 10.1101/gr.3567505 (cit. on pp. 23, 24).
- HAHN, MW, JP DEMUTH, and SG HAN (2007). **Accelerated Rate of Gene Gain and Loss in Primates.** *Genetics* 177.3, pp. 1941–1949. DOI: 10.1534/genetics.107.080077 (cit. on pp. 23, 24).
- HAHN, MW, MV HAN, and SG HAN (2007). **Gene Family Evolution across 12 Drosophila Genomes.** *PLoS Genet* 3.11, e197. DOI: 10.1371/journal.pgen.0030197 (cit. on pp. 19, 21–24).
- HAN, K, Zf LI, R PENG, Lp ZHU, T ZHOU, Lg WANG, Sg LI, Xb ZHANG, W HU, Zh WU, N QIN, and Yz LI (2013). **Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu.** *Scientific Reports* 3, p. 2101. DOI: 10.1038/srep02101 (cit. on p. 17).
- HARDIE, DC and PD HEBERT (2004). **Genome-size evolution in fishes.** *Canadian Journal of Fisheries and Aquatic Sciences* 61.9, pp. 1636–1646. DOI: 10.1139/f04-106 (cit. on p. 16).
- HEDGES, SB and S KUMAR (2002). **Vertebrate Genomes Compared.** *Science* 297.5585, pp. 1283–1285 (cit. on p. 12).
- HENRICSON, A, K FORSLUND, and E SONNHAMMER (2010). **Orthology confers intron position conservation.** *BMC Genomics* 11.1, p. 412 (cit. on p. 8).
- HICKEY, AJR and KD CLEMENTS (2005). **Genome Size Evolution in New Zealand Triplefin Fishes.** *Journal of Heredity* 96.4, pp. 356–362. DOI: 10.1093/jhered/esi061 (cit. on p. 16).
- HOLLAND, PWH, F MARLÉTAZ, I MAESO, TL DUNWELL, and J PAPS (2017). **New genes from old: asymmetric divergence of gene duplicates and the evolution of**

- development.** *Phil. Trans. R. Soc. B* 372.1713, p. 20150480. DOI: 10.1098/rstb.2015.0480 (cit. on p. 22).
- HOWE, K, PH SCHIFFER, J ZIELINSKI, T WIEHE, GK LAIRD, JC MARIONI, O SOYLEMEZ, F KONDRASHOV, and M LEPTIN (2016). **Structure and evolutionary history of a large family of NLR proteins in the zebrafish.** *Open Biology* 6.4, p. 160009. DOI: 10.1098/rsob.160009 (cit. on p. 23).
- HUANG, S, J DING, D DENG, W TANG, H SUN, D LIU, L ZHANG, X NIU, X ZHANG, M MENG, J YU, J LIU, Y HAN, W SHI, D ZHANG, S CAO, Z WEI, Y CUI, Y XIA, H ZENG, K BAO, L LIN, Y MIN, H ZHANG, M MIAO, X TANG, Y ZHU, Y SUI, G LI, H SUN, J YUE, J SUN, F LIU, L ZHOU, L LEI, X ZHENG, M LIU, L HUANG, J SONG, C XU, J LI, K YE, S ZHONG, BR LU, G HE, F XIAO, HL WANG, H ZHENG, Z FEI, and Y LIU (2013). **Draft genome of the kiwifruit *Actinidia chinensis*.** *Nature Communications* 4. DOI: 10.1038/ncomms3640 (cit. on p. 19).
- HUGHES, AL and R FRIEDMAN (2008). **Genome Size Reduction in the Chicken Has Involved Massive Loss of Ancestral Protein-Coding Genes.** *Molecular Biology and Evolution* 25.12, pp. 2681–2688. DOI: 10.1093/molbev/msn207 (cit. on p. 18).
- I5K CONSORTIUM (2013). **The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.** *Journal of Heredity* 104.5, pp. 595–600. DOI: 10.1093/jhered/est050 (cit. on p. 12).
- JIA, J, S ZHAO, X KONG, Y LI, G ZHAO, W HE, R APPELS, M PFEIFER, Y TAO, X ZHANG, R JING, C ZHANG, Y MA, L GAO, C GAO, M SPANNAGL, KFX MAYER, D LI, S PAN, F ZHENG, Q HU, X XIA, J LI, Q LIANG, J CHEN, T WICKER, C GOU, H KUANG, G HE, Y LUO, B KELLER, Q XIA, P LU, J WANG, H ZOU, R ZHANG, J XU, J GAO, C MIDDLETON, Z QUAN, G LIU, J WANG, H YANG, X LIU, Z HE, L MAO, and J WANG (2013). ***Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation.** *Nature* 496.7443, pp. 91–95. DOI: 10.1038/nature12028 (cit. on p. 18).
- JIANG, X and R ASSIS (2017). **Natural Selection Drives Rapid Functional Evolution of Young *Drosophila* Duplicate Genes.** *Molecular Biology and Evolution* 34.12, pp. 3089–3098. DOI: 10.1093/molbev/msx230 (cit. on p. 22).
- JO, BS and SS CHOI (2015). **Introns: The Functional Benefits of Introns in Genomes, Introns: The Functional Benefits of Introns in Genomes.** *Genomics & Informatics, Genomics & Informatics* 13.4, pp. 112–118. DOI: 10.5808/GI.2015.13.4.112 (cit. on p. 21).
- JOHANNSEN, W (1909). *Elemente der exakten Erblchkeitslehre.* Google-Books-ID: \_4mFB-wAAQBAJ. Gustav Fischer. 529 pp. (cit. on p. 4).
- JORDAN, IK, Y WOLF, and E KOONIN (2004). **Duplicated genes evolve slower than singletons despite the initial rate increase.** *BMC Evolutionary Biology* 4.1, p. 22 (cit. on p. 22).
- KAPUSTA, A, A SUH, and C FESCHOTTE (2017). **Dynamics of genome size evolution in birds and mammals.** *Proceedings of the National Academy of Sciences*, p. 201616702. DOI: 10.1073/pnas.1616702114 (cit. on p. 17).

- KELLEY, JL, AP BROWN, NO THERKILDSEN, and AD FOOTE (2016). **The life aquatic: advances in marine vertebrate genomics.** *Nature Reviews. Genetics* 17.9, pp. 523–534. DOI: 10.1038/nrg.2016.66 (cit. on p. 12).
- KELLEY, JL, JT PEYTON, AS FISTON-LAVIER, NM TEETS, MC YEE, JS JOHNSTON, CD BUSTAMANTE, RE LEE, and DL DENLINGER (2014). **Compact genome of the Antarctic midge is likely an adaptation to an extreme environment.** *Nature Communications* 5. DOI: 10.1038/ncomms5611 (cit. on p. 17).
- KHALTURIN, K, G HEMMRICH, S FRAUNE, R AUGUSTIN, and TC BOSCH (2009). **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends in Genetics* 25.9, pp. 404–413. DOI: 10.1016/j.tig.2009.07.006 (cit. on pp. 19, 23).
- KJER, KM, JL WARE, J RUST, T WAPPLER, R LANFEAR, LS JERMIIN, X ZHOU, H ASPÖCK, U ASPÖCK, RG BEUTEL, A BLANKE, A DONATH, T FLOURI, PB FRANDSEN, P KAPLI, AY KAWAHARA, H LETSCH, C MAYER, DD MCKENNA, K MEUSEMANN, O NIEHUIS, RS PETERS, BM WIEGMANN, DK YEATES, BMv REUMONT, A STAMATAKIS, and B MISOF (2015). **Response to Comment on “Phylogenomics resolves the timing and pattern of insect evolution”.** *Science* 349.6247, pp. 487–487. DOI: 10.1126/science.aaa7136 (cit. on p. 27).
- KJER, KM, C SIMON, M YAVORSKAYA, and RG BEUTEL (2016). **Progress, pitfalls and parallel universes: a history of insect phylogenetics.** *Journal of The Royal Society Interface* 13.121, p. 20160363. DOI: 10.1098/rsif.2016.0363 (cit. on p. 13).
- KLENK, HP, RA CLAYTON, JF TOMB, O WHITE, KE NELSON, KA KETCHUM, RJ DODSON, M GWINN, EK HICKEY, JD PETERSON, DL RICHARDSON, AR KERLAVAGE, DE GRAHAM, NC KYRPIDES, RD FLEISCHMANN, J QUACKENBUSH, NH LEE, GG SUTTON, S GILL, EF KIRKNESS, BA DOUGHERTY, K MCKENNEY, MD ADAMS, B LOFTUS, S PETERSON, CI REICH, LK MCNEIL, JH BADGER, A GLODEK, L ZHOU, R OVERBEEK, JD GOCAYNE, JF WEIDMAN, L McDONALD, T UTTERBACK, MD COTTON, T SPRIGGS, P ARTIACH, BP KAINE, SM SYKES, PW SADOW, KP D’ANDREA, C BOWMAN, C FUJII, SA GARLAND, TM MASON, GJ OLSEN, CM FRASER, HO SMITH, CR WOESE, and JC VENTER (1997). **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 390.6658, pp. 364–370. DOI: 10.1038/37052 (cit. on p. 10).
- KNOWLES, DG and A MCLYSAGHT (2006). **High Rate of Recent Intron Gain and Loss in Simultaneously Duplicated Arabidopsis Genes.** *Molecular Biology and Evolution* 23.8, pp. 1548–1557. DOI: 10.1093/molbev/msl017 (cit. on p. 22).
- KONDRASHOV, FA, IB ROGOZIN, YI WOLF, and EV KOONIN (2002). **Selection in the evolution of gene duplications.** *Genome Biol* 3.2, pp. 8–1 (cit. on pp. 22, 23).
- KOONIN, EV (2005). **Orthologs, Paralogs and Evolutionary Genomics.** *Annual Review of Genetics* 39, pp. 309–338 (cit. on p. 8).
- KOONIN, EV (2009). **Darwinian evolution in the light of genomics.** *Nucleic Acids Research* 37.4, pp. 1011–1034. DOI: 10.1093/nar/gkp089 (cit. on pp. 8, 12).

- KOZAK, M (1991). **An analysis of vertebrate mRNA sequences: intimations of translational control.** *The Journal of Cell Biology* 115.4, pp. 887–903. DOI: 10.1083/jcb.115.4.887 (cit. on p. 7).
- KUZNIAR, A, RC van HAM, S PONGOR, and JA LEUNISSEN (2008). **The quest for orthologs: finding the corresponding gene across genomes.** *Trends in Genetics* 24.11, pp. 539–551. DOI: 10.1016/j.tig.2008.08.009 (cit. on p. 8).
- LAND, M, L HAUSER, SR JUN, I NOOKAEW, MR LEUZE, TH AHN, T KARPINETS, O LUND, G KORA, T WASSENAAR, S POUDEL, and DW USSERY (2015). **Insights from 20 years of bacterial genome sequencing.** *Functional & Integrative Genomics* 15.2, pp. 141–161. DOI: 10.1007/s10142-015-0433-4 (cit. on p. 12).
- LANDER, ES *et al.* (2001). **Initial sequencing and analysis of the human genome.** *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062 (cit. on p. 10).
- LEDERBERG, J and AT MCCRAY (2001). **'Ome Sweet 'Omics– A Genealogical Treasury of Words.** *The Scientist* 15.7, p. 8 (cit. on p. 11).
- LEFÉBURE, T, C MORVAN, F MALARD, C FRANÇOIS, L KONECNY-DUPRÉ, L GUÉGUEN, M WEISS-GAYET, A SEGUIN-ORLANDO, L ERMINI, CD SARKISSIAN, NP CHARRIER, D EME, F MERMILLOD-BLONDIN, L DURET, C VIEIRA, L ORLANDO, and C DOUADY (2017). **Less effective selection leads to larger genomes.** *Genome Research*, gr.212589.116. DOI: 10.1101/gr.212589.116 (cit. on pp. 16, 17).
- LESPINET, O, YI WOLF, EV KOONIN, and L ARAVIND (2002). **The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes.** *Genome Research* 12.7, pp. 1048–1059. DOI: 10.1101/gr.174302 (cit. on p. 23).
- LI, SW, L FENG, and DK NIU (2007). **Selection for the miniaturization of highly expressed genes.** *Biochemical and Biophysical Research Communications* 360.3, pp. 586–592. DOI: 10.1016/j.bbrc.2007.06.085 (cit. on p. 20).
- LYNCH, M and JS CONERY (2003). **“The evolutionary demography of duplicate genes”.** *Genome Evolution*. Springer, Dordrecht, pp. 35–44. DOI: 10.1007/978-94-010-0263-9\_4 (cit. on p. 23).
- MAHNER, M and M KARY (1997). **What Exactly Are Genomes, Genotypes and Phenotypes? And What About Phenomes?** *Journal of Theoretical Biology* 186.1, pp. 55–63. DOI: 10.1006/jtbi.1996.0335 (cit. on p. 4).
- MALMSTROM, M, R BRITZ, M MATSCHINER, OK TORRESEN, RK HADIATY, N YAAKOB, HH TAN, KS JAKOBSEN, W SALZBURGER, and L RUBER (2017). **The most developmentally truncated fishes show extensive Hox gene loss and miniaturized genomes.** *bioRxiv*, p. 160168. DOI: 10.1101/160168 (cit. on pp. 17, 18).
- MCCOY, RC, RW TAYLOR, TA BLAUWKAMP, JL KELLEY, M KERTESZ, D PUSHKAREV, DA PETROV, and AS FISTON-LAVIER (2014). **Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements.** *PLoS ONE* 9.9. Ed. by N SINGH, e106689. DOI: 10.1371/journal.pone.0106689 (cit. on p. 5).
- MIKKELSEN, TS (2004). **What makes us human?** *Genome Biology* 5.8, p. 238. DOI: 10.1186/gb-2004-5-8-238 (cit. on p. 18).



- MISOF, B *et al.* (2014). **Phylogenomics resolves the timing and pattern of insect evolution.** *Science* 346.6210, pp. 763–767. DOI: 10.1126/science.1257570 (cit. on p. 27).
- MISRA, S, MA CROSBY, CJ MUNGALL, BB MATTHEWS, KS CAMPBELL, P HRADECKY, Y HUANG, JS KAMINKER, GH MILLBURN, SE PROCHNIK, CD SMITH, JL TUPY, EJ WHITFIELD, L BAYRAKTAROGLU, BP BERMAN, BR BETTENCOURT, SE CELNIKER, AD de GREY, RA DRYSDALE, NL HARRIS, J RICHTER, S RUSSO, AJ SCHROEDER, S SHU, M STAPLETON, C YAMADA, M ASHBURNER, WM GELBART, GM RUBIN, and SE LEWIS (2002). **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biology* 3.12, research0083.1–83.22. DOI: 10.1186/gb-2002-3-12-research0083 (cit. on p. 7).
- MURPHY, WJ, PA PEVZNER, and SJ O'BRIEN (2004). **Mammalian phylogenomics comes of age.** *Trends in Genetics* 20.12, pp. 631–639. DOI: 10.1016/j.tig.2004.09.005 (cit. on p. 12).
- NEAFSEY, DE and SR PALUMBI (2003). **Genome Size Evolution in Pufferfish: A Comparative Analysis of Diodontid and Tetraodontid Pufferfish Genomes.** *Genome Research* 13.5, pp. 821–830. DOI: 10.1101/gr.841703 (cit. on pp. 17, 18).
- NEAFSEY, DE *et al.* (2015). **Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes.** *Science* 347.6217, p. 1258522. DOI: 10.1126/science.1258522 (cit. on p. 24).
- NEHRT, NL, WT CLARK, P RADIVOJAC, and MW HAHN (2011). **Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals.** *PLoS Computational Biology* 7.6. Ed. by A RZHETSKY, e1002073. DOI: 10.1371/journal.pcbi.1002073 (cit. on p. 8).
- NEMBAWARE, V, K CRUM, J KELSO, and C SEOIGHE (2002). **Impact of the Presence of Paralogs on Sequence Divergence in a Set of Mouse-Human Orthologs.** *Genome Research* 12.9, pp. 1370–1376. DOI: 10.1101/gr.270902 (cit. on p. 22).
- NEWTON, LE (2007). **What Makes Us Human?** *Bioscience Reports* 27.4, pp. 185–187. DOI: 10.1007/s10540-007-9048-x (cit. on p. 18).
- NIEHUIS, O, JD GIBSON, MS ROSENBERG, BA PANNEBAKKER, T KOEVOETS, AK JUDSON, CA DESJARDINS, K KENNEDY, D DUGGAN, LW BEUKEBOOM, Lvd ZANDE, DM SHUKER, JH WERREN, and J GADAU (2010). **Recombination and Its Impact on the Genome of the Haplodiploid Parasitoid Wasp *Nasonia*.** *PLOS ONE* 5.1, e8597. DOI: 10.1371/journal.pone.0008597 (cit. on p. 21).
- NIERMAN, WC, JA EISEN, RD FLEISCHMANN, and CM FRASER (2000). **Genome data: what do we learn?** *Current Opinion in Structural Biology* 10.3, pp. 343–348. DOI: 10.1016/S0959-440X(00)00094-4 (cit. on p. 11).
- O'BRIEN, SJ and R STANYON (1999). **Phylogenomics: Ancestral primate viewed.** *Nature* 402.6760, pp. 365–366. DOI: 10.1038/46450 (cit. on p. 11).
- OLIVER, SG *et al.* (1992). **The complete DNA sequence of yeast chromosome III.** *Nature* 357.6373, pp. 38–46. DOI: 10.1038/357038a0 (cit. on pp. 10, 11).

- PELLICER, J, MF FAY, and IJ LEITCH (2010). **The largest eukaryotic genome of them all?** *Botanical Journal of the Linnean Society* 164.1, pp. 10–15. DOI: 10.1111/j.1095-8339.2010.01072.x (cit. on p. 16).
- PETERS, RS, L KROGMANN, C MAYER, A DONATH, S GUNKEL, K MEUSEMANN, A KOZLOV, L PODSIADLOWSKI, M PETERSEN, R LANFEAR, PA DIEZ, J HERATY, KM KJER, S KLOPFSTEIN, R MEIER, C POLIDORI, T SCHMITT, S LIU, X ZHOU, T WAPPLER, J RUST, B MISOF, and O NIEHUIS (2017). **Evolutionary History of the Hymenoptera.** *Current Biology* 27.7, pp. 1013–1018. DOI: 10.1016/j.cub.2017.01.027 (cit. on p. 27).
- PETERSON, ML, MB BRYMAN, M PEITER, and C COWAN (1994). **Exon size affects competition between splicing and cleavage-polyadenylation in the immunoglobulin mu gene.** *Molecular and Cellular Biology* 14.1, pp. 77–86. DOI: 10.1128/MCB.14.1.77 (cit. on p. 20).
- PETROV, DA (2001). **Evolution of genome size: new approaches to an old problem.** *TRENDS in Genetics* 17.1, pp. 23–28 (cit. on p. 16).
- PETROV, DA, TA SANGSTER, JS JOHNSTON, DL HARTL, and KL SHAW (2000). **Evidence for DNA Loss as a Determinant of Genome Size.** *Science* 287.5455. ArticleType: research-article / Full publication date: Feb. 11, 2000 / Copyright © 2000 American Association for the Advancement of Science, pp. 1060–1062. DOI: 10.2307/3074222 (cit. on pp. 16, 17).
- PEVSNER, J (2009). *Bioinformatics and Functional Genomics*. 2nd ed. Wiley-Blackwell. 992 pp. (cit. on p. 10).
- PHILIPS, S, HY WU, and L LI (2017). **Using machine learning algorithms to identify genes essential for cell survival.** *BMC Bioinformatics* 18.11, p. 397. DOI: 10.1186/s12859-017-1799-1 (cit. on p. 5).
- PINGAULT, L, F CHOLET, A ALBERTI, N GLOVER, P WINCKER, C FEUILLET, and E PAUX (2015). **Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome.** *Genome Biology* 16.1. DOI: 10.1186/s13059-015-0601-9 (cit. on p. 20).
- POLLARD, KS (2009). **What makes us human?** *Scientific American* 300.5, pp. 44–49 (cit. on p. 18).
- PRACHUMWAT, A, L DEVINCENTIS, and MF PALOPOLI (2004). **Intron Size Correlates Positively With Recombination Rate in *Caenorhabditis elegans*.** *Genetics* 166.3, pp. 1585–1590. DOI: 10.1534/genetics.166.3.1585 (cit. on p. 20).
- RAO, YS, ZF WANG, XW CHAI, GZ WU, M ZHOU, QH NIE, and XQ ZHANG (2010). **Selection for the compactness of highly expressed genes in *Gallus gallus*.** *Biology Direct* 5, p. 35. DOI: 10.1186/1745-6150-5-35 (cit. on p. 20).
- RAPPOPORT, N and M LINIAL (2015). **Trends in genome dynamics among major orders of insects revealed through variations in protein families.** *BMC Genomics* 16.1, p. 583. DOI: 10.1186/s12864-015-1771-2 (cit. on p. 24).
- RICHARDS, S (2015). **It's More Than Stamp Collecting: How Genome Sequencing Can Unify Biological Research.** *Trends in genetics : TIG* 31.7, pp. 411–421. DOI: 10.1016/j.tig.2015.04.007 (cit. on p. 13).

- RÖDELSPERGER, C, A STREIT, and RJ SOMMER (2013). "Structure, Function and Evolution of The Nematode Genome". *eLS*. Ed. by JOHN WILEY & SONS, LTD. John Wiley & Sons, Ltd (cit. on p. 19).
- ROGOZIN, IB (2014). **Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing**. *Genetics Research International* 2014, e516508. DOI: 10.1155/2014/516508 (cit. on p. 12).
- ROGOZIN, IB, L CARMEL, M CSUROS, and EV KOONIN (2012). **Origin and evolution of spliceosomal introns**. *Biol Direct* 7.11, pp. 6150–7 (cit. on p. 21).
- ROGOZIN, IB and IB ROGOZIN (2014). **Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing, Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing**. *Genetics Research International, Genetics Research International* 2014, 2014, e516508. DOI: 10.1155/2014/516508, 10.1155/2014/516508 (cit. on p. 8).
- ROUX, J, E PRIVMAN, S MORETTI, JT DAUB, M ROBINSON-RECHAVI, and L KELLER (2014). **Patterns of Positive Selection in Seven Ant Genomes**. *Molecular Biology and Evolution*, msu141. DOI: 10.1093/molbev/msu141 (cit. on p. 24).
- ROY, M, N KIM, Y XING, and C LEE (2008). **The effect of intron length on exon creation ratios during the evolution of mammalian genomes**. *RNA* 14.11, pp. 2261–2273. DOI: 10.1261/rna.1024908 (cit. on p. 20).
- ROY, SW and D PENNY (2007). **Patterns of Intron Loss and Gain in Plants: Intron Loss-Dominated Evolution and Genome-Wide Comparison of *O. sativa* and *A. thaliana***. *Molecular Biology and Evolution* 24.1, pp. 171–181. DOI: 10.1093/molbev/msl159 (cit. on p. 21).
- RUBIN, GM, MD YANDELL, JR WORTMAN, GL GABOR, MIKLOS, CR NELSON, IK HARIHARAN, ME FORTINI, PW LI, R APWEILER, W FLEISCHMANN, JM CHERRY, S HENIKOFF, MP SKUPSKI, S MISRA, M ASHBURNER, E BIRNEY, MS BOGUSKI, T BRODY, P BROKSTEIN, SE CELNIKER, SA CHERVITZ, D COATES, A CRAVCHIK, A GABRIELIAN, RF GALLE, WM GELBART, RA GEORGE, LSB GOLDSTEIN, F GONG, P GUAN, NL HARRIS, BA HAY, RA HOSKINS, J LI, Z LI, RO HYNES, SJM JONES, PM KUEHL, B LEMAITRE, JT LITTLETON, DK MORRISON, C MUNGALL, PH O'FARRELL, OK PICKERAL, C SHUE, LB VOSSHALL, J ZHANG, Q ZHAO, XH ZHENG, F ZHONG, W ZHONG, R GIBBS, JC VENTER, MD ADAMS, and S LEWIS (2000). **Comparative Genomics of the Eukaryotes**. *Science* 287.5461, pp. 2204–2215. DOI: 10.1126/science.287.5461.2204 (cit. on p. 11).
- RYABOV, Y and M GRIBSKOV (2008). **Spontaneous symmetry breaking in genome evolution**. *Nucleic Acids Research* 36.8, pp. 2756–2763. DOI: 10.1093/nar/gkn086 (cit. on p. 20).
- SANN, M, O NIEHUIS, RS PETERS, C MAYER, A KOZLOV, L PODSIADLOWSKI, S BANK, K MEUSEMANN, B MISOF, C BLEIDORN, and M OHL (2018). **Phylogenomic analysis of Apoidea sheds new light on the sister group of bees**. *BMC Evolutionary Biology* 18.1. DOI: 10.1186/s12862-018-1155-8 (cit. on p. 27).

- SCHIFFER, PH, J GRAVEMEYER, M RAUSCHER, and T WIEHE (2016). **Ultra Large Gene Families: A Matter of Adaptation or Genomic Parasites?** *Life* 6.3. DOI: 10.3390/life6030032 (cit. on p. 24).
- SHINOZAKI, K, M OHME, M TANAKA, T WAKASUGI, N HAYASHIDA, T MATSUBAYASHI, N ZAITA, J CHUNWONGSE, J OBOKATA, K YAMAGUCHI-SHINOZAKI, C OHTO, K TORAZAWA, BY MENG, M SUGITA, H DENO, T KAMOGASHIRA, K YAMADA, J KUSUDA, F TAKAIWA, A KATO, N TOHDOH, H SHIMADA, and M SUGIURA (1986). **The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression.** *The EMBO Journal* 5.9, pp. 2043–2049. DOI: 10.1002/j.1460-2075.1986.tb04464.x (cit. on p. 10).
- SMITH, DR, LA DOUCETTE-STAMM, C DELOUGHERY, H LEE, J DUBOIS, T ALDREDGE, R BASHIRZADEH, D BLAKELY, R COOK, K GILBERT, D HARRISON, L HOANG, P KEAGLE, W LUMM, B POTHIER, D QIU, R SPADAFORA, R VICAIRE, Y WANG, J WIERZBOWSKI, R GIBSON, N JIWANI, A CARUSO, D BUSH, and JN REEVE (1997). **Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics.** *Journal of Bacteriology* 179.22, pp. 7135–7155. DOI: 10.1128/jb.179.22.7135-7155.1997 (cit. on p. 10).
- SMITH, EM and TR GREGORY (2009). **Patterns of genome size diversity in the ray-finned fishes.** *Hydrobiologia* 625.1, pp. 1–25. DOI: 10.1007/s10750-009-9724-x (cit. on p. 16).
- SNYDER, M and M GERSTEIN (2003). **Defining genes in the genomics era.** *Science* 300.5617, pp. 258–260 (cit. on p. 4).
- STEIN, L (2001). **Genome annotation: from sequence to biology.** *Nature Reviews Genetics* 2.7, pp. 493–503. DOI: 10.1038/35080529 (cit. on p. 5).
- STUDER, RA and M ROBINSON-RECHAVI (2009). **How confident can we be that orthologs are similar, but paralogs differ?** *Trends in Genetics* 25.5, pp. 210–216. DOI: 10.1016/j.tig.2009.03.004 (cit. on p. 8).
- SUN, C, JRL ARRIAZA, and RL MUELLER (2012). **Slow DNA Loss in the Gigantic Genomes of Salamanders.** *Genome Biology and Evolution* 4.12, pp. 1340–1348. DOI: 10.1093/gbe/evs103 (cit. on p. 17).
- TAUTZ, D and T DOMAZET-LOŠO (2011). **The evolutionary origin of orphan genes.** *Nature Reviews Genetics* 12.10, pp. 692–702. DOI: 10.1038/nrg3053 (cit. on p. 19).
- THE ARABIDOPSIS GENOME INITIATIVE (2000). **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 408.6814, pp. 796–815. DOI: 10.1038/35048692 (cit. on p. 10).
- THE C. ELEGANS SEQUENCING CONSORTIUM (1998). **Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology.** *Science* 282.5396, pp. 2012–2018 (cit. on pp. 10, 11).
- THE MOUSE GENOME SEQUENCING CONSORTIUM (2002). **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 420.6915, pp. 520–562. DOI: 10.1038/nature01262 (cit. on p. 10).

- THOMAS, CA (1971). **The Genetic Organization of Chromosomes.** *Annual Review of Genetics* 5.1, pp. 237–256. DOI: 10.1146/annurev.ge.05.120171.001321 (cit. on p. 16).
- THOMAS, PD, V WOOD, CJ MUNGALL, SE LEWIS, JA BLAKE, and ON BEHALF OF THE GENE ONTOLOGY CONSORTIUM (2012). **On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report.** *PLoS Computational Biology* 8.2. Ed. by PE BOURNE, e1002386. DOI: 10.1371/journal.pcbi.1002386 (cit. on p. 8).
- VENTER, JC *et al.* (2001). **The Sequence of the Human Genome.** *Science* 291.5507, pp. 1304–1351. DOI: 10.1126/science.1058040 (cit. on pp. 10, 11).
- VIDAL, NM, AL GRAZZIOTIN, LM IYER, L ARAVIND, and TM VENANCIO (2016). **Transcription factors, chromatin proteins and the diversification of Hemiptera.** *Insect biochemistry and molecular biology* 69, pp. 1–13. DOI: 10.1016/j.ibmb.2015.07.001 (cit. on p. 23).
- VINOGRADOV, AE (1995). **Nucleotypic Effect in Homeotherms: Body-Mass-Corrected Basal Metabolic Rate of Mammals Is Related to Genome Size.** *Evolution* 49.6, pp. 1249–1259. DOI: 10.1111/j.1558-5646.1995.tb04451.x (cit. on p. 16).
- VINOGRADOV, AE (2004). **Evolution of genome size: multilevel selection, mutation bias or dynamical chaos?** *Current Opinion in Genetics & Development* 14.6, pp. 620–626. DOI: 10.1016/j.gde.2004.09.007 (cit. on p. 17).
- VOGEL, C and C CHOTHIA (2006). **Protein Family Expansions and Biological Complexity.** *PLoS Computational Biology* 2.5. DOI: 10.1371/journal.pcbi.0020048 (cit. on p. 23).
- WANG, H, D NETTLETON, and K YING (2014). **Copy number variation detection using next generation sequencing read counts.** *BMC Bioinformatics* 15.1, p. 109. DOI: 10.1186/1471-2105-15-109 (cit. on p. 16).
- WANG, X, X FANG, P YANG, X JIANG, F JIANG, D ZHAO, B LI, F CUI, J WEI, C MA, Y WANG, J HE, Y LUO, Z WANG, X GUO, W GUO, X WANG, Y ZHANG, M YANG, S HAO, B CHEN, Z MA, D YU, Z XIONG, Y ZHU, D FAN, L HAN, B WANG, Y CHEN, J WANG, L YANG, W ZHAO, Y FENG, G CHEN, J LIAN, Q LI, Z HUANG, X YAO, N LV, G ZHANG, Y LI, J WANG, J WANG, B ZHU, and L KANG (2014). **The locust genome provides insight into swarm formation and long-distance flight.** *Nature Communications* 5. DOI: 10.1038/ncomms3957 (cit. on pp. 17, 20).
- WANG, Z, D ZARLENGA, J MARTIN, S ABUBUCKER, and M MITREVA (2012). **Exploring metazoan evolution through dynamic and holistic changes in protein families and domains.** *BMC Evolutionary Biology* 12, p. 138. DOI: 10.1186/1471-2148-12-138 (cit. on p. 23).
- WATERHOUSE, RM (2015). **A maturing understanding of the composition of the insect gene repertoire.** *Current Opinion in Insect Science* 7. DOI: 10.1016/j.cois.2015.01.004 (cit. on p. 19).
- WATERHOUSE, RM, EM ZDOBNOV, and EV KRIVENTSEVA (2011). **Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in**

- Vertebrates, Arthropods, and Fungi.** *Genome Biology and Evolution* 3, pp. 75–86. DOI: 10.1093/gbe/evq083 (cit. on p. 22).
- WATSON, JD and FH CRICK (1953). **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 171.4356, pp. 737–738 (cit. on p. 4).
- WENDEL, JF, RC CRONN, I ALVAREZ, B LIU, RL SMALL, and DS SENCHINA (2002). **Intron Size and Genome Size in Plants.** *Molecular Biology and Evolution* 19.12, pp. 2346–2352 (cit. on p. 17).
- WENDEL, JF, SA JACKSON, BC MEYERS, and RA WING (2016). **Evolution of plant genome architecture.** *Genome Biology* 17, p. 37. DOI: 10.1186/s13059-016-0908-1 (cit. on p. 12).
- WILBRANDT, J, B MISOF, and O NIEHUIS (2017). **COGNATE: comparative gene annotation characterizer.** *BMC Genomics* 18.1, p. 535. DOI: 10.1186/s12864-017-3870-8 (cit. on p. 4).
- WILSON, EB (1925). *The Cell In Development And Heredity* (cit. on p. 4).
- WINKLER, H (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche.* In collab. with MBLWHOI LIBRARY. Jena : G. Filscher. 252 pp. (cit. on p. 4).
- WISSLER, L, J GADAU, DF SIMOLA, M HELMKAMPF, and E BORNBERG-BAUER (2013). **Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes.** *Genome Biology and Evolution* 5.2, pp. 439–455. DOI: 10.1093/gbe/evt009 (cit. on pp. 19, 24).
- WOODY, JL and RC SHOEMAKER (2011). **Gene Expression: Sizing It All Up.** *Frontiers in Genetics* 2. DOI: 10.3389/fgene.2011.00070 (cit. on p. 20).
- WRIGHT, NA, TR GREGORY, and CC WITT (2014). **Metabolic ‘engines’ of flight drive genome size reduction in birds.** *Proceedings of the Royal Society of London B: Biological Sciences* 281.1779, p. 20132780. DOI: 10.1098/rspb.2013.2780 (cit. on p. 16).
- WURM, Y (2015). **Arthropod genomics beyond fruit flies: bridging the gap between proximate and ultimate causation.** *Briefings in Functional Genomics* 14.6, pp. 381–383. DOI: 10.1093/bfpg/ev034 (cit. on p. 12).
- YANDELL, M, CJ MUNGALL, C SMITH, S PROCHNIK, J KAMINKER, G HARTZELL, S LEWIS, and GM RUBIN (2006). **Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes.** *PLoS Comput Biol* 2.3, e15. DOI: 10.1371/journal.pcbi.0020015 (cit. on pp. 20, 21).
- YANG, YF, T ZHU, and DK NIU (2013). **Association of Intron Loss with High Mutation Rate in Arabidopsis: Implications for Genome Size Evolution.** *Genome Biology and Evolution* 5.4, pp. 723–733. DOI: 10.1093/gbe/evt043 (cit. on p. 16).
- YANG, S, JR ARGUELLO, X LI, Y DING, Q ZHOU, Y CHEN, Y ZHANG, R ZHAO, F BRUNET, L PENG, M LONG, and W WANG (2008). **Repetitive Element-Mediated Recombination as a Mechanism for New Gene Origination in Drosophila.** *PLoS Genet* 4.1, e3. DOI: 10.1371/journal.pgen.0040003 (cit. on p. 24).
- ZACHARIAS, H, B ANOKHIN, K KHALTURIN, and TC BOSCH (2004). **Genome sizes and chromosomes in the basal metazoan Hydra.** *Zoology* 107.3, pp. 219–227. DOI: 10.1016/j.zool.2004.04.005 (cit. on p. 16).

- ZDOBNOV, EM and P BORK (2007). **Quantification of insect genome divergence.** *Trends in Genetics* 23.1, pp. 16–20. DOI: 10.1016/j.tig.2006.10.004 (cit. on p. 24).
- ZDOBNOV, EM, Cv MERING, I LETUNIC, D TORRENTS, M SUYAMA, RR COPLEY, GK CHRISTOPHIDES, D THOMASOVA, RA HOLT, GM SUBRAMANIAN, HM MUELLER, G DIMOPOULOS, JH LAW, MA WELLS, E BIRNEY, R CHARLAB, AL HALPERN, E KOKOZA, CL KRAFT, Z LAI, S LEWIS, C LOUIS, C BARILLAS-MURY, D NUSSKERN, GM RUBIN, SL SALZBERG, GG SUTTON, P TOPALIS, R WIDES, P WINCKER, M YANDELL, FH COLLINS, J RIBEIRO, WM GELBART, FC KAFATOS, and P BORK (2002). **Comparative Genome and Proteome Analysis of *Anopheles gambiae* and *Drosophila melanogaster*.** *Science* 298.5591, pp. 149–159. DOI: 10.1126/science.1077061 (cit. on p. 24).
- ZDOBNOV, EM, C von MERING, I LETUNIC, and P BORK (2005). **Consistency of genome-based methods in measuring Metazoan evolution.** *FEBS Letters* 579.15, pp. 3355–3361. DOI: 10.1016/j.febslet.2005.04.006 (cit. on p. 24).
- ZHANG, G, B LI, C LI, MT GILBERT, ED JARVIS, J WANG, and THE AVIAN GENOME CONSORTIUM (2014). **Comparative genomic data of the Avian Phylogenomics Project.** *GigaScience* 3.1, p. 26. DOI: 10.1186/2047-217X-3-26 (cit. on pp. 17, 18).
- ZHANG, MQ (1998). **Statistical Features of Human Exons and Their Flanking Regions.** *Human Molecular Genetics* 7.5, pp. 919–932. DOI: 10.1093/hmg/7.5.919 (cit. on p. 20).
- ZHANG, MQ (2002). **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 3.9, pp. 698–709. DOI: 10.1038/nrg890 (cit. on p. 5).
- ZHANG, Q and SV EDWARDS (2012). **The Evolution of Intron Size in Amniotes: A Role for Powered Flight?** *Genome Biology and Evolution* 4.10, pp. 1033–1043. DOI: 10.1093/gbe/evs070 (cit. on p. 16).
- ZHANG, SQ, LH CHE, Y LI, D LIANG, H PANG, A ŚLIPIŃSKI, and P ZHANG (2018). **Evolutionary history of Coleoptera revealed by extensive sampling of genes and species.** *Nature Communications* 9. DOI: 10.1038/s41467-017-02644-4 (cit. on p. 27).
- ZHAO, J, AI TEUFEL, DA LIBERLES, and L LIU (2015). **A generalized birth and death process for modeling the fates of gene duplication.** *BMC Evolutionary Biology* 15.1, p. 275. DOI: 10.1186/s12862-015-0539-2 (cit. on pp. 22, 23).





---

# COGNATE: COMPARATIVE GENE ANNOTATION CHARACTERIZER

---

The following text is the author's version (including minor edits like additional subheadings) of the published article:

**WILBRANDT J, MISOF B, NIEHUIS O (2017). COGNATE: comparative gene annotation characterizer. *BMC Genomics* 18:535.**

**DOI: 10.1186/s12864-017-3870-8**

Authors' contributions to the original article:

Software development, figures: JW; manuscript design and writing: JW, BM, ON.



---

## Abstract

---

**B**ACKGROUND: The comparison of gene and genome structures across species has the potential to reveal major trends of genome evolution. However, such a comparative approach is currently hampered by a lack of standardization (e.g., ELLIOTT and GREGORY, 2015). For example, testing the hypothesis that the total amount of coding sequences is a reliable measure of potential proteome diversity (WANG *et al.*, 2011) requires the application of standardized definitions of coding sequence and genes to create both comparable and comprehensive data sets and corresponding summary statistics. However, such standard definitions either do not exist or are not consistently applied. These circumstances call for a standard at the descriptive level using a minimum of parameters as well as an undeviating use of standardized terms, and for software that infers the required data under these strict definitions. The acquisition of a comprehensive, descriptive, and standardized set of parameters and summary statistics for genome publications and further analyses can thus greatly benefit from the availability of an easy to use standard tool.

**R**ESULTS: We developed a new open-source command-line tool, COGNATE (Comparative Gene Annotation Characterizer), which uses a given genome assembly and its annotation of protein-coding genes for a detailed description of the respective gene and genome structure parameters. Additionally, we revised the standard definitions of gene and genome structures and provide the definitions used by COGNATE as a working draft suggestion for further reference. Complete parameter lists and summary statistics are inferred using this set of definitions to allow down-stream analyses and to provide an overview of the genome and gene repertoire characteristics. COGNATE is written in Perl and freely available at the ZFMK homepage (<https://www.zfmk.de/en/COGNATE>) and on github (<https://github.com/ZFMK/COGNATE>).

**C**ONCLUSION: The tool COGNATE allows comparing genome assemblies and structural elements on multiple levels (*e.g.*, scaffold or contig sequence, gene). It clearly enhances comparability between analyses. Thus, COGNATE can provide the important standardization of both genome and gene structure parameter disclosure as well as data acquisition for future comparative analyses. With the establishment of comprehensive descriptive standards and the extensive availability of genomes, an encompassing database will become possible.

**K**EYWORDS: Comparative genomics, Protein-coding genes, Gene annotation, Gene repertoires, Gene structure, Standardization

---

## Introduction

---

**A**S MORE AND MORE sequenced genomes become available, studying the commonalities and differences in the structure of genes and genomes has become an exciting and a rapidly expanding research field. Examples of comparative studies of intron size are those published by YANDELL *et al.* (2006), MOSS *et al.* (2011), and ZIMMER *et al.* (2013), who found that intron length evolution behaves clocklike, that ancient bursts of repetitive elements can be responsible for an unusual intron length distribution, and that there is a trend towards shorter introns in the evolution of land plants, respectively. These studies were restricted to a rather unrepresentative selection of animal, fish, and plants species, respectively, due to the lack of genome sequences. Studies with much larger species numbers and a broader taxonomic coverage are becoming feasible.

### 2.1 A lack of standards in gene structure characterization

ELLIOTT and GREGORY (2015) recently published a seminal meta-analysis of the genome and gene summary statistics of animals, land plants, fungi, and

'protists', relying on 521 species. The large number of species and genomes considered in their analysis allowed the authors to robustly detect statistical trends in genome evolution, such as a positive correlation between genome size and both gene and intron content, while taking phylogenetic relationships into account. These trends have been previously observed (*e.g.*, HOU and LIN, 2009; Q ZHANG and EDWARDS, 2012), but were based on a much smaller taxonomic sampling. Yet, despite the evidently improved availability of sequenced genomes, ELLIOTT and GREGORY (2015) struggled with a lack of standards in the disclosure of genome characteristics when compiling data for their analyses; they evaluated 28 parameters of the genomes of 521 species (see Supplement of ELLIOTT and GREGORY, 2015), for which only 48 % of all possible values were provided in the publications to the respective genomes (*cf.* Fig. II.2) and thus available for the meta-analysis.

The lack of standardization in the publication of gene structure characteristics is a general problem. Not only are some basic gene content and structure statistics frequently presented in a non-standardized manner, it often remains unclear whether or not terms describing gene structure were consistently applied to achieve comparability between analyses. For example, gene counts may or may not be inferred from tallying all predicted transcripts, thus bearing the risk of including alternative transcripts or isoforms as pseudo-replicates in meta-analyses. Furthermore, GC content may be reckoned without respect to IUPAC base-calling ambiguity in the total sequence lengths, which predicates the resulting value on sequencing and assembly quality. Finally, it can be difficult to trace inconsistencies in the use of terms, like 'exon' versus 'coding sequence (CDS)' despite existing standard vocabularies like the Sequence Ontology (EILBECK *et al.*, 2005). Clearly, comparability and traceability of published data can greatly benefit from standardized analyses of genome organization and gene structure (see also GREGORY, 2005).

A partial explanation for the lack of a standardized analysis and presentation of fundamental genomic features referring to protein-coding genes is a lack of software that infers the desired statistics. Available tool suites like BEDtools

(QUINLAN and HALL, 2010), genomeTools (GREMME *et al.*, 2013), AEGeAN<sup>II.1</sup>, and gfftools<sup>II.2</sup> are mostly intended for processing rather than describing annotations. While various programming libraries, such as BioPerl<sup>II.3</sup> and SeqAn (DÖRING *et al.*, 2008) provide suitable methods, their usage is demanding to researchers without programming experience and fosters the development of custom scripts by researchers with programming skills. The former likely limits the number of scientists who can infer the desired statistics, while the latter increases the risk of inferring incompatible results due to errors and/or misconceptions in analyses and definitions. Thus, there is a need for easy to use software that provides the facility to examine genome annotations for a wealth of structural features of the protein-coding gene repertoire in a concise way and that provides basic and standardized statistics as well as results suitable for downstream applications.

## 2.2 Why another tool?

The tool COGNATE, a Comparative Gene Annotation Characterizer, fills the above identified gap of software for structural characterization of the annotated protein-coding gene repertoire of a genome. COGNATE allows a quick and easy extraction of basic genome features and gene repertoire data; it is thus a tool to primarily describe a genome and its annotated protein-coding gene repertoire, which is an essential prerequisite for comparative analyses. Given the ongoing genome sequencing efforts, especially by large consortia like 10 k (KOEPLI *et al.*, 2015) and i5k (I5K CONSORTIUM, 2013), there is an increasing demand for a standardization of large-scale comparisons of genome and gene structure.

---

II.1 Standage DS. AEGeAn: an integrated toolkit for analysis and evaluation of annotated genomes. 2010-2015. <http://standage.github.io/AEGeAn>. Last accessed 20 March 2017

II.2 GitHub: Holmes I. gfftools. 2011. <https://github.com/ihh/gfftools>. Last accessed 20 March 2017

II.3 The BioPerl Project. 2016. <http://bioperl.org>. Last accessed 20 March 2017.





---

## Methods and implementation

---

WITH COGNATE, we promote a tool to simultaneously analyze a given protein-coding gene annotation and the corresponding assembled sequences of a genome, here referred to as scaffold or contig sequence (SCS). An overview of the software's input, work flow, analyzed parameters, and output is visualized in Fig. II.1. A complete list of analyzed parameters is given in Additional file ??, a glossary with the definitions of all terms used in this publication and by COGNATE is provided in Table B.1.1.

### 3.1 Input and running

COGNATE requires as input: (1) a gff file in GFF3 format<sup>II.1</sup> containing the annotation of protein-coding genes; (2) a fasta file, containing the corresponding genomic nucleotide sequences, which are exploited to infer the length, GC

---

<sup>II.1</sup> The Generic Model Organism Database: GFF format definition. 2016.  
<http://gmod.org/wiki/GFF3>. Last accessed 20 March 2017

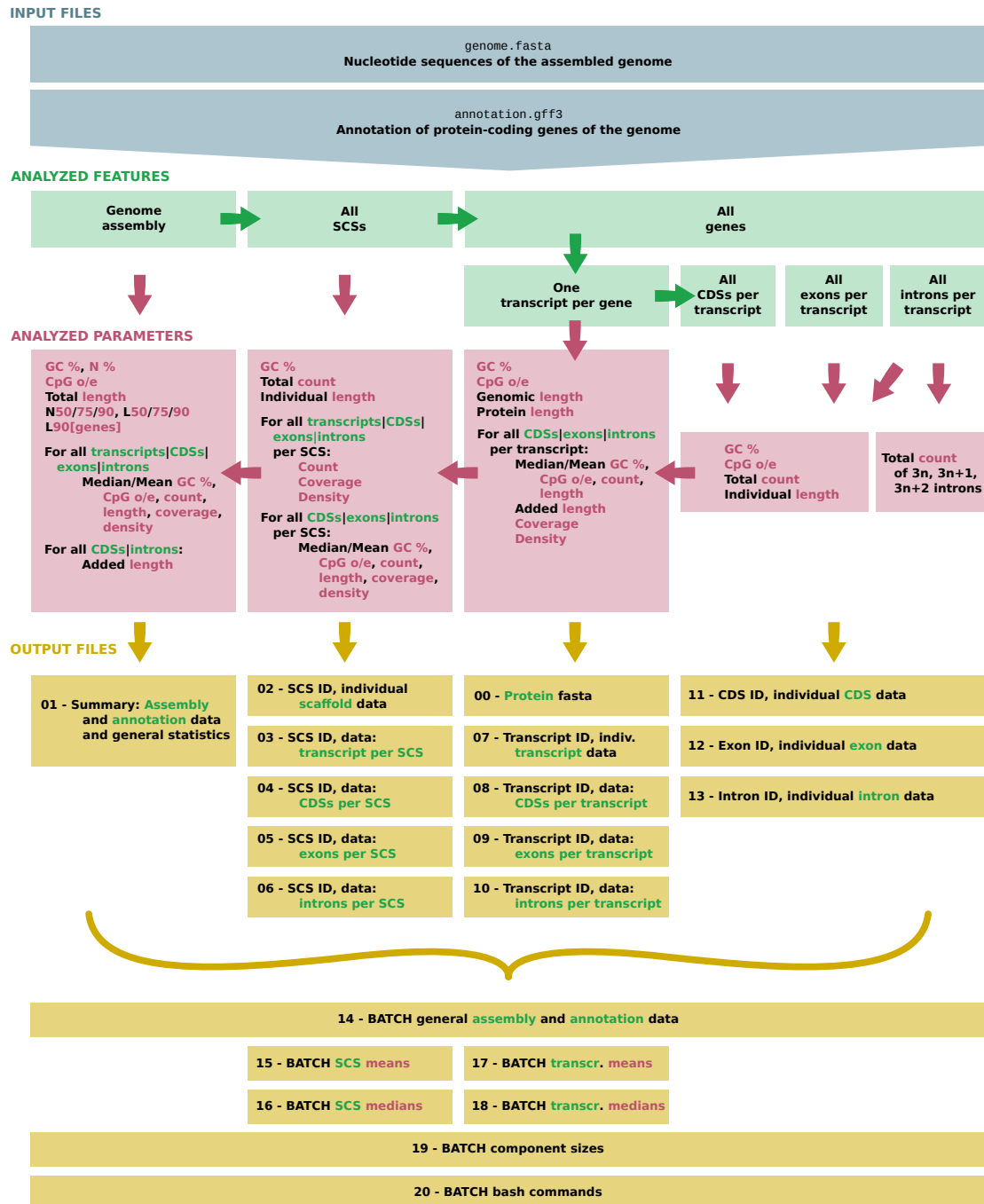


Figure II.1 – Overview of the information flow in the software package COGNATE.  
(Continued on next page.)

**Figure II.1 – Overview of the information flow in the software package COGNATE.**

(Continued)

The Perl script COGNATE requires two files per run as input (blue): a fasta file containing the assembled nucleotide sequences and a GFF3 file with the protein-coding gene annotation information. The input (blue) is used to analyze genomic and genic features (green) on the level of assembly, SCSs, transcripts, CDSs, exons, and introns. Each complex of analyzed features is evaluated individually and the analyzed parameters are condensed in a step-wise manner by calculating means and medians (red). As output (yellow), 21 files are generated, of which all except two are in TSV format (the exceptions are: 00, protein fasta; 20, bash commands). The output files are split according to the analyzed features and parameters. All data files (02–13) are ordered by the ID of the respective feature. BATCH files (14–20) contain one entry line per genome and thus data of multiple COGNATE runs to facilitate direct comparisons of genomes.

CDS: CoDing Sequence; GFF: Generic Feature Format; SCS: Scaffold or Contig Sequence; TSV: Tab-Separated Values

content, and amino acid sequences of the assembled SCSs and of the predicted protein-coding genes, respectively. The gene annotation has to include at least the features 'gene', 'mRNA', and 'exon', as provided by, for example, BRAKER1 (HOFF *et al.*, 2016) and MAKER2 (HOLT and YANDELL, 2011). Thus, the analysis of partial and pseudogenes depends on their annotation in the analyzed gff file; non-coding genes (*i.e.*, genes without mRNA) are not considered in the analysis. Further technical requirements are several standard Perl libraries as well as the `GAL::Annotation` and `GAL::List` libraries to allow gff-handling. The latter two libraries are available from the Sequence Ontology Project<sup>II.2</sup> and are also included in the COGNATE software package. COGNATE is written in Perl and has been tested under Ubuntu 12.04 and 14.04. COGNATE analyzes one genome at a time. Providing multiple genomes (*i.e.*, a batch) for serial processing is possible with a special input file (see README in B.2.4). Serial, single-threaded processing leads to a linear relationship of processed genomes and required time. As a gauge, the analysis of the latest *Apis mellifera* gene set (see II.4), which has a genome size of 250.3 Mb and 10,733 annotated protein-coding genes, takes with COGNATE up to 4 h, using up to 600 MiB RAM. For comparison, COGNATE requires a very similar amount of time for the

II.2 The Genome Annotation Library. 2016.  
<http://www.sequenceontology.org/software/GAL.html>. Last accessed 20 March 2017

analysis of the gene set<sup>II.3</sup> of *Ixodes scapularis* (genome size: 1765.4 Mb, 20,467 annotated protein-coding genes). A benchmark comparison of COGNATE to other software, such as genomeTools (KOEPLI *et al.*, 2015), AEGeAN<sup>II.1</sup>, or gfftools<sup>II.2</sup>, is not meaningful due to major differences between these software packages in focus and aim. At the moment, no tool yields the wide array of metrics that COGNATE delivers by default.

## 3.2 Output

COGNATE infers the following major metrics (for a full list of the 296 parameters, see Additional file B.2.1):

- summary counts of the analyzed features, including L90pcG<sup>II.4</sup>, *i.e.*, the number of SCSs needed to cover 90 % of all annotated protein-coding genes;
- strandedness of transcripts and features (CDSs, exons, and introns);
- lengths and length statistics (nucleotide/amino acid sequences), including N50/L50, 75/L75, N90/L90;
- intron length distribution (ROY and PENNY, 2007);
- percental GC content statistics in two different ways, namely

II.3 Data of *Ixodes scapularis*: NCBI: FTP directory of the *Ixodes scapularis* genome version JCVI\_ISG\_i3\_1.0 and the corresponding protein-coding gene annotation (NCBI RefSeq). 2017. [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/208/615/GCF\\_000208615.1\\_JCVI\\_ISG\\_i3\\_1.0/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/208/615/GCF_000208615.1_JCVI_ISG_i3_1.0/). Last accessed 20 March 2017.

II.4 L90pcG, the count of SCSs necessary to cover 90 % of the annotated protein-coding genes in an assembly, is here, to our knowledge, explicitly termed for the first time. Similar metrics have been described in other publications, for example, the “number of whole genome CARs [Contiguous Ancestral Regions] that cover 90 % of one-to-one orthologous families” (NEAFSEY *et al.*, 2015, supplementary online material, p. 57). Although the notation of L for a number (instead of a length) appears to be counter-intuitive, we deliberately decided to follow the already established convention of N50 and L50, with N50 designating “maximum length L such that 50 % of all nucleotides lie in contigs (or scaffolds) of size at least L” (LANDER *et al.*, 2001), and L50 designating the “number of sequences evaluated at the point when the sum length exceeds 50 % of the assembly site” (Bradnam K. ACGT. 2015. <http://www.acgt.me/blog/2015/6/11/150-vs-n50-thats-another-fine-mess-that-bioinformatics-got-us-into>. Last accessed 23 May 2017).

- ⇒ using a calculation that explicitly considers IUPAC ambiguity codes (G, C, S per total length excluding N, R, Y, K, M, B, D, H, V);
- ⇒ using the previously prevailing calculation of GC per total length, which is inappropriate for genome comparisons due to its dependence on assembly quality;
- statistics of CpG dinucleotide depletion (CpG observed/expected), normalized by C and G content of the respective region (ELANGO *et al.*, 2009);
- density statistics (ratio of the length of a feature covered by another, number-wise);
- coverage statistics (ratio of the length of a feature covered by another, length-wise).

In summary, the output parameters can be classified as computations of the eight above major metrics or feature types, some with child types (*e.g.*, added length), of six structural entities (*e.g.*, assembly/annotation, SCSs, introns). In other words, parameters are inferred on several levels. For example, the total count of CDSs in analyzed transcripts is given for the entire assembly as well as on a per transcript basis. For the latter, COGNATE also calculates the mean and median count of CDSs per transcript as well as the mean/median of these medians over all transcripts. As another example, the intron density of a gene is calculated as the total number of introns divided by the length of the gene (*i.e.*, genomic length of the transcript, including introns and exons) and also given as mean/median intron density per gene over the whole annotation. For each gene, only one representative (optionally the longest [default], shortest, or median-length) transcript is evaluated. The analysis is independent of homology hypotheses (*i.e.*, not limited to gene families), thus comprising information on a genome's entire annotated protein-coding gene repertoire.

As output, COGNATE provides various result tables in TSV format:

- a concise overview (summary) of measured variables;
- lists of all measured variables referring to features of
  - ⇒ a given SCS, transcript, or individual CDSs, exons, or introns, respectively;
  - ⇒ summary, the output parameters can be classified

- 'batch' files, which contain one line of summary statistics per analyzed genome. There are individual files for general genome data and means and medians of SCS and transcript data, respectively;
- a component size overview (*i.e.*, the added length [in bp] of all coding and intron sequences, respectively), which offers a basis for a comparison of these values with statistics of other genomic features inferred with other tools, for example non-coding elements.

All above specified files (except the ones providing an overview) facilitate tests for correlations between parameters within and among genomes. The output files are formatted specifically to allow easy import in statistical software, such as R (R CORE TEAM, 2017) and SPSS (*IBM SPSS Statistics for Windows* 2013). COGNATE also provides a fasta file ('analyzed\_transcripts') containing the predicted amino acid sequences inferred from the CDSs of the one analyzed transcript per gene. This file can be used, for example, as input for BUSCO (SIMÃO *et al.*, 2015) to test for the completeness of the gene set, which is facilitated by the ready-made bash commands supplied in the 'bash commands' text file. The generation of all output files can be controlled directly by the user. The output of COGNATE can be used in manifold analyses, ranging from a descriptive characterization to an in-depth comparative analysis of gene organization across multiple genomes. This is further exemplified in the discussion.

---

## Results and discussion

---

### 4.1 Applicability of COGNATE

IT IS AN ESSENTIAL FEATURE of COGNATE to provide not only descriptive statistics but also the complete primary data, since “an over-reliance on simple summary statistics [...] can obscure real biological trends and differences” (MOSS *et al.*, 2011, p. 1191). Apart from other already mentioned potential applications, COGNATE output can be used to study the variability of gene structure within a genome and to compare it with that in other genomes. In such an instance, the list of transcript features can be exploited to analyze the range of exon lengths, intron lengths, and their distribution over genes of a certain GC content. Another example would be a comparison of GC content in coding and non-coding regions of genes across a genome. Having the characteristics of a gene repertoire at hand, they can be compared to those of other species and used in phylogenomic analyses (*e.g.*, NIEHUIS *et al.*, 2012). COGNATE results can also serve as a starting point to find genes of interest and relate them to functions, *e.g.*, looking for very long or short genes or investigating genes containing exactly two CDSs. Hypotheses like “Flying birds

have shorter introns than birds of non-volant sister lineages due to energetic demands of powered flight” (Q ZHANG and EDWARDS, 2012), “Evolutionary changes in intron lengths correlate with co-expression of genes” (KEANE and SEOIGHE, 2016), or “Strategies of splice-site recognition are influenced by differences in GC content between exons and introns” (AMIT *et al.*, 2012) could thereby be tested in more detail. Thus, COGNATE provides data to facilitate downstream analyses, and in addition, provides summary statistics that can help standardizing genome parameter disclosure.

## 4.2 Standardization problems: missing values, fuzzy terminology, and inaccuracy

Missing standardization in comparative genomics can easily lead to problems in meta-analyses and consequently result in biased conclusions. As ELLIOTT and GREGORY (2015) noted during their tremendous effort of data compilation, there are problems of standardization in terms of parameter listing and source disclosure as well as of definitions of descriptive terms. Some of these subtle and sometimes deemphasized problems are elucidated here in more detail to raise and sustain the awareness for them.

One problem in compiling data for meta-analyses are missing values. The data matrix compiled by ELLIOTT and GREGORY (2015) (Supplement of ELLIOTT and GREGORY, 2015<sup>II.1</sup>) contains overall 52 % missing values due to incomplete data disclosure by publications or missing entries in databases. This lack of data introduces a potential bias in correlative analyses of genome structures, which has not been systematically investigated. Thus, without in-depth parameter disclosure, the enormous effort of collecting data from open sources for genome and gene structure comparison potentially yields unreliable results. The general distribution of missing data in the matrix compiled by ELLIOTT and GREGORY (2015) is noteworthy in that the GC content is almost always given while values

---

<sup>II.1</sup> Supplement 1 – Genome data used in the analyses. 2015.  
[http://rstb.royalsocietypublishing.org/highwire/filestream/32237/field\\_highwire\\_adjunct\\_files/0/rstb20140331supp1.xlsx](http://rstb.royalsocietypublishing.org/highwire/filestream/32237/field_highwire_adjunct_files/0/rstb20140331supp1.xlsx). Last accessed 20 March 2017



related to gene structure including intron size values are missing for half of the genomes in the data matrix (see Fig. II.2). It is surprising to find that for 38 % of the genomes in their dataset no assembly genome size was included in the original publications or databases. To further illustrate the problem of missing data in comparative genomics, we analyzed the genome (version 4.5, downloaded 31 August 2015, from NCBI<sup>II.2</sup>) and latest protein-coding gene annotation (release 103, downloaded 20 March 2017 from NCBI<sup>II.3</sup>) of *Apis mellifera*. Compared to the 144 values recorded by COGNATE that can readily be given as a single number, the publications covering the official gene sets 1 (MORIOKA *et al.*, 2006) and 3.2 (ELSIK *et al.*, 2014) offer only eight and nine comparable values, respectively; NCBI offers a report site<sup>II.4</sup> for the most recent annotation release (103), where we found 14 comparable values (Additional file B.2.1, sheet 2). The obtained values differ on a small scale (for example, the count of protein-coding genes differs by 5 for a total of circa 10,730), most likely due to the different annotation versions or deviating definitions. Generally, COGNATE can help to mitigate the problem of missing values by easing their acquisition and has the benefit of providing tractable values with a transparent method.

Problems of fuzzy terminology become apparent when, for example, the coding amount (*i.e.*, the total length of protein-coding sequences within a genome) is given in exonic megabases (Mb) (Fig. II.2; ELLIOTT and GREGORY, 2015). Given the functional and structural similarity of exons and CDSs and their often complete overlap in automated annotations, it is an understandable, yet potentially misleading lack of differentiation. In contrast to CDSs, annotated exons can include untranslated regions (UTRs) and stop codons; not every exon is a coding sequence (MQ ZHANG, 2002). Most of the automated annotations

---

II.2 NCBI: FTP directory of the *Apis mellifera* genome version 4.5 (NCBI RefSeq). 2016.

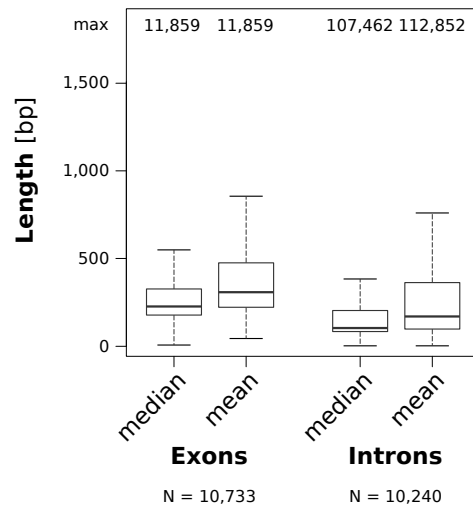
[ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/195/GCF\\_000002195.4\\_Amel\\_4.5/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/195/GCF_000002195.4_Amel_4.5/). Genome file downloaded 31 August 2015. Last accessed 20 March 2017

II.3 NCBI: FTP directory of the *Apis mellifera* annotation release 103. 2017.

[ftp://ftp.ncbi.nlm.nih.gov/genomes/Apis\\_mellifera/GFF/](ftp://ftp.ncbi.nlm.nih.gov/genomes/Apis_mellifera/GFF/). Annotation file downloaded 20 March 2017. Last accessed 20 March 2017

II.4 NCBI: NCBI *Apis mellifera* Annotation Release 103 report site. 2016. [https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Apis\\_mellifera/103](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Apis_mellifera/103)

accessed 20 March 2017



**Figure II.2** – Comparison of means and medians of exon and intron length in the genome of *Apis mellifera*. We applied COGNATE with default options (thus using the longest of each gene’s alternative transcripts) to the genome and gene annotation of *A. mellifera*, version 4.5. The respective data were downloaded from NCBI<sup>II.2 II.3</sup> and constitute a RefSeq annotation, predicted by the NCBI Eukaryotic Genome Annotation Pipeline. Shown is the comparison between the mean and median values of the exon length and of the intron length per transcript in bp, respectively. COGNATE considered 10,733 transcripts, comprising of 76,276 exons and 65,543 introns. N = 10,733 for mean and median exon lengths and N = 10,240 for mean and median intron lengths. The means of exon lengths are 355.32 (medians) and 452.99 (means) bp, means of intron lengths are 613.08 (medians) and 1502.31 (means) bp. The primary data are provided in Additional file B.2.3

do not include UTRs, which are difficult to delineate *de novo* (e.g., GRIFFITH *et al.*, 2008; MIGNONE *et al.*, 2005); nevertheless, a future project is to include the analysis of UTR annotations in COGNATE. Thus, in this instance, it remains unclear in which form exons and CDSs were evaluated and contributed to a summary statistic. With the above example, we are illustrating why we stress the importance of clear definitions and applications of these to genome and gene structure characterizations. Accordingly, COGNATE differentiates between CDSs and exons, but it can only be as accurate as the given annotation.

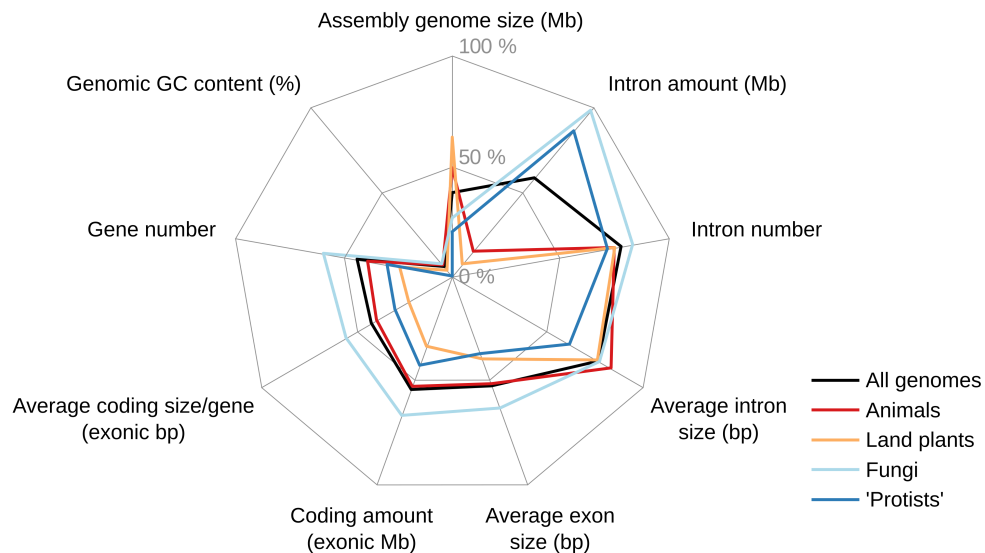
For a complete list of our definitions, compared to Sequence Ontology terms<sup>II.5</sup>, see the glossary in section B.1.1. The problems of defining a universally needed term such as ‘gene’ (described in GERSTEIN *et al.*, 2007) as well as the various ways and needs of gene annotation (MUDGE and HARROW, 2016) render the ongoing efforts of finding precise and useful definitions both essential and exacting.

Another problem of terminological and methodological nature is the widespread use of means as descriptive summary statistic. Since many gene structure features are not normally distributed within a genome, the mean is an inappropriate summary statistic of these features. Yet, in many investigations, only the mean is calculated as a summary statistic of gene structure features (see (ELLIOTT and GREGORY, 2015) as well as the publications cited therein). Doing so can bias analyses and severely mislead comparisons between genomes, especially when one is represented by a mean, the other by a median. To illustrate this, we used results of COGNATE from analyzing the latest gene set of *Apis mellifera* (see above) and compared the obtained values of mean and median of exon size and intron size per transcript, respectively (Fig. II.3, data in Additional file B.2.3). In normally distributed data, means and medians are expected to be (nearly) identical, which is clearly not the case in *A. mellifera*. COGNATE calculates both means and medians for a wealth of parameters.

A third example of unclear usage of terms relates to the evaluation of intron density. The two above evaluated parameters – exon size and intron size per transcript – together with intron density per transcript can be understood as a proxy for gene structure, as demonstrated by YANDELL *et al.* (2006), and are thus of great interest in structural gene characterization. Note however that intron density as calculated by YANDELL *et al.* (2006) relates to protein length (*i.e.*, count of introns/protein length). We advocate (and implemented in COGNATE) the relation of intron density to gene length as described above, since proteins as well as mature mRNAs are spliced and thus intron-free.

---

<sup>II.5</sup> The Sequence Ontology. 2016. <http://www.sequenceontology.org>. Last accessed 20 March 2017.



**Figure II.3** – Amount of missing data [%] in nine selected parameters analyzed by ELLIOTT and GREGORY (2015). We selected nine parameters evaluated by ELLIOTT and GREGORY (2015), namely those that are directly comparable to the parameters evaluated by COGNATE. These parameters are: (1) the size of the assembled genome (in Mb); (2) the GC content of the assembled genome in % (COGNATE provides here two values, taking Ns in the sequence into account and excluding them, respectively); (3) gene number (total gene count in COGNATE); (4) average coding size/gene in exonic bp (mean added CDS length per transcript in COGNATE); (5) coding amount (total added length of all CDSs in COGNATE); (6) the average exon size in bp (mean exon length in COGNATE); (7) the average intron size in bp (mean intron length in COGNATE); (8) intron number (total intron count in COGNATE); (9) intron amount (total added length of all introns in COGNATE). Please note that we applied the same parameter terminology in the figure as ELLIOTT and GREGORY (2015).

Values of these parameters were taken from the supplement of ELLIOTT and GREGORY (2015), including all genomes of the original set and partitioned by kingdoms (animals, red; land plants, orange; fungi, light blue; 'protists', dark blue). Values referring to all genomes are depicted by a black line.

The plot shows the amount of missing data, *i.e.*, for each parameter, the count of missing values per count of potential values was determined. Thus, 0 % of missing data means that all values of the genome set under scrutiny were present, as is nearly the case for GC content. bp: basepairs; CDS: CoDing Sequence; Mb: Megabases

### 4.3 Standardization suggestions

Aside from reporting important insights, ELLIOTT and GREGORY (2015) advocated the need for standardization in large-scale comparisons of genomes. The inevitable problems of analyzing datasets with missing data could, in the future, be extenuated by a common, comprehensive set of basic parameters published together with genomic data. When publishing a genome and its annotation of protein-coding genes, it would be most beneficial to attach the complete set of COGNATE results to it to avoid problems resulting from changing versions of genomes and/or annotations. A set of standard metrics to advance standardization of parameter publishing was proposed by ELLIOTT and GREGORY (2015), including “details of base pair composition, gene number, intron number and size, total repeat content, and TE abundance, diversity and activity” (ELLIOTT and GREGORY, 2015, p. 8). Many other parameters can and should be used to describe the features of a genome completely, most of which go beyond the scope of COGNATE (*e.g.*, properties of repetitive elements). Regarding protein-coding genes, we suggest to cover the descriptive parameters more broadly and to provide the following parameters as a minimum:

- assembly size (*i.e.*, total added length of all SCSs, with and without Ns),
- assembly GC content (with and without ambiguity),
- gene count,
- median transcript length (tallying one representative transcript per gene),
- median CDS length,
- median CDS count per transcript (*i.e.*, density),
- median CDS length per gene (*i.e.*, coverage),
- coding amount (*i.e.*, total added length of all CDSs),
- intron count,
- median intron length,
- median intron count per transcript (*i.e.*, density),
- median intron length per gene (*i.e.*, coverage),
- intron amount (*i.e.*, total added length of all introns).

Following the establishment of standard parameters of gene model properties and the institution of a standard tool to acquire these, the next desirable step is the constitution of a “curated, user-friendly, open-access database [to] make

this information accessible and usable in large-scale comparative analyses” (ELLIOTT and GREGORY, 2015, p. 8).

Finally, we would like to draw the readers’ awareness also to a frequently encountered problem in comparative genomics: the source of primary sequence data or the version of gene annotations are often not clearly stated, which hampers reproducibility of the published analyses. Therefore, we emphasize the need for disclosing used databases, genome versions, and other source information in combination with data and results.

---

## Conclusion

---

COMPARATIVE META-ANALYSES of gene and genome characteristics, testing, for example, whether potential proteome diversity is reliably reflected by the total amount of coding sequences (WANG *et al.*, 2011), rely on descriptive statistics of primary genome sequences and gene annotations. However, comprehensive standard statistics of genome organization and gene structure have not been fully or consistently defined with the effect that they are inconsistently collected or often incomplete. Due to this problem, comparative meta-analyses of gene and genome characteristics can be severely handicapped and are potentially unreliable. Obviously, this problem can be solved with the routine application of standard tools. The here presented software COGNATE allows effortless and flexible parameter disclosure as well as genome comparisons within its designated scope. Its merits include the comprehensive evaluation of an extensive set of standard and non-standard parameters of protein-coding genes, the provision of both primary data and summary statistics, and the use of explicit term definitions. COGNATE was developed in the hope to further promote and ease comparative studies, which should eventually yield insights into the evolution of genomes and gene repertoires.





---

## Additional publication information

---

### 6.1 Availability and requirements

COGNATE is provided as a package, including source code, helper scripts (*e.g.*, to check the presence of required Perl libraries), example data, GAL libraries, and manual at the ZFMK website and together with this publication as Additional file B.2.4.

**Project name:** COGNATE

**Project home page:** <https://www.zfmk.de/en/COGNATE>  
and <https://github.com/ZFMK/COGNATE>

**Operating system(s):** platform independent

**Programming language:** Perl

**Other requirements:** GAL libraries (included)

**License:** GNU GPLv3

The datasets analyzed during the current study are available in the NCBI RefSeq repositories<sup>II.2 II.3</sup> and from the supplement<sup>II.1</sup> of ELLIOTT and GREGORY (2015).

## **6.2 Acknowledgements**

We thank MALTE PETERSEN, JAN PHILIP OEYEN, and TANJA ZIESMANN for beta-testing COGNATE and JOSHUA D GIBSON as well as three anonymous reviewers for helpful feedback on the manuscript. We further acknowledge the students of the Leibniz Graduate School on Genomic Biodiversity Research, the i5K community, and especially ANNA CHILDERS, for valuable feedback on ideas put forth in this study. JW thanks BARRY MOORE for help implementing GAL in the software package COGNATE. Finally, BM, JW and ON thank the German Research Foundation for support of this study and acknowledge the Leibniz Association for funding the Graduate School on Genomic Biodiversity Research.

## **6.3 Funding**

This study was supported by the Leibniz Graduate School on Genomic Biodiversity Research and by the German Research Foundation (MI 649/16–1, NI-1387/3–1). The funding agencies did not influence the design of the study, the collection, analysis, and interpretation of data, or the manuscript writing.

## **6.4 Authors' contributions**

JW conceived this study. BM, JW, and ON designed the study. JW developed the software package. BM, JW, and ON wrote the manuscript. All authors read and approved the final manuscript.

---

## Bibliography II

---

- AMIT, M, M DONYO, D HOLLANDER, A GOREN, E KIM, S GELFMAN, G LEV-MAOR, D BURSTEIN, S SCHWARTZ, B POSTOLSKY, T PUPKO, and G AST (2012). **Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition.** *Cell Reports* 1.5, pp. 543–556. DOI: 10.1016/j.celrep.2012.03.013 (cit. on p. 64).
- DÖRING, A, D WEESE, T RAUSCH, and K REINERT (2008). **SeqAn An efficient, generic C++ library for sequence analysis.** *BMC Bioinformatics* 9, p. 11. DOI: 10.1186/1471-2105-9-11 (cit. on p. 55).
- EILBECK, K, SE LEWIS, CJ MUNGALL, M YANDELL, L STEIN, R DURBIN, and M ASHBURNER (2005). **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biology* 6.5, R44. DOI: 10.1186/gb-2005-6-5-r44 (cit. on p. 54).
- ELANGO, N, BG HUNT, MAD GOODISMAN, and SV YI (2009). **DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*.** *Proceedings of the National Academy of Sciences* 106.27, pp. 11206–11211. DOI: 10.1073/pnas.0900301106 (cit. on p. 61).
- ELLIOTT, TA and TR GREGORY (2015). **What's in a genome? The C-value enigma and the evolution of eukaryotic genome content.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678, p. 20140331. DOI: 10.1098/rstb.2014.0331 (cit. on pp. 51, 53, 54, 64, 65, 67–70, 73).
- ELSIK, CG, KC WORLEY, AK BENNETT, M BEYE, F CAMARA, CP CHILDERS, DCd GRAAF, G DEBYSER, J DENG, B DEVRESE, E ELHAIK, JD EVANS, LJ FOSTER, D GRAUR, R GUIGO, \f \f \$AUTHOR.LASTNAME, KJ HOFF, ME HOLDER, ME HUDSON, GJ HUNT, H JIANG, V JOSHI, RS KHETANI, P KOSAREV, CL KOVAR, J MA, R MALESZKA, RFA MORITZ, MC MUNOZ-TORRES, TD MURPHY, DM MUZNY, IF NEWSHAM, JT REESE, HM ROBERTSON, GE ROBINSON, O RUEPELLE, V SOLOVYEV, M STANKE, E STOLLE, JM TSURUDA, MV VAERENBERGH, RM WATERHOUSE, DB WEAVER, CW WHITFIELD, Y WU, EM ZDOBNOV, L ZHANG, D ZHU, RA GIBBS, and

- \f \ \$AUTHOR.LASTNAME (2014). **Finding the missing honey bee genes: lessons learned from a genome upgrade.** *BMC Genomics* 15.1, p. 86. DOI: 10.1186/1471-2164-15-86 (cit. on p. 65).
- GERSTEIN, MB, C BRUCE, JS ROZOWSKY, D ZHENG, J DU, JO KORBEL, O EMANUELSSON, ZD ZHANG, S WEISSMAN, and M SNYDER (2007). **What is a gene, post-ENCODE? History and updated definition.** *Genome Research* 17.6, pp. 669–681. DOI: 10.1101/gr.6339607 (cit. on p. 67).
- GREGORY, TR (2005). **Synergy between sequence and size in Large-scale genomics.** *Nature Reviews Genetics* 6.9, pp. 699–708. DOI: 10.1038/nrg1674 (cit. on p. 54).
- GREMME, G, S STEINBISS, and S KURTZ (2013). **GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10.3, pp. 645–656. DOI: 10.1109/TCBB.2013.68 (cit. on p. 55).
- GRIFFITH, OL, SB MONTGOMERY, B BERNIER, B CHU, K KASAIAN, S AERTS, S MAHONY, MC SLEUMER, M BILENKY, M HAEUSSLER, M GRIFFITH, SM GALLO, B GIARDINE, B HOOGHE, PV LOO, E BLANCO, A TICOLL, S LITHWICK, E PORTALES-CASAMAR, IJ DONALDSON, G ROBERTSON, C WADELIUS, PD BLESER, D Vlieghe, MS HALFON, W WASSERMAN, R HARDISON, CM BERGMAN, SJM JONES, and TORA CONSORTIUM (2008). **ORegAnno: an open-access community-driven resource for regulatory annotation.** *Nucleic Acids Research* 36 (suppl 1), pp. D107–D113. DOI: 10.1093/nar/gkm967 (cit. on p. 66).
- HOFF, KJ, S LANGE, A LOMSADZE, M BORODOVSKY, and M STANKE (2016). **BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.** *Bioinformatics* 32.5, pp. 767–769. DOI: 10.1093/bioinformatics/btv661 (cit. on p. 59).
- HOLT, C and M YANDELL (2011). **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 12.1, p. 491. DOI: 10.1186/1471-2105-12-491 (cit. on p. 59).
- HOU, Y and S LIN (2009). **Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes.** *PLOS ONE* 4.9, e6978. DOI: 10.1371/journal.pone.0006978 (cit. on p. 54).
- I5K CONSORTIUM (2013). **The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.** *Journal of Heredity* 104.5, pp. 595–600. DOI: 10.1093/jhered/est050 (cit. on p. 55).
- IBM SPSS Statistics for Windows (2013). Version 22.0 (cit. on p. 62).
- KEANE, PA and C SEOIGHE (2016). **Intron Length Coevolution across Mammalian Genomes.** *Molecular Biology and Evolution* 33.10, pp. 2682–2691. DOI: 10.1093/molbev/msw151 (cit. on p. 64).
- KOEPFLI, KP, B PATEN, THE GENOME 10K COMMUNITY OF SCIENTISTS, and SJ O'BRIEN (2015). **The Genome 10K Project: A Way Forward.** *Annual Review of Animal Biosciences* 3.1, pp. 57–111. DOI: 10.1146/annurev-animal-090414-014900 (cit. on pp. 55, 60).

- LANDER, ES *et al.* (2001). **Initial sequencing and analysis of the human genome.** *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062 (cit. on p. 60).
- MIGNONE, F, G GRILLO, F LICCIULLI, M IACONO, S LIUNI, PJ KERSEY, J DUARTE, C SACCONI, and G PESOLE (2005). **UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs.** *Nucleic Acids Research* 33 (suppl 1), pp. D141–D146. DOI: 10.1093/nar/gki021 (cit. on p. 66).
- MORIOKA, M *et al.* (2006). **Insights into social insects from the genome of the honeybee *Apis mellifera*.** *Nature* 443.7114, pp. 931–949. DOI: 10.1038/nature05260 (cit. on p. 65).
- MOSS, SP, DA JOYCE, S HUMPHRIES, KJ TINDALL, and DH LUNT (2011). **Comparative Analysis of Teleost Genome Sequences Reveals an Ancient Intron Size Expansion in the Zebrafish Lineage.** *Genome Biology and Evolution* 3.0, pp. 1187–1196. DOI: 10.1093/gbe/evr090 (cit. on pp. 53, 63).
- MUDGE, JM and J HARROW (2016). **The state of play in higher eukaryote gene annotation.** *Nature Reviews Genetics* 17.12, pp. 758–772. DOI: 10.1038/nrg.2016.119 (cit. on p. 67).
- NEAFSEY, DE *et al.* (2015). **Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes.** *Science* 347.6217, p. 1258522. DOI: 10.1126/science.1258522 (cit. on p. 60).
- NIEHUIS, O, G HARTIG, S GRATH, H POHL, J LEHMANN, H TAFER, A DONATH, V KRAUSS, C EISENHARDT, J HERTEL, M PETERSEN, C MAYER, K MEUSEMANN, RS PETERS, PF STADLER, RG BEUTEL, E BORNBERG-BAUER, DD MCKENNA, and B MISOF (2012). **Genomic and Morphological Evidence Converge to Resolve the Enigma of Strepsiptera.** *Current Biology* 22.14, pp. 1309–1313. DOI: 10.1016/j.cub.2012.05.018 (cit. on p. 63).
- QUINLAN, AR and IM HALL (2010). **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033 (cit. on p. 55).
- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing (cit. on p. 62).
- ROY, SW and D PENNY (2007). **Intron length distributions and gene prediction.** *Nucleic Acids Research* 35.14, pp. 4737–4742. DOI: 10.1093/nar/gkm281 (cit. on p. 60).
- SIMÃO, FA, RM WATERHOUSE, P IOANNIDIS, EV KRIVENTSEVA, and EM ZDOBNOV (2015). **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 31.19, pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351 (cit. on p. 62).
- WANG, M, CG KURLAND, and G CAETANO-ANOLLÉS (2011). **Reductive evolution of proteomes and protein structures.** *Proceedings of the National Academy of Sciences* 108.29, pp. 11954–11958. DOI: 10.1073/pnas.1017361108 (cit. on pp. 51, 71).
- YANDELL, M, CJ MUNGALL, C SMITH, S PROCHNIK, J KAMINKER, G HARTZELL, S LEWIS, and GM RUBIN (2006). **Large-Scale Trends in the Evolution of Gene**

- Structures within 11 Animal Genomes.** *PLoS Comput Biol* 2.3, e15. DOI: 10.1371/journal.pcbi.0020015 (cit. on pp. 53, 67).
- ZHANG, MQ (2002). **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 3.9, pp. 698–709. DOI: 10.1038/nrg890 (cit. on p. 65).
- ZHANG, Q and SV EDWARDS (2012). **The Evolution of Intron Size in Amniotes: A Role for Powered Flight?** *Genome Biology and Evolution* 4.10, pp. 1033–1043. DOI: 10.1093/gbe/evs070 (cit. on pp. 54, 64).
- ZIMMER, AD, D LANG, K BUCHTA, S ROMBAUTS, T NISHIYAMA, M HASEBE, Y VAN DE PEER, SA RENSING, and R RESKI (2013). **Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions.** *BMC Genomics* 14, p. 498. DOI: 10.1186/1471-2164-14-498 (cit. on p. 53).

---

# **AUTOMATICALLY GENERATED VS MANUALLY CURATED MODELS**

---

The following text is the author's version of the submitted article:

**WILBRANDT J, MISOF B, PANFILIO KA, NIEHUIS O (2018). How suitable are automatically inferred gene models for uncovering taxon-specific gene structural differences in gene repertoires? *BMC Genomics*, in rev.**

Authors' contributions to the original article:

Study design: BM, JW; data acquisition and analysis: JW, KAP; additional contributions to data analysis and interpretation: BM, KAP, ON; manuscript writing: JW; manuscript editing: BM, KAP, ON.





---

## Abstract

---

**B**ACKGROUND: The location and modular structure of eukaryotic protein-coding genes in genomic sequences can be automatically predicted by gene annotation algorithms. The resulting predictions are the basis for comparative studies on gene structure, gene repertoires, and genome evolution. However, such algorithms do not perform perfectly and it has been argued that only human review of the predictions ensures reliability. This implies that comparative genomics is futile until the enormous effort of manual curation of on average 21,500 gene models per eukaryotic species can be handled.

**R**ESULTS: Here we outline the prospects and limits of both automated and manual annotation in a comparative study focused on gene structure variability between species. Specifically, we compare the effect of manual curation on predicted structural properties of protein-coding genes by analyzing annotated gene sets from seven insect species sequenced by the i5k initiative. We find that the properties of automatically generated gene models and their manually curated replacements do not differ extensively, and major correlative trends

regarding gene structures can be recovered from both sets. This holds even when comparing the results of different algorithms.

**C**ONCLUSIONS: Our analyses indicate that research questions and data quality should be guideposts for the extent to which manual curation is fruitful and necessary. We anticipate further developments and benchmarks of prediction routines to improve the basis for protein-coding gene repertoire analyses.

**K**EYWORDS: Gene prediction, annotation, manual curation, exon-intron structure, insects

---

## Introduction

---

**E**UKARYOTIC PROTEIN-CODING GENE STRUCTURE is characterized by a modular organization of introns and exons (the latter being composed of a combination of coding sequence [CDS] and/or untranslated regions [UTRs]; MQ ZHANG, 2002), which are commonly identified in genome sequences using automated *in silico* gene annotation procedures (Supplementary Note C.1.2). The configuration of exons and introns — GC content, length, and number of exons and introns — varies among and within species, as well as by gene type. Various hypotheses have been proposed to explain the evolutionary persistence and variance of gene structure organization. For example, it has been hypothesized that in regions of low GC content in mammalian genomes, differential GC content of exons and introns constitutes a marker for exon recognition during splicing and is thus a factor in the stabilization of exon-intron boundaries (AMIT *et al.*, 2012; GELFMAN and AST, 2013). Further example hypotheses on gene structure organization evolution state that introns are generated by the insertion of non-autonomous DNA-transposons (HUFF *et al.*, 2016) and that in the case of birds intron size selection is driven by the evolution of powered flight (Q ZHANG and EDWARDS, 2012). The foundation of these propositions and observations is the structural description of protein-coding

gene repertoires derived from automated annotation, with all its strength and limits.

## 2.1 Prevalence of automated annotation and its problems

The development of automated procedures for unsupervised annotation of gene structures has been pursued since the 1980s and constantly improved (reviewed by, for example, BRENT, 2005, 2008; BURGE and KARLIN, 1998), but it is still not error free (KÖNIG *et al.*, 2016; X ZHANG *et al.*, 2012). The most commonly encountered errors are false positive and false negative identifications of protein-coding sequences (DENTON *et al.*, 2014; GOODSWEN *et al.*, 2012), non-coding sequence retention in coding exons (DRĂGAN *et al.*, 2016), wrong exon and gene boundaries (GOODSWEN *et al.*, 2012; GUIGÓ *et al.*, 2000), and fragmented or merged genes (DRĂGAN *et al.*, 2016; GUIGÓ *et al.*, 2006; HARROW *et al.*, 2009). With increasing gene size and complexity (*i.e.*, increasing exon count), these errors are more likely to occur and automated annotations can thus be expected to be increasingly less accurate (FRANCIS and WÖRHEIDE, 2017; GUIGÓ *et al.*, 2000; PANFILIO *et al.*, 2017). Another factor negatively influencing the results of automated annotation is draft assembly quality (TREANGEN and SALZBERG, 2012). The more fragmented an assembly is, the less likely it is to find a complete gene on one fragment (FRANCIS and WÖRHEIDE, 2017; YANDELL and ENCE, 2012). This problem becomes exacerbated as the size of the sequenced and analyzed genome increases, since larger genomes harbor a greater repetitive content load (GREGORY, 2005). The gene set predicted by automated annotation procedures also depends on whether or not extrinsic evidence (*i.e.*, alignments of homologous proteins and other amino acid or DNA sequences from species other than the one being annotated) is used for gene sequence delineation: algorithms that incorporate extrinsic evidence will likely more reliably predict genes with conserved coding or amino acid sequence (YANDELL *et al.*, 2005). However, genes that do not resemble the provided extrinsic evidence, being potentially taxon-specific and/or fast evolving, could be missing from the annotation, as such algorithms may demand a certain amount of evidence support for prediction (CANTAREL *et al.*, 2008). Annotation results thus depend on the availability and quality of evidence in terms of sequence diversity, sequencing approach, and taxonomic coverage. This

dependence becomes a problem when little evidence is available, for example, when annotating genome sequences of previously unsequenced taxa. Despite these caveats, advantages of automated gene annotation are the speed and ease of application on (multiple) genome assemblies, error consistency, and the theoretical feasibility of error tracking due to the application of explicit algorithms. With an expected average number of 21,500 protein-coding genes in a eukaryotic genome (ELLIOTT and GREGORY, 2015), the automated approach is the method of choice to comprehensively annotate gene repertoires despite the risk of erroneous models.

## 2.2 Manual curation as corrective

Structural annotation errors can pose a serious problem in comparative analyses and have been held responsible for false positive detection of clade-specific genes (BÁNYAI and PATTHY, 2016), inference of incorrect gene copy numbers (DENTON *et al.*, 2014), and misleading functional annotations based on erroneous coding sequences (PROSDOCIMI *et al.*, 2012). Consequently, many authors rely on manual curation, that is, the critical review and correction by hand, of automatically generated gene annotations, to study gene structure evolution (*e.g.*, GOODSWEN *et al.*, 2012; HARROW *et al.*, 2009; MISRA *et al.*, 2002; PANFILO *et al.*, 2017; YANDELL *et al.*, 2005). Accordingly, manual curation is often termed the ‘gold standard’ in annotation, as coined by GUIGÓ *et al.* (2006). However, manual curation is not free of problems either. Although curators follow guidelines (as stated, for example, by i5k<sup>III.1</sup>), their personal background knowledge, training, experience, and interpretation likely influence the result. Manual curation thus depends on individual proficiency, on the underlying research questions and aims, as well as on the (sometimes implicit) use of additional evidence. This can lead to potentially high variation in quantity and reliability of curated gene annotations (MISRA *et al.*, 2002). Additionally, a limited number of experts involved in many different annotation projects might bias the selection and revision of curated genes. The impact of curator

---

III.1 i5k Workspace @ NAL. Manual Curation Overview.  
<https://i5k.nal.usda.gov/manual-curation-overview>. Accessed 18 December 2017

experience and diversity within an annotation project has to our knowledge not been systematically addressed. This impact is especially interesting in cases of conflicting evidence, when curators are confronted with dilemmas and (in)consistent decisions can bias annotation results. Finally, the often incomplete documentation of the review processes, including records of changes and used sources, hampers tracking annotation errors. Nonetheless, automatically annotated models generally benefit from manual curation (MISRA *et al.*, 2002; PANFILIO *et al.*, 2017).

### **2.3 Implication of ‘gold standard’ manual curation**

The prevalence of manual curation as the gold standard goes to such lengths that some researchers endorse the view that only manually curated protein-coding gene models are of sufficient quality and thus imperative to infer hypotheses on the evolution of genomes and genes (*e.g.*, AMID *et al.*, 2009; MISRA *et al.*, 2002). Typically only a minor, most likely specifically selected, fraction of the gene repertoire is manually curated. Thus, the implied limitation of comparative genomics to only manually curated gene models for all approaches would confine this research area to meaninglessness. Consequently, it is vital to assess the reliability of the currently available automated structural annotation approaches and to investigate the differences between automated and manual curation in a systematic way.

### **2.4 Benchmarking manual versus automated annotation**

Fifteen years ago, MISRA *et al.* (2002) assessed the effects of manual curation in comparison to the previous, computationally generated gene annotation of the fruit fly *Drosophila melanogaster* (Diptera). Although their seminal findings regarding the shortcomings of gene prediction, including inaccurate structures and failure to delineate complex genes, most likely influenced the development of prediction algorithms thereafter, comparable studies of the current state of the art are lacking. To our knowledge, there has been no benchmark study published that compares properties or quality of manual versus automated structural annotation of protein-coding genes across species

and that includes recent algorithmic developments. In contrast, in addressing the reliability of automatically generated functional annotations, ŠKUNCA *et al.* (2012) documented an improvement of automated functional assignments to a quality comparable to that of manually curated functional annotations. Functional annotation considers the role or function of a gene's protein product, while structural annotation regards solely the location of a gene's components within the genome (see Supplementary Note C.1.2).

In summary, automated structural annotations are suspected to mislead or reduce the power of analyses of gene repertoire evolution, while manual curation, although being considered the gold standard, is not only labor-intensive and restricted to a fraction of a species' protein-coding gene set but is also potentially afflicted with subjectivity. Beyond spot-check vetting of automatically predicted models, the decision of when to invest in manual curation requires background knowledge not only on the error-based limitations of automated annotation, but also on the effects of manual curation (and factors influencing it, like personal experience). Using available data, we can provide first insights into the potential impact of manual curation on structural annotations of protein-coding genes, and thus also on downstream analyses, and illustrate in which cases the expenditure of human review appears to be not only fruitful but necessary.

## 2.5 Comparing automatically generated and manually curated gene models

In this study, we present a comparative analysis of manually curated gene models and their automatically annotated predecessors as well as of the gene sets containing these. We examine the extent of manual curation effects on the structural properties of protein-coding gene sets of seven insect species (Supplementary file III.1) generated within the i5k initiative (15K CONSORTIUM, 2013). The scrutinized species represent taxonomically distant insect clades (last common ancestor ca. 370 million years ago; MISOF *et al.*, 2014) and differ in genome size and assembly quality from each other (Supplementary Note C.1.2). Furthermore, only for these species were both automated and manually curated gene sets available, with on average 5.3 % of the original

genes having been manually curated (Table III.1. As our aim is to investigate global properties of protein-coding gene sets, we focus this analysis deliberately on structural features (*i.e.*, we disregard protein functions or roles) in insect genomes. Genomes of insects appear to be more difficult to assemble and annotate than those of vertebrates due to the faster pace with which insect genomes seem to evolve, resulting in a comparatively low similarity between insect genomes in general (ZDOBNOV and BORK, 2007, but see SIEPEL *et al.*, 2005].

Here, we survey the potential influence of individual human experts on the whole process of manual curation and highlight the curators' dilemmas. We also check the degree to which gene models selected for curation are representative for the whole gene repertoire regarding structural parameters. Following this, we assess the general changes in five structural parameters of gene models induced by manual curation. Furthermore, to substantiate the universality of our findings, for three insect species we additionally compare our primary data set of automatically predicted gene models and manually curated versions with automated re-annotations from a second pipeline. Finally, we explore whether reported correlations of gene counts and coverages (from ELLIOTT and GREGORY, 2015) as well as of gene composition (from ZHU *et al.*, 2009) can be recovered using uncurated and curated annotations alike.

Although manual validation of models adds value to annotations, the limitation by work force requires a directed allocation of efforts. We argue that uncurated annotations can be sufficient to investigate certain research questions, especially regarding properties of protein-coding genes on the structural level in large-scale repertoire-wide comparisons irrespective of functional classification. Examples of such questions are: 'Can we identify and explain evolutionary patterns of exon length variation (cf. YANDELL *et al.*, 2006) across insect orders?' or 'Is there a reduction of structural gene complexity in accordance with small genome size in parasites?'



---

## Results

---

### 3.1 Curator experience varies

A SYSTEMATIC ASSESSMENT of the diversity and proficiency of curators and their effects on manual curation processes is difficult, as there is no formalization or standardized documentation of this aspect to rely on. Similarly, it is currently not feasible to systematically quantify quality gain due to manual curation. For now, we decided to gauge the potential impact of personal experience in reviewing protein-coding gene models by evaluating the number of annotation projects in which each curator participated (section III.5).

For our sample of seven species, 7.6 % of all curators participated in four or more of the analyzed projects (here regarded as being 'experienced') and provided 24.4 % of all curated models. One of the two curators who participated in all projects handled 22 times more gene models than the average of all curators (Fig. III.1 a; detailed results in Supplementary Note C.1.3, Additional File C.2.2). Our results suggest that few experienced curators have a relatively large influence on selection and handling of gene models during curation.

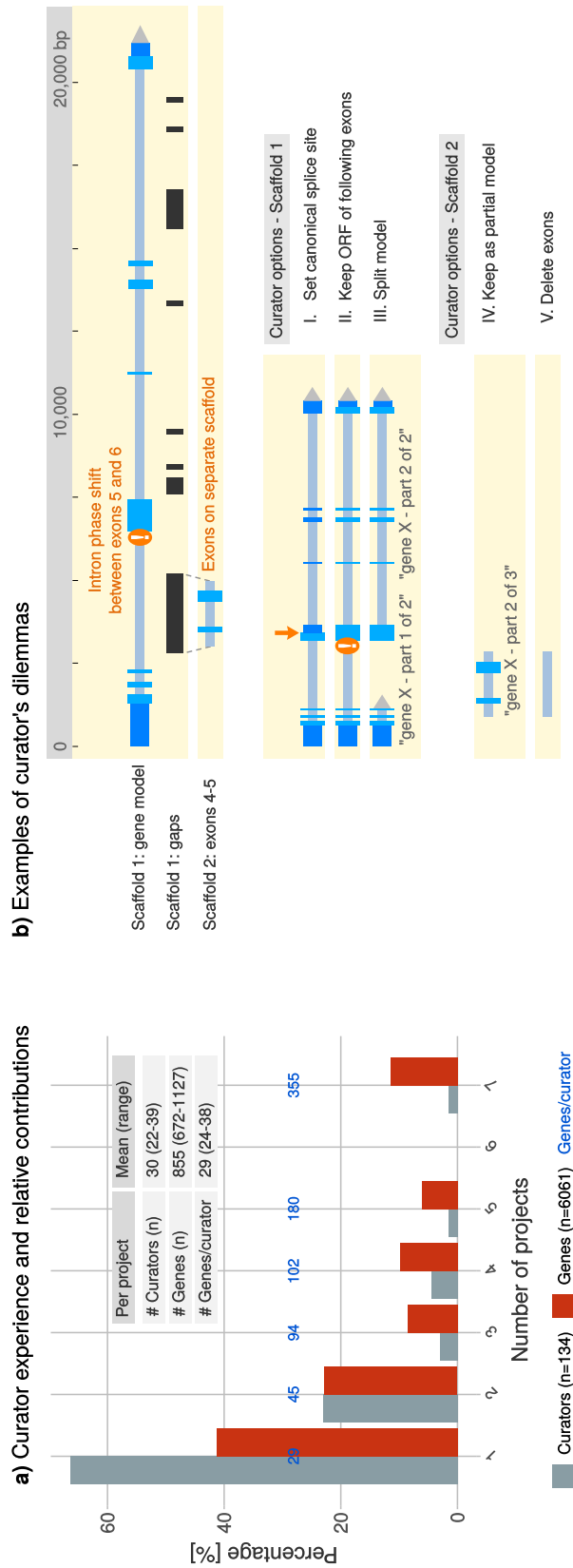


Figure III.1 – Curator experience and dilemmas. (Continued on next page.)

**Figure III.1 – Curators.** (Continued)

**a) Curator experience and relative contributions.** Percentage of curators that participated in a given number of projects (*i.e.*, an 'experience group') in relation to the total number of curators in all projects ( $n = 132$ ) (gray), and the percentage of manually curated gene models that were revised by the curators of a given 'experience group' in relation to the total count of curated gene models ( $n = 6057$ ) (red). Numbers (blue) indicate the average count of genes revised per curator in the respective group. The two curators that participated in all seven annotation projects contributed 643 and 67 gene model revisions, respectively. Data in the inserted table refers to individual projects.

**b) Examples of curator's dilemmas.** Theoretical gene X consists of ten exons, split across two scaffolds in the genome assembly. Protein-coding regions are represented by light blue boxes, untranslated regions (UTRs) by thinner, dark blue boxes. Curation option I retains a correct exon structure, but thereby induces a frame shift with a premature stop codon and 3'UTR, losing ORF and protein information within the model on Scaffold 1. Option II retains ORF and protein information, but necessitates the use of an incorrect, non-canonical acceptor splice site. Option III, splitting the model in two parts, also retains correct exon structure and ORF information, but yields incomplete models, where the functional protein domain may be split or only occur in one part, which limits domain/homology recognition. Curation option IV implies that the true exons are included in the official gene set, but within a model that lacks start/stop codons and UTRs and may lack conserved protein sequence for homology recognition, inflating the gene count and potentially the number of apparent lineage-specific proteins. Option V, ignoring exons 4-5 by not annotating a *de novo* model or actively deleting automatic predictions, leads to fewer incomplete models in the annotation, but also to information loss regarding protein-coding regions.

## 3.2 Curator's dilemmas

Experience from i5k genome annotation projects discloses not only that automatically generated gene models can be grossly wrong, but also under which circumstances both automated and manual annotation meet their limits. Problems during manual curation and decision consequences are rarely made visible (but see PANFILIO *et al.*, 2017). Thus, we highlight here an example dilemma with multiple possible outcomes, stemming from intron phase shift combined with mis-assembly of a realistic gene model, within the confines of manual edits that can be documented within gff files for assembled scaffolds (Fig. III.1 b).

We observed that the MAKER pipeline allows non-canonical translation start codons (results in Additional File C.2.3). Of the proteins encoded by MAKER-

predicted gene models, between 11.8 % (in *Cimex lectularius*, Hemiptera) and 29.9 % (in *Frankliniella occidentalis*, Thysanoptera) do not start with methionine (M). In almost all species, some protein sequences even start with X (IUPAC ambiguity code serving as a wild-card for any amino acid). While uncurated MAKER models retain non-canonical start codons, the act of curation through the Apollo graphical web interface automatically reconfigures the predicted ORF (open reading frame) to strictly require a methionine start, which may be inappropriate for a fragmented gene model and thus necessitate advanced curator actions.

### **3.3 Gene models selected for manual curation are representative**

The very restricted number of experienced curators together with the small number of gene models selected for manual revision in comparison to the whole gene set raises concerns that this selection may not be representative regarding certain gene structure parameters (*e.g.*, if manual annotation was focused on models with very long introns, as suggested by DRÄGAN *et al.*, 2016).

		Median of 'analyzed' set							
		Assembly size (% det. nucs)	Curated models: count (% of OGS)	Curation actions: # of splits merges	Transcript length: a / m	Protein length: a / m	Exon count: a / m	Exon length: a / m	
Holometabolous	Coleoptera	<i>Anoplophora glabripennis</i>	707.7 (85.1)	771 (3.5)	45 / 61	6,183 / 5,789.5	358 / 389	4 / 4	1,210 / 1345.5
		<i>Leptinotarsa decemlineata</i>	1170.2 (58.0)	933 (3.8)	22 / 125	8,562.5 / 9,280	255 / 300	4 / 4	984 / 1127
	Hymenoptera	<i>Athalia rosae</i>	163.8 (95.7)	163.8 (95.7)	825 (6.9)	28 / 34	4,340 / 3,208	445 / 423	6 / 5
		<i>Orussus abietinus</i>	201.2 (92.7)	672 (6.1)	17 / 41	5,200 / 3,996	430 / 419	5 / 5.5	2,151 / 1,828
Hemimetabolous	Hemiptera	<i>Cimex lectularius</i>	650.5 (79.0)	780 (5.5)	23 / 95	4,362 / 4,360	358 / 372.5	5 / 5	1,200 / 1,194.5
		<i>Oncopeltus fasciatus</i>	1098.7 (70.4)	945 (4.8)	14 / 160	9,324 / 11,244	257 / 320	4 / 4	1,086 / 1,347
	Thysanoptera	<i>Frankliniella occidentalis</i>	415.8 (63.4)	1,127 (6.3)	39 / 64	5,001.5 / 4064	419.5 / 419	6 / 6	1,807.5 / 1,755

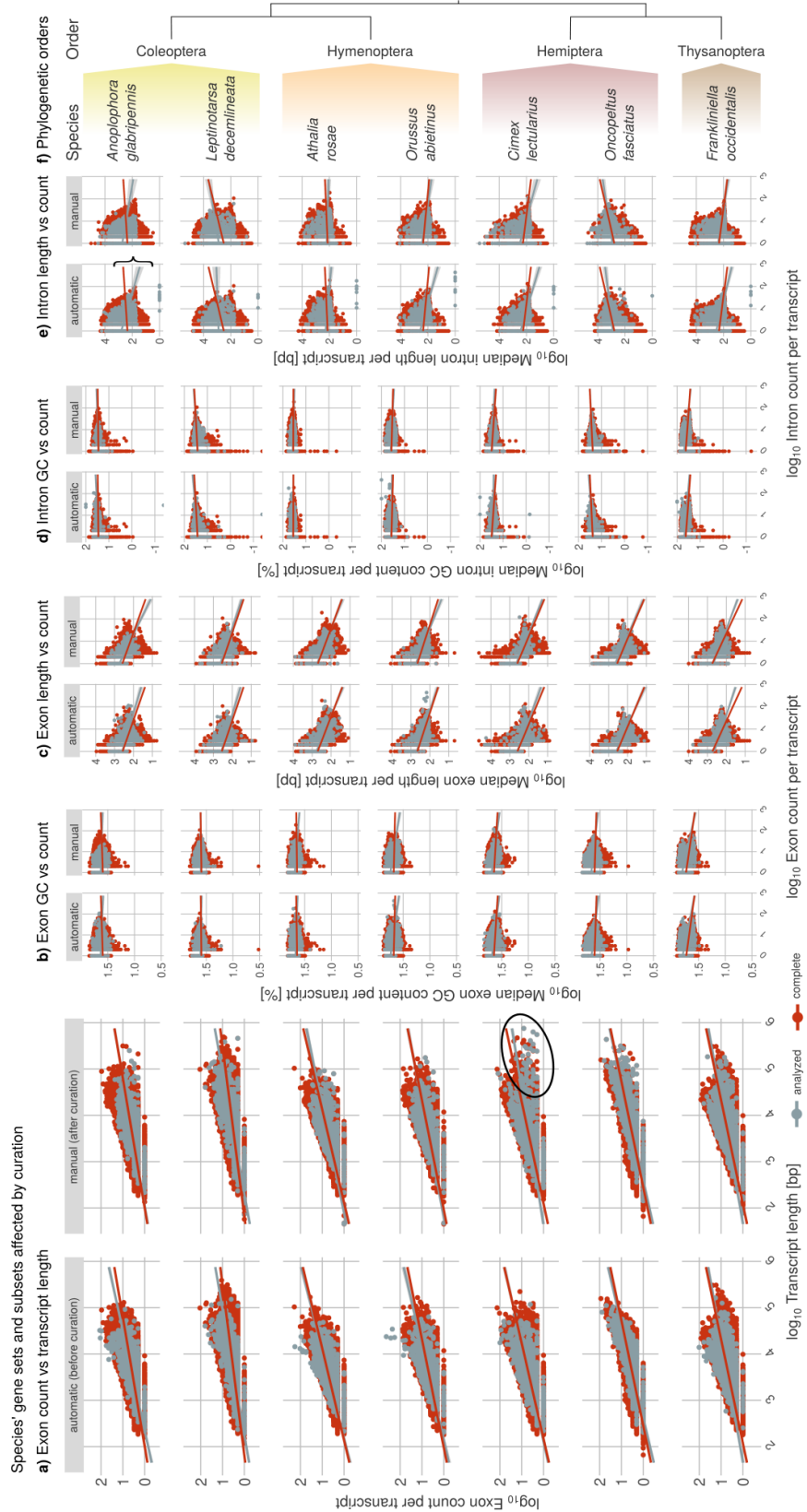
**Table III.1 – Overview of basic parameters.**

Basic data on assemblies and manual curation actions for each species and selected parameter values of each 'analyzed' set regarding gene models that were manually curated ('manual') and their predecessors ('automatic'). Assembly size is given in Mb, transcript length and exon length in bp, and protein length in aa; exon count and exon length are given as (median) per transcript. 'a' is short for 'automatic', 'm' for 'manual'

Using correlative plots (section III.5), we observed that genes selected for manual curation (subsets termed ‘analyzed automatic’) generally show a similar distribution of exon count per transcript (pre-mRNAa) in relation to transcript length as the entire automatically predicted set (‘complete automatic’ sets) (Fig. III.2 a: left column; Supplementary Note C.1.3). Likewise, the curated gene models (‘analyzed manual’ subsets) cover the data cloud of annotations in the complete official gene set (OGS, non-redundant merge of automatic and manual models) (‘complete manual’ sets) for these parameters, although with a (minor) enrichment in curated models with few exons and large transcripts, notably in the Hemiptera (Fig. III.2 a: right column, example highlighted with black circle). Evidently, the subset of manually curated gene models within the i5k project is in general not biased to certain structural parameter values, compared to the entire set of genes within each OGS. This holds also for comparisons of count, GC content, and lengths of introns and exons (Fig. III.2 b–e). The sole exception, an inversion of the trend line (smoothed conditional mean) of intron count vs. intron length in ‘analyzed’ subsets in comparison to ‘complete’ sets in the Asian long-horned beetle *Anoplophora glabripennis*, indicates that gene models chosen for curation were not representative in this species for these features (Fig. III.2 e, black brace).

### 3.4 The effect of manual curation on structural parameters of gene repertoires

We analyzed five parameters to compare automatically generated and manually curated gene models: transcript length, protein length, exon count per transcript, and median exon and intron lengths per transcript (see section III.5). Following the expectations outlined in the introduction, it is predicted that manual curation leads to an increase of transcript, protein, intron, and exon length as well as intron/exon count that is more pronounced in gene models from large genomes than in those from small genomes. Our data provide some support for this: manual curation in the two species with the largest genome assemblies in our dataset, *Leptinotarsa decemlineata* (Coleoptera; 1,170.2 Mb) and *Oncopeltus fasciatus* (Hemiptera; 1,098.7 Mb), led to a slight increase in the counts of longer transcripts, of transcripts with more exons, and of longer proteins compared to the statistics inferred from the automatically generated gene set;



**Figure III.2 – Automatically generated and manually curated annotation comparison I.** (Continued on next page.)

**Figure III.2 – Automatically generated and manually curated annotation comparison I. (Continued)**

**a) Species' gene sets and subsets affected by curation.** Logarithmic display of transcript length vs exon count per transcript as dispersion coverage plot. The values of automatically generated models that overlap manually curated versions as well as their manually curated versions ('analyzed', gray) are distributed within the range of values of the complete gene model sets including these subsets ('complete', red). Note that the outer rims of the data clouds are not covered by 'analyzed' gene models, indicating that models with extreme values were generally not manually annotated, although there are exceptions (*e.g.*, several gene models with many exons in *A. rosae* and *O. abietinus* that were apparently removed or split into several curated models). Values are given for the longest transcript per gene.

**b) – e) Correlative trends of exon-intron structure variation.** Logarithmic display of exon/intron content per transcript vs median exon/intron GC content per transcript (**b, d**) and median exon/intron length per transcript (**c, e**) as dispersion coverage plots. Values are given for the longest transcript per gene.

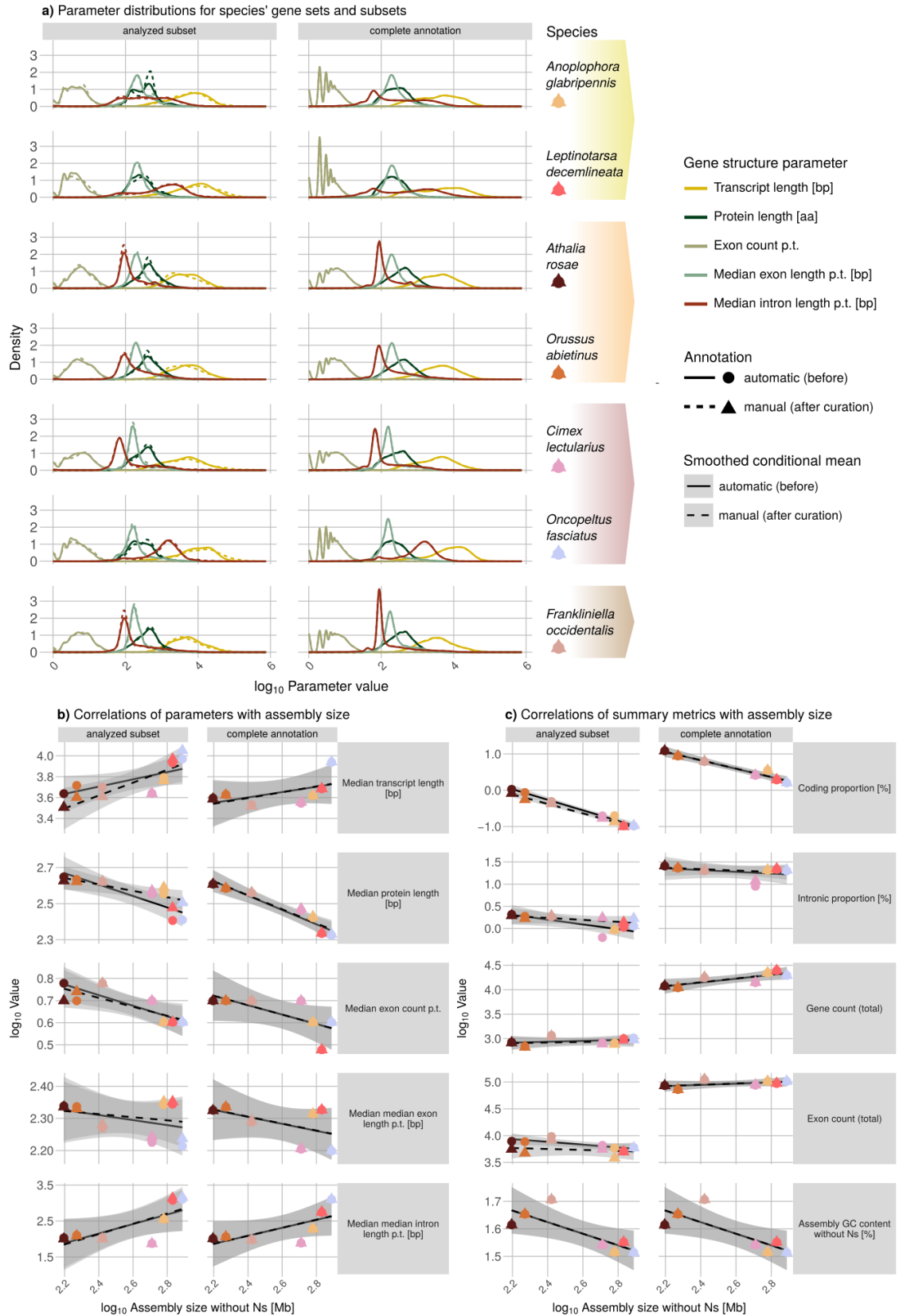
**f) Phylogenetic orders.** Color codes represent the insect orders Coleoptera (yellow), Hymenoptera (orange), Hemiptera (burgundy), and Thysanoptera (brown). The right side tree illustrates the order-level phylogenetic relationships (after MISOF *et al.*, 2014).

Lines represent the smoothed conditional mean for the 'analyzed' (gray) and 'complete' (red) set. Facets show the two annotations ('automatic' and 'manual') as columns and the seven species in rows (*Anoplophora glabripennis* [Coleoptera], *Athalia rosae* [Hymenoptera], *Frankliniella occidentalis* [Thysanoptera], *Leptinotarsa decemlineata* [Coleoptera], *Oncopeltus fasciatus* [Hemiptera], *Orussus abietinus* [Hymenoptera]). Black tags indicate special cases described in the text.

note that due to the small number of manually curated models, this effect is only visible in the subsets of 'analyzed' models (Fig. III.3 a and b, left column).

The most pronounced increase is found in median protein length (from 250 to ca. 320 aa; Table III.1). Contrariwise, in the two species with the smallest genome sizes in our sample, manual curation led to antagonistic results (Fig. III.3 b): after manual curation the repertoire-wide median exon count decreased in *Athalia rosae* (Hymenoptera; 163.8 Mb) and increased in *Orussus abietinus* (Hymenoptera; 201.2 Mb), while the median transcript length and median protein length decreased for both species. Note that the two hymenopterans do not only stand out by having the smallest genomes in our sample, but also because most of the curators involved in the manual curation of their annotations had only participated in at most two other projects of our sample (Supplementary Note C.1.3, Additional File C.2.4).





**Figure III.3 – Automatically generated and manually curated annotation comparison II.** (Continued)

**a) The effect of manual curation on structural parameters of gene repertoires.** Density distributions of four gene structure parameters per genome (semi-logarithmic): transcript length [bp], protein length [aa], exon count p.t., median exon length p.t. [bp], median intron length p.t. [bp].

**b) Correlations of assembly size** (without Ns, in Mb) to (in rows from top to bottom) median transcript length [bp], median protein length [aa], median exon count p.t., median median exon length p.t. [bp], and median median intron length p.t. [bp] (logarithmic).

**c) Correlations of assembly size** (without Ns, in Mb) to (in rows from top to bottom) coding proportion [%] (*i.e.*, the summed lengths of all coding sequences in the annotation in relation to genome size), intronic proportion [%] (*i.e.*, the summed lengths of all intronic sequences in the annotation in relation to genome size), total gene count, total exon count, and assembly GC content without ambiguity [%]. Most pronounced changes due to manual curation are found in transcript and protein lengths of the ‘analyzed’ set. Values are derived from the longest predicted transcript per gene.

Line types in (a) indicate the annotation type (‘automatic’, solid; ‘manual’, dashed), in (b) and (c) lines types indicate the smoothed conditional mean for each annotation type. Facets show the two sets (‘analyzed’ and ‘complete’) as columns and in (a) the seven species in rows (*Anoplophora glabripennis* [Coleoptera], *Athalia rosae* [Hymenoptera], *Cimex lectularius* [Hemiptera], *Frankliniella occidentalis* [Thysanoptera], *Leptinotarsa decemlineata* [Coleoptera], *Oncopeltus fasciatus* [Hemiptera], *Orussus abietinus* [Hymenoptera]). Taxonomic orders are color-coded according to Fig. III.2 f. aa: amino acids; bp: basepairs; Mb: megabasepairs; p.t.: per transcript.

### 3.5 Comparison to another automated annotation procedure

We investigate whether using a different tool to predict primary gene models leads to changes in our conclusions regarding the structural similarity to manually curated genes. For this, we compared the automatically predicted gene repertoires derived from a re-annotation using the BRAKER pipeline (HOFF *et al.*, 2016) with the original gene repertoires derived from application of the MAKER pipeline (HOLT and YANDELL, 2011) as used within i5k projects and with the MAKER-derived manually curated gene models (5), detailed results in Supplementary Note C.1.3). The genomes of three of the seven species in our sample (*A. rosae*, *O. abietinus* [both Hymenoptera], and *O. fasciatus* [Hemiptera]) were chosen for re-annotation with BRAKER due to data availability and

to represent a balance of small holometabolan and larger hemimetabolan genomes.

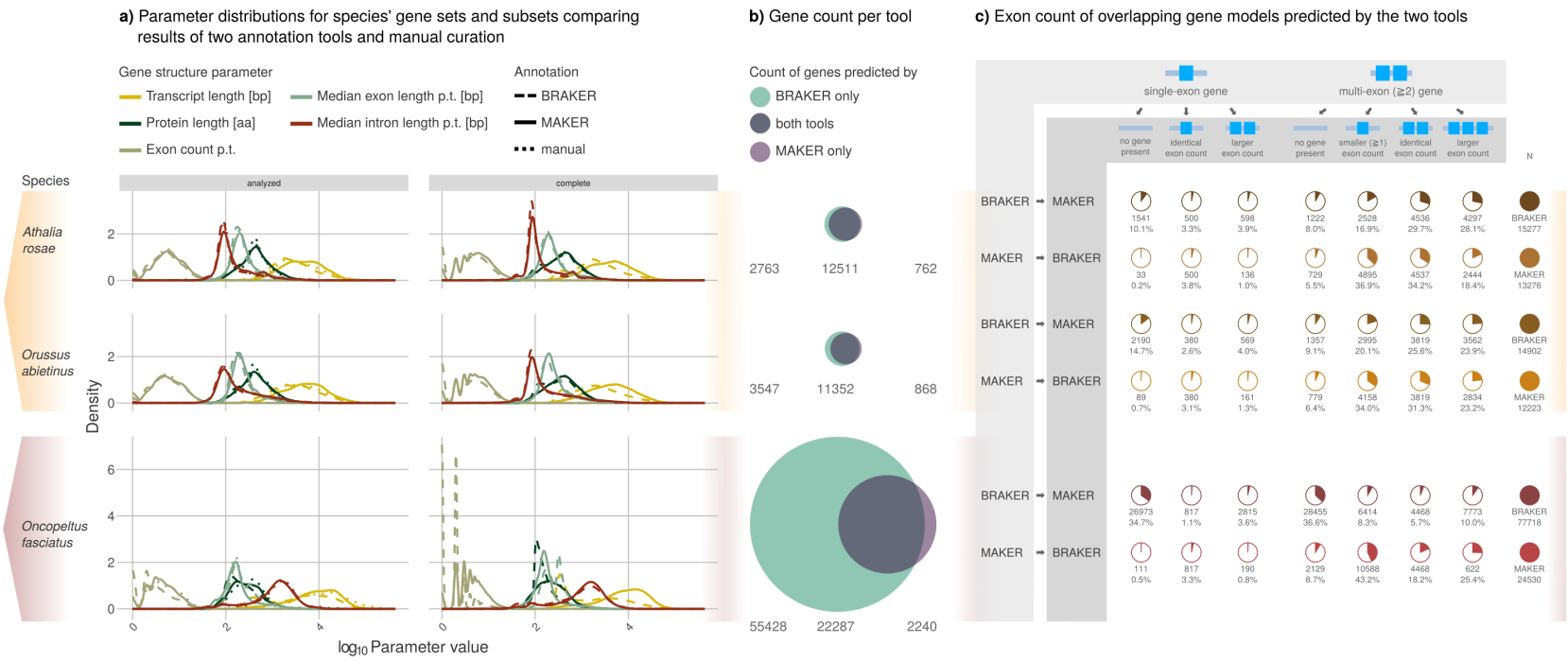
Where BRAKER and MAKER gene predictions overlap, the models are very similar in structure as measured by transcript, protein, intron, and exon length as well as exon count (Fig. III.4 a, Additional File C.2.5), although on average only a third of these models is predicted with identical exon count by both tools (Fig. III.4 c). Additionally, manually curated versions of these overlapping MAKER gene models are also in general structurally similar to both automatically predicted versions (Fig. III.4 a). However, it is conspicuous that BRAKER predicts a large number of genes not found by MAKER (Fig. III.4 b). About half of these BRAKER-only gene models consist of a single exon (Fig. III.4 c). Most of these single-exon BRAKER predictions were not well supported gene models judged from *ab initio*, RNA expression, and/or protein homology evidence (Supplementary Note C.1.3, Additional File C.2.6). Some differences between the annotations by both tools seem to correlate with genome assembly properties. Specifically, in the two small genomes we observed in some cases of BRAKER-only gene models that both BRAKER and MAKER predicted genes on both sides of the region it lies in. In the large genome of *O. fasciatus*, in contrast, the BRAKER-only models often occurred in poorly assembled scaffold regions comprised of very short contigs and many assembly gaps. It will be interesting to disentangle the effect of lineage-specificity from genome size, gene density, and assembly quality on automated gene prediction (see chapter III.4).

Despite the outlined systematic differences between the predicted gene sets of the compared tools, we find that overall structural parameters do not differ extensively from those of manually curated gene models. We thus claim that our comparisons and conclusions especially on the following previously reported trends (see below) are valid not only when using MAKER as the primary automated prediction pipeline.

### **3.6 Correlative trends of gene counts and coverages are not altered by manual curation**

ELLIOTT and GREGORY (2015) detected a statistically well supported negative correlation between coding proportion (*i.e.*, total length of respective DNA

Figure III.4 – Comparison of different annotation pipelines. (Continued on next page.)



**Figure III.4 – Comparison of different annotation pipelines.** (Continued)

**a) Parameter distributions for species' gene sets and subsets comparing results of two annotation tools and manual curation.** Semi-logarithmic display of 625 gene structure parameters (transcript length [bp], protein length [aa], exon count p.t., median exon length p.t., median intron length p.t.) for three annotation types (automatically predicted by BRAKER and MAKER, manually curated). BRAKER predicts more short genes with fewer exons than MAKER (median transcript length ca. 1,995 bp; ca. 3,980 bp with MAKER).

**b) Gene count per tool.** Euler diagrams of the number of genes predicted by MAKER and BRAKER, including overlapping gene models.

**c) Exon count of overlapping gene models predicted by the two tools.** Bi-directional comparison of gene complexity. Given in pies, counts and percentages is how many of the query genes, that have either one (single-exon gene) or more (multi-exon gene) exons, overlap what kind of gene in the target annotation (either no gene at all, one with the same exon count, or with more or less exons). The full pie charts give the total N for each automated annotation of each species.

The rows in all three figure parts refer to (from top to bottom) the species *Athalia rosae*, *Orussus abietinus* (both Hymenoptera), and *Oncopeltus fasciatus* (Hemiptera). The taxonomic orders are color-coded according to Fig. III.2 f. aa: amino acid; bp: basepairs; p.t.: per transcript.

sequences relative to genome size) and genome size among analyzed eukaryotic genomes, as well as a positive correlation between gene count and intronic proportion.

We compared the sets of automated and manually curated genes with respect to these correlations (coding proportion  $\propto 1/\text{genome size}$ , gene count  $\propto$  intronic proportion) (5). The analysis shows that the coding proportion in the genomes of our seven insect species is negatively correlated with genome size (Fig. III.3 c, first row). In conflict with the positive correlation between intronic proportion and genome size determined across eukaryotes by ELLIOTT and GREGORY (2015), we find no or only a weakly negative correlation of these parameters (Fig. III.3 c, second row). These trends are observed not only when all automatically generated models or the whole OGS are examined (Fig. III.3 c, right column), but also irrespective of whether only the subsets of manually curated gene models or their automatically inferred predecessors are taken into account (Fig. III.3 c, left column).

### 3.7 Correlative trends of gene composition found in uncurated and curated models

ZHU *et al.* (2009) studied the variation of intron-exon structures (*e.g.*, intron and exon lengths, ordinal position, GC content) in 13 eukaryotic genomes and consistently observed, among others, a negative correlation between exon and intron count per gene and both their respective lengths and GC content. Additionally, they detected that introns and exons shorter than average had either a higher or lower than average GC content. The GC content of introns/exons longer than average was intermediate. As explanations, ZHU *et al.* (2009) tentatively offered splicing requirements, a relation to the isochores structure of the genome, and “common factor(s) [...] shaping exons and introns” (ZHU *et al.*, 2009, p. 11).

Comparing the count of exons and introns per transcript with their median length and GC content per transcript (see section III.5), we can corroborate some of the proposed correlations (*e.g.*, increasing exon count correlates with decreasing median exon length), but we find mixed trends regarding others (*e.g.*, the relationship of intron count to median intron length) (Fig. III.2 b–e; detailed results in Supplementary Note C.1.3). We thus cannot corroborate all patterns of exon-intron composition regarding correlations between length and count as proposed by ZHU *et al.* (2009). However, these results hold for manually curated gene models (Fig. III.2 b–e, left columns) as well as their uncurated predecessors (Fig. III.2 b–e, right columns).

---

## Discussion

---

**E**RRORS IN AUTOMATED ANNOTATIONS can mislead analyses of gene family evolution (DENTON *et al.*, 2014), protein innovation rates (BÁNYAI and PATTHY, 2016), or the interpretation of gene function (PROSDOCIMI *et al.*, 2012). Algorithms and pipelines to annotate protein-coding genes are still unable to find and delineate all genes in a genome with perfect accuracy, thus their improvement is an important goal of comparative genomics. Meanwhile, manual curation can identify and correct many errors in automatically generated gene models, but this is costly. The aim of our investigation was to check whether automatic annotations are reliable enough to be the basis of (certain) comparative genomic analyses. Before we review this general problem, we discuss critical aspects of manual curation as well as the influence of manual curation on gene structure and of tool choice on the universality of our findings.

## 4.1 Reflections on manual curation

The curators' experiences and gene (family) specific knowledge, which are difficult to formalize, will have an influence on annotation quality after manual curation. However, the extent of their effects is difficult to assess, as several facets remain elusive: which person executed the curations indicated by a (shared) owner-tag, how many projects did this person participate in, and how much effort was expended to ensure the correctness of the annotations? Besides this, curation results may be difficult to incorporate into annotations. In complex cases, as exemplified in Fig. III.1 b, where part(s) of a gene lie on another scaffold and the annotation could be improved by re-assembly, thorough documentation of sequences and models to reflect their split nature is often infeasible. Manual revision of a given gene model can in such cases help to overcome assembly problems that currently inhibit correct automated prediction.

We observed in our study that assembly size and quality determines the frequency of split or merged gene models after manual curation insofar that in large genomes automatically predicted gene models are more likely to be split and thus require merging via manual curation (Table III.1). Setting the potential cause of this effect aside, it becomes clear that automatic annotation procedures need to be aware of assembly/genome size to be adjusted for the observed systematic error. This can be summed up as 'know your genome before you start annotating it'. Furthermore, given recently identified trends in gene structure (exon size and intron number) that distinguish different insect orders (PANFILO *et al.*, 2017), taxonomy could also inform starting assumptions for gene prediction.

It is open how often reading frame errors occur in automatically predicted gene models and how often they are corrected by manual curation. Given the importance of a correct coding sequence for downstream analysis of, for example, protein function (MUDGE and HARROW, 2016; PROSDOCIMI *et al.*, 2012), further research on this aspect is required. As exemplified in Fig. III.1 b, the correction of phase shifts may lead to the loss of ORF information for a gene model. The resulting dilemma forces a decision that will in either case lead to a partly incorrect gene model. Thus, the curator's experience as well as decision consistency can bias manually curated annotations, depending on



whether structural or functional correction is prioritized in resolving dilemmas. We observed that the MAKER pipeline as used by i5k permits non-canonical start codons (Supplementary Note C.1.4, Additional File C.2.3), which may lead to decision-problems if setting a canonical start codon leads to the loss of the first exon. A low dilemma frequency should be a major goal of annotation pipeline development as well as a benchmarking criterion.

## 4.2 The influence of manual curation on gene structure

We analyzed the effects of manual curation on five structural parameters (transcript and protein length, exon count, median length of introns and exons per transcript) in seven species of four insect orders. Note that models that were selected for manual curation are representative regarding structural parameters of the whole repertoire before and after curation (Fig. III.2), but probably not representative regarding functions or roles (*e.g.*, PANFILIO *et al.*, 2017). Furthermore, the small sample size and small taxonomic coverage as well as the non-normal distribution of parameter values limits statistical power of our analysis. However, we aimed for methodological uniformity when selecting the sample: for the seven analyzed species, genome sequencing and annotation had been performed consistently and gene sets before and after manual curation were available to us at the time of writing. All annotations and the derived statistics are based on more or less fragmented assemblies and thus can only be considered preliminary (as reviewed by FRANCIS and WÖRHEIDE, 2017). Since assessing the biological correctness of gene models remains difficult without a validated benchmark set (GUIGÓ *et al.*, 2000; REESE *et al.*, 2000), we based our comparisons on the assumption that manually curated gene models are correct. Encouragingly, we find that the overall gene structure of the gene models predicted by automated pipelines is very similar to that resulting from manual curation (Figs. III.2, III.3). Thus, the influence of manual curation on overall gene structure parameter distributions – insofar as curator changes can be documented in genome gff files – is rather small, even if individual models may be significantly changed in detail (*e.g.*, by adding an exon or by merging with another model, Table III.1). We conclude that automated annotation by MAKER is reliable regarding gene structure, when investigating structural parameters of the protein-coding gene repertoire in general.

### 4.3 The influence of tool choice

We investigated whether different automated annotation approaches yield structural predictions deviating to different degrees from manually curated models. For that goal, we compared the automated predictions generated by the MAKER pipeline within i5k to those generated by the BRAKER pipeline. The MAKER pipeline (HOLT and YANDELL, 2011) incorporates extrinsic evidence (see III.2 for implications) and is expected to yield conservative gene predictions, while the BRAKER pipeline (HOFF *et al.*, 2016) relies solely on intrinsic evidence (alignments of RNA sequences from the same species), which should result in more lineage-specific gene predictions. We observed systematic differences between the predictions of both tools, outlined in the following.

As expected, the BRAKER pipeline predicts consistently more protein-coding genes in the re-annotated genome assemblies (Fig. III.4). However, it is currently not possible to reliably differentiate between false positives and adequate additional gene annotations in this approach. It is suspicious that many of the BRAKER-predicted gene models do not overlap with gene models that are part of the MAKER-inferred gene set. Of these 'BRAKER-only' models, about half are single-exon genes. Some of these lie in regions devoid of supporting evidence, while others are flanked by gene models predicted by both pipelines. Such cases highlight the importance of making algorithm-specific cutoffs, thresholds, and gene discrimination criteria available to facilitate comparative quality assessment of gene prediction. As both tools predict genes not found by the other, we underline that the comparison of gene sets from different automated prediction routines and the elucidation of reasons for these differences may facilitate further algorithm developments. To date, the vast majority of sequenced insect genomes are for the Holometabola (Diptera and Hymenoptera). For hemimetabolous species (*e.g.*, Hemiptera and Thysanoptera), limited extrinsic evidence may indeed be a consideration until denser genome sequencing sampling is achieved, while large genome size will be an increasingly important challenge for hemimetabolous insect genome projects (HANRAHAN and JOHNSTON, 2011; PANFILIO and ANGELINI, 2017, p. 7).

Generally, apart from the set of BRAKER's single-exon gene predictions, the overall gene structures predicted by both tools are rather similar (Fig. III.4 a).

Also, parameter distributions of gene sets from either annotation tool are similar to those of manually curated gene models. We conclude that choice of tools for gene prediction is not of primary importance when the goal is a comparative analysis of gene structures.

#### 4.4 Elucidating trends using automatically predicted gene models

Given the general reliability of automatically predicted gene structures (by MAKER and BRAKER, likely also by other tools), we are confident that the observed corroborations and disparities from previously reported correlations of structural parameters are not an artifact of the annotation procedure. ELLIOTT and GREGORY (2015) reported a negative correlation of coding proportion and genome size, which was corroborated by FRANCIS and WÖRHEIDE (2017) and which we also observe in our data (Fig. III.3 c). This fits the hypothesis that genome size is mainly driven by increasing content of repetitive elements rather than by an increase of the gene count (GREGORY, 2005). On the other hand, ELLIOTT and GREGORY (2015) as well as FRANCIS and WÖRHEIDE (2017) found a correlation between intronic proportion and genome size, which we do not observe in our insect-specific data. However, the original observations were made across four (GREGORY, 2005) and six (FRANCIS and WÖRHEIDE, 2017) phyla of Eukaryota, respectively, and thus an insect-specific deviation cannot be excluded. If a different pattern of intron evolution can be corroborated in insects, assumptions on general genome evolution would have to be re-evaluated. Elucidating the exact relationships thus requires further investigation with a larger, denser, and more insect-specific taxon sample.

ZHU *et al.* (2009) reported that intron count negatively correlates with intron length and that exon count negatively correlates with exon length in a similar manner. Accordingly, the authors (ZHU *et al.*, 2009) proposed common selective pressures that shape both exon and intron structure within genes in similar directions. The similarity (or consistent presence) in correlations is not corroborated by our investigation in insect genomes (Fig. III.2 c and e), which might again point towards insect-specific modes of evolution. The vertebrate-biased taxon sample of ZHU *et al.* (2009) (nine vertebrates, two plants, one worm, and one insect, namely *D. melanogaster*) barely allowed generalizations for insects. While an amniote-specific positive correlation of intron and genome size has been shown and discussed in relation to avian powered flight (Q ZHANG and EDWARDS, 2012), it has yet to be unveiled whether intron size evolves neutrally in insects or is afflicted by other constraints than in amniotes. We conclude that elucidating commonalities, differences, and driving forces of

gene structure evolution requires a denser taxon sample but not necessarily manual curation: the trends are clearly visible in both automatically generated predictions (BRAKER and MAKER), as well as in manually curated models.

## 4.5 Conclusion

The impact of many potentially confounding factors affecting annotation quality remains to be assessed: automated predictions critically depend on assembly quality, algorithm criteria, training data, and the use of intrinsic and extrinsic evidence (FRANCIS and WÖRHEIDE, 2017; Q ZHANG and EDWARDS, 2012); manual curation quality, in contrast, depends not only on the underlying data (automatically generated models), but also on selection criteria (which genes are curated) and curator experience (who curated).

Detailed studies that are based on, for example, intron position homology (*e.g.*, KEANE and SEOIGHE, 2016; PANFILIO *et al.*, 2017) or protein function (*e.g.*, SIMAKOV *et al.*, 2015), depend on exact gene models, as achieved by manual curation. We show, however, that for the study of gene structure in a framework of comparative genomics, such as further research to explain the described insect-specific patterns, automated annotations can be considered sufficiently reliable. However, the goal remains to automatically infer biologically correct gene models and thus to eliminate the unresolved problems of automated annotation.

## 4.6 Future directions

We propose two questions to serve as guides for improving automated annotations of protein-coding genes and their usability.

(1) How can annotations be improved? Ideally, additional annotation information could be given as header of the gff file and should include declarations on data generation, prediction rationales (*e.g.*, alternative transcripts handled as individual genes), and applied criteria (*e.g.*, coding sequence begins with translation start codon). Distinguishing UTRs from coding sequence and the explicit usage of definitions will be worthwhile to consider (FRANCIS and

WÖRHEIDE, 2017; WILBRANDT *et al.*, 2017). Integrating annotations from different pipelines can further enhance accuracy (ZICKMANN and RENARD, 2015).

(2) Which criteria can be used to establish confidence in automatically generated gene models? Algorithm benchmarks to estimate accuracy depend on the provision of verified gene sets. For each gene model, confidence can be estimated with an index of the amount of supporting evidence (MISRA *et al.*, 2002), which is represented as the annotation edit distance in some pipelines (HOLT and YANDELL, 2011). Confidence should increase when characterizing criteria of coding sequence (MUDGE and HARROW, 2016) and protein structure (DRĂGAN *et al.*, 2016) are met. To assess the completeness of annotated gene sets, which is *a priori* difficult (FRANCIS and WÖRHEIDE, 2017), only few metrics have been proposed: the ratio of present single-copy orthologs (PARRA *et al.*, 2007; SIMÃO *et al.*, 2015) or conserved domains (DOHMEN *et al.*, 2016) expected in all organisms, the ratio of RNAseq reads mapping to the annotation (FRANCIS and WÖRHEIDE, 2017), and the ratio of intronic to intergenic regions (FRANCIS and WÖRHEIDE, 2017). Further research will show whether additional metrics and criteria can be applied to gauge the quality of gene annotations.

---

## Material and Methods

---

### 5.1 Data sample

WE ANALYZED ANNOTATIONS AND ASSEMBLIES of seven insect species of four orders (Coleoptera: *Anoplophora glabripennis*, *Leptinotarsa decemlineata*; Hemiptera: *Cimex lectularius*, *Oncopeltus fasciatus*; Hymenoptera: *Athalia rosae*, *Orussus abietinus*; Thysanoptera: *Frankliniella occidentalis*; Additional file C.2.1) that were sequenced and annotated within the i5k initiative (I5K CONSORTIUM, 2013). Additional file C.2.1 lists the sources of primary datasets.

### 5.2 Set preparation

We prepared four (sub)sets of data from the available annotations of each species. The set 'complete automatic' corresponds to the original annotation produced by the i5k initiative, while 'complete manual' refers to the official gene set (OGS, the non-redundant merge of manually curated gene models and unmodified predictions). 'Analyzed automatic' designates the subset of gene

models within the original annotation that overlap with the manually curated gene models ('analyzed manual').

### 5.3 Non-canonical start positions in MAKER predictions

A custom script was used to determine the frequencies of those amino acids, proteins encoded by gene models predicted by MAKER started with. The amino acid fasta files of the version BCM\_version\_0.5.3-Primary\_Gene\_Set were obtained for all seven species from <https://i5k.nal.usda.gov/content/data-downloads>. Results are displayed in Additional file C.2.3.

### 5.4 Curator experience analysis

Participating curators were identified by the owner-tag in their reviewed models in the OGS file. For each annotation project, we counted the number of participating curators and the number of gene models reviewed by them using a custom script. We assessed the number of projects any contributor participated in by hand. The count of curated gene models refers to the subsets 'analyzed manual' of each species.

### 5.5 Structural parameter and correlative trend analysis

Values of genic structural properties for each set of each species were obtained using COGNATE (WILBRANDT *et al.*, 2017) version 1.0 with default parameter (*i.e.*, COGNATE considered only the longest transcript per gene). Additional file C.2.7 holds all COGNATE result sets generated during the current study, except those of *F. occidentalis*, which are available upon request. Comparative plots were generated using custom scripts (R, R CORE TEAM, 2017; ggplot2, WICKHAM, 2009).



## 5.6 BRAKER-vs-MAKER analysis

We re-annotated the two hymenopteran species in our dataset, *Athalia rosae* and *Orussus abietinus*, as well as the hemipteran *Oncopeltus fasciatus* for a balancing of genome size, assembly quality, and curation efforts. We identified overlapping predictions lying on the same strand using a custom script. We spot-checked the annotation situation for single-exon genes using the Apollo web browser interface hosted by i5K@NAL. Values of genic structural properties for each set of each species were obtained as described above.



---

## Additional publication information

---

### 6.1 Availability of data and materials

All primary data (assemblies, annotations, RNA-seq reads) analyzed during this study are available from repositories as listed in Additional file C.2.1. All datasets generated during this study (except of *F. occidentalis*) are included in this published article and its supplementary information files (Additional files C.2.2–C.2.6) or available from the Dryad repository (ID<sup>III.1</sup>). The datasets of *F. occidentalis* generated during the current study are not publicly available due to ongoing research but are available from the corresponding author on reasonable request.

---

<sup>III.1</sup> A Dryad repository will be made available upon publication.

## 6.2 Funding

This work was partially funded by the German Research Foundation (DFG): MI 649/16-1 granted to BM, NI 1387/3-1 granted to ON, and SFB 680 Project A12 granted to KAP. The funding agencies did not influence the design of the study, the collection, analysis, and interpretation of data, or the manuscript writing.

## 6.3 Authors' contributions

JW conceived this study. BM and JW designed the study. JW and KAP acquired and analyzed the data. BM, KAP, and ON contributed to data analysis and interpretation. JW wrote the manuscript, BM, KAP, and ON critically revised it. All authors read and approved the final manuscript.

## 6.4 Acknowledgements

The authors are grateful to the i5k consortium and STEPHEN RICHARDS for championing insect genome sequencing at an unprecedented scale. We also thank MONICA POELCHEAU for patient help with all questions concerning i5k data. Invaluable is the joint effort of all curators who provided manually revised gene models. Furthermore, we thank DORITH ROTENBURG (coordinator of the *F. occidentalis* project) as well as YOLANDA CHEN and SEAN SCHOVILLE (coordinators of the *L. decemlineata* project) and ALEXIE PAPANICOLAOU for graciously providing data. JW thanks HANNES JÄKEL, JAN PHILIP OEYEN, MALTE PETERSEN, PANOS PROVATARIS, and TANJA ZIESMANN for fruitful discussions and ruthless feedback on the manuscript. Finally, BM, JW and ON thank the German Research Foundation for support of this study.

---

## Bibliography III

---

- AMID, C, LM REHAUME, KL BROWN, JG GILBERT, G DOUGAN, RE HANCOCK, and JL HARROW (2009). **Manual annotation and analysis of the defensin gene cluster in the C57BL/6J mouse reference genome.** *BMC Genomics* 10, p. 606. DOI: 10.1186/1471-2164-10-606 (cit. on p. 86).
- AMIT, M, M DONYO, D HOLLANDER, A GOREN, E KIM, S GELFMAN, G LEV-MAOR, D BURSTEIN, S SCHWARTZ, B POSTOLSKY, T PUPKO, and G AST (2012). **Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition.** *Cell Reports* 1.5, pp. 543–556. DOI: 10.1016/j.celrep.2012.03.013 (cit. on p. 83).
- BÁNYAI, L and L PATTHY (2016). **Putative extremely high rate of proteome innovation in lancelets might be explained by high rate of gene prediction errors.** *Scientific Reports* 6. DOI: 10.1038/srep30700 (cit. on pp. 85, 103).
- BRENT, MR (2005). **Genome annotation past, present, and future: How to define an ORF at each locus.** *Genome Research* 15.12, pp. 1777–1786. DOI: 10.1101/gr.3866105 (cit. on p. 84).
- (2008). **Steady progress and recent breakthroughs in the accuracy of automated genome annotation.** *Nature Reviews Genetics* 9.1, pp. 62–73. DOI: 10.1038/nrg2220 (cit. on p. 84).
- BURGE, C and S KARLIN (1998). **Finding the genes in genomic DNA.** *Current Opinion in Structural Biology* 8.3, pp. 346–354. DOI: 10.1016/S0959-440X(98)80069-9 (cit. on p. 84).
- CANTAREL, BL, I KORF, SMC ROBB, G PARRA, E ROSS, B MOORE, C HOLT, AS ALVARADO, and M YANDELL (2008). **MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Research* 18.1, pp. 188–196. DOI: 10.1101/gr.6743907 (cit. on p. 84).

- DENTON, JF, J LUGO-MARTINEZ, AE TUCKER, DR SCHRIDER, WC WARREN, and MW HAHN (2014). **Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies.** *PLOS Computational Biology* 10.12, e1003998. DOI: 10.1371/journal.pcbi.1003998 (cit. on pp. 84, 85, 103).
- DOHMEN, E, LPM KREMER, E BORNBERG-BAUER, and C KEMENA (2016). **DOGMA: domain-based transcriptome and proteome quality assessment.** *Bioinformatics* 32.17, pp. 2577–2581. DOI: 10.1093/bioinformatics/btw231 (cit. on p. 110).
- DRĂGAN, MA, I MOGHUL, A PRIYAM, C BUSTOS, and Y WURM (2016). **GeneValidator: identify problems with protein-coding gene predictions.** *Bioinformatics*, btw015. DOI: 10.1093/bioinformatics/btw015 (cit. on pp. 84, 92, 110).
- ELLIOTT, TA and TR GREGORY (2015). **What's in a genome? The C-value enigma and the evolution of eukaryotic genome content.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678, p. 20140331. DOI: 10.1098/rstb.2014.0331 (cit. on pp. 85, 88, 99, 101, 108).
- FRANCIS, WR and G WÖRHEIDE (2017). **Similar Ratios of Introns to Intergenic Sequence across Animal Genomes.** *Genome Biology and Evolution* 9.6, pp. 1582–1598. DOI: 10.1093/gbe/evx103 (cit. on pp. 84, 105, 108–110).
- GELFMAN, S and G AST (2013). **When epigenetics meets alternative splicing: the roles of DNA methylation and GC architecture.** *Epigenomics* 5.4, pp. 351–353. DOI: 10.2217/epi.13.32 (cit. on p. 83).
- GOODSWEN, SJ, PJ KENNEDY, and JT ELLIS (2012). **Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques.** *PLOS ONE* 7.11, e50609. DOI: 10.1371/journal.pone.0050609 (cit. on pp. 84, 85).
- GREGORY, TR (2005). **Synergy between sequence and size in Large-scale genomics.** *Nature Reviews Genetics* 6.9, pp. 699–708. DOI: 10.1038/nrg1674 (cit. on pp. 84, 108).
- GUIGÓ, R, P AGARWAL, JF ABRIL, M BURSET, and JW FICKETT (2000). **An Assessment of Gene Prediction Accuracy in Large DNA Sequences.** *Genome Research* 10.10, pp. 1631–1642. DOI: 10.1101/gr.122800 (cit. on pp. 84, 105).
- GUIGÓ, R, P FLICEK, JF ABRIL, A REYMOND, J LAGARDE, F DENOEUDE, S ANTONARAKIS, M ASHBURNER, VB BAJIC, E BIRNEY, R CASTELO, E EYRAS, C UCLA, TR GINGERAS, J HARROW, T HUBBARD, SE LEWIS, and MG REESE (2006). **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biology* 7.1, S2. DOI: 10.1186/gb-2006-7-s1-s2 (cit. on pp. 84, 85).
- HANRAHAN, SJ and JS JOHNSTON (2011). **New genome size estimates of 134 species of arthropods.** *Chromosome Research* 19.6, pp. 809–823. DOI: 10.1007/s10577-011-9231-6 (cit. on p. 106).
- HARROW, J, A NAGY, A REYMOND, T ALIOTO, L PATTHY, SE ANTONARAKIS, and R GUIGÓ (2009). **Identifying protein-coding genes in genomic sequences.** *Genome Biology* 10.1, p. 201. DOI: 10.1186/gb-2009-10-1-201 (cit. on pp. 84, 85).
- HOFF, KJ, S LANGE, A LOMSADZE, M BORODOVSKY, and M STANKE (2016). **BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and**

- AUGUSTUS**. *Bioinformatics* 32.5, pp. 767–769. DOI: 10.1093/bioinformatics/btv661 (cit. on pp. 98, 106).
- HOLT, C and M YANDELL (2011). **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects**. *BMC Bioinformatics* 12.1, p. 491. DOI: 10.1186/1471-2105-12-491 (cit. on pp. 98, 106, 110).
- HUFF, JT, D ZILBERMAN, and SW ROY (2016). **Mechanism for DNA transposons to generate introns on genomic scales**. *Nature* 538.7626, p. 533. DOI: 10.1038/nature20110 (cit. on p. 83).
- I5K CONSORTIUM (2013). **The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment**. *Journal of Heredity* 104.5, pp. 595–600. DOI: 10.1093/jhered/est050 (cit. on pp. 87, 111).
- KEANE, PA and C SEOIGHE (2016). **Intron Length Coevolution across Mammalian Genomes**. *Molecular Biology and Evolution* 33.10, pp. 2682–2691. DOI: 10.1093/molbev/msw151 (cit. on p. 109).
- KÖNIG, S, LW ROMOTH, L GERISCHER, and M STANKE (2016). **Simultaneous gene finding in multiple genomes**. *Bioinformatics* 32.22, pp. 3388–3395. DOI: 10.1093/bioinformatics/btw494 (cit. on p. 84).
- MISOFF, B *et al.* (2014). **Phylogenomics resolves the timing and pattern of insect evolution**. *Science* 346.6210, pp. 763–767. DOI: 10.1126/science.1257570 (cit. on pp. 87, 96).
- MISRA, S, MA CROSBY, CJ MUNGALL, BB MATTHEWS, KS CAMPBELL, P HRADECKY, Y HUANG, JS KAMINKER, GH MILLBURN, SE PROCHNIK, CD SMITH, JL TUPY, EJ WHITFIELD, L BAYRAKTAROGLU, BP BERMAN, BR BETTENCOURT, SE CELNIKER, AD de GREY, RA DRYSDALE, NL HARRIS, J RICHTER, S RUSSO, AJ SCHROEDER, S SHU, M STAPLETON, C YAMADA, M ASHBURNER, WM GELBART, GM RUBIN, and SE LEWIS (2002). **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review**. *Genome Biology* 3.12, research0083.1–83.22. DOI: 10.1186/gb-2002-3-12-research0083 (cit. on pp. 85, 86, 110).
- MUDGE, JM and J HARROW (2016). **The state of play in higher eukaryote gene annotation**. *Nature Reviews Genetics* 17.12, pp. 758–772. DOI: 10.1038/nrg.2016.119 (cit. on pp. 104, 110).
- PANFILIO, KA and D ANGELINI (2017). **By land, air, and sea: Hemipteran diversity through the genomic lens**. *Current Opinion in Insect Science* in press (cit. on p. 106).
- PANFILIO, KA, IMV JENTZSCH, JB BENOIT, D EREZYILMAZ, Y SUZUKI, S COLELLA, HM ROBERTSON, MF POELCHAU, RM WATERHOUSE, P IOANNIDIS, MT WEIRAUCH, DST HUGHES, SC MURALI, JH WERREN, CGC JACOBS, EJ DUNCAN, D ARMISÉN, BMI VREEDE, P BAA-PUYOULET, CS BERGER, Cc CHANG, H CHAO, MJM CHEN, YT CHEN, CP CHILDERS, AD CHIPMAN, AG CRIDGE, AJJ CRUMIÈRE, PK DEARDEN, EM DIDION, H DINH, H DODDAPANENI, A DOLAN, S DUGAN-PEREZ, CG EXTAVOUR, G FEBVVAY, M FRIEDRICH, N GINZBURG, Y HAN, P HEGER, T HORN, Ym HSIAO, EC JENNINGS, JS JOHNSTON, TE JONES, JW JONES, A KHILA, S KOELZER, V KOVACOVA, M LEASK, SL LEE, CY LEE, MR LOVEGROVE, HL LU, Y LU, PJ MOORE, MC MUNOZ-TORRES, DM MUZNY, SR PALLI, N PARISOT, L PICK, M PORTER, J QU, PN REFKI,

- R RICHTER, R RIVERA-POMAR, AJ ROSENDALE, S ROTH, L SACHS, ME SANTOS, J SEIBERT, E SGHAIER, JN SHUKLA, RJ STANCLIFFE, O TIDSWELL, L TRAVERSO, Mvd ZEE, S VIALA, KC WORLEY, EM ZDOBNOV, RA GIBBS, and S RICHARDS (2017). **Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome.** *bioRxiv*, p. 201731. DOI: 10 . 1101/201731 (cit. on pp. 84–86, 91, 104, 105, 109).
- PARRA, G, K BRADNAM, and I KORF (2007). **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 23.9, pp. 1061–1067. DOI: 10 . 1093/bioinformatics/btm071 (cit. on p. 110).
- PROSDOCIMI, F, B LINARD, P PONTAROTTI, O POCH, and JD THOMPSON (2012). **Controversies in modern evolutionary biology: the imperative for error detection and quality control.** *BMC Genomics* 13, p. 5. DOI: 10 . 1186 / 1471 - 2164 - 13 - 5 (cit. on pp. 85, 103, 104).
- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing (cit. on p. 112).
- REESE, MG, G HARTZELL, NL HARRIS, U OHLER, JF ABRIL, and SE LEWIS (2000). **Genome Annotation Assessment in *Drosophila melanogaster*.** *Genome Research* 10.4, pp. 483–501. DOI: 10 . 1101/gr . 10 . 4 . 483 (cit. on p. 105).
- SIEPEL, A, G BEJERANO, JS PEDERSEN, AS HINRICHS, M HOU, K ROSENBLUM, H CLAWSON, J SPIETH, LW HILLIER, S RICHARDS, GM WEINSTOCK, RK WILSON, RA GIBBS, WJ KENT, W MILLER, and D HAUSSLER (2005). **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Research* 15.8, pp. 1034–1050. DOI: 10 . 1101/gr . 3715005 (cit. on p. 88).
- SIMAKOV, O, T KAWASHIMA, F MARLÉTAZ, J JENKINS, R KOYANAGI, T MITROS, K HISATA, J BREDESON, E SHOGUCHI, F GYOJA, JX YUE, YC CHEN, RM FREEMAN, A SASAKI, T HIKOSAKA-KATAYAMA, A SATO, M FUJIE, KW BAUGHMAN, J LEVINE, P GONZALEZ, C CAMERON, JH FRITZENWANKER, AM PANI, H GOTO, M KANDA, N ARAKAKI, S YAMASAKI, J QU, A CREE, Y DING, HH DINH, S DUGAN, M HOLDER, SN JHANGIANI, CL KOVAR, SL LEE, LR LEWIS, D MORTON, LV NAZARETH, G OKWUONU, J SANTIBANEZ, R CHEN, S RICHARDS, DM MUZNY, A GILLIS, L PESHKIN, M WU, T HUMPHREYS, YH SU, NH PUTNAM, J SCHMUTZ, A FUJIYAMA, JK YU, K TAGAWA, KC WORLEY, RA GIBBS, MW KIRSCHNER, CJ LOWE, N SATOH, DS ROKHSAR, and J GERHART (2015). **Hemichordate genomes and deuterostome origins.** *Nature* 527.7579, pp. 459–465. DOI: 10 . 1038/nature16150 (cit. on p. 109).
- SIMÃO, FA, RM WATERHOUSE, P IOANNIDIS, EV KRIVENTSEVA, and EM ZDOBNOV (2015). **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 31.19, pp. 3210–3212. DOI: 10 . 1093 / bioinformatics/btv351 (cit. on p. 110).
- ŠKUNCA, N, A ALTENHOFF, and C DESSIMOZ (2012). **Quality of Computationally Inferred Gene Ontology Annotations.** *PLOS Computational Biology* 8.5, e1002533. DOI: 10 . 1371/journal.pcbi . 1002533 (cit. on p. 87).



- TREANGEN, TJ and SL SALZBERG (2012). **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nature Reviews Genetics* 13.1, p. 36. DOI: 10.1038/nrg3117 (cit. on p. 84).
- WICKHAM, H (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York (cit. on p. 112).
- WILBRANDT, J, B MISOF, and O NIEHUIS (2017). **COGNATE: comparative gene annotation characterizer.** *BMC Genomics* 18.1, p. 535. DOI: 10.1186/s12864-017-3870-8 (cit. on pp. 110, 112).
- YANDELL, M, AM BAILEY, S MISRA, S SHU, C WIEL, M EVANS-HOLM, SE CELNIKER, and GM RUBIN (2005). **A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome.** *Proceedings of the National academy of Sciences of the United States of America* 102.5, pp. 1566–1571 (cit. on pp. 84, 85).
- YANDELL, M and D ENCE (2012). **A beginner's guide to eukaryotic genome annotation.** *Nature Reviews Genetics* 13.5, pp. 329–342. DOI: 10.1038/nrg3174 (cit. on p. 84).
- YANDELL, M, CJ MUNGALL, C SMITH, S PROCHNIK, J KAMINKER, G HARTZELL, S LEWIS, and GM RUBIN (2006). **Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes.** *PLoS Comput Biol* 2.3, e15. DOI: 10.1371/journal.pcbi.0020015 (cit. on p. 88).
- ZDOBNOV, EM and P BORK (2007). **Quantification of insect genome divergence.** *Trends in Genetics* 23.1, pp. 16–20. DOI: 10.1016/j.tig.2006.10.004 (cit. on p. 88).
- ZHANG, MQ (2002). **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 3.9, pp. 698–709. DOI: 10.1038/nrg890 (cit. on p. 83).
- ZHANG, Q and SV EDWARDS (2012). **The Evolution of Intron Size in Amniotes: A Role for Powered Flight?** *Genome Biology and Evolution* 4.10, pp. 1033–1043. DOI: 10.1093/gbe/evs070 (cit. on pp. 83, 108, 109).
- ZHANG, X, J GOODSELL, and RB NORGRÉN (2012). **Limitations of the rhesus macaque draft genome assembly and annotation.** *BMC Genomics* 13, p. 206. DOI: 10.1186/1471-2164-13-206 (cit. on p. 84).
- ZHU, L, Y ZHANG, W ZHANG, S YANG, JQ CHEN, and D TIAN (2009). **Patterns of exon-intron architecture variation of genes in eukaryotic genomes.** *BMC Genomics* 10, p. 47. DOI: 10.1186/1471-2164-10-47 (cit. on pp. 88, 102, 108).
- ZICKMANN, F and BY RENARD (2015). **IPred - integrating ab initio and evidence based gene predictions to improve prediction accuracy.** *BMC Genomics* 16, p. 134. DOI: 10.1186/s12864-015-1315-9 (cit. on p. 110).



---

# DESCRIPTIVE ASSESSMENT OF STRUCTURAL FEATURES WITHIN HYMENOPTERAN GENE REPERTOIRES

---

The following text is the author's version of the contribution to the supplementary notes of an article intended for publication in *Current Biology*:

**OEYEN JP et al. (2018) The genomes of *Athalia* and *Orussus*. *Current Biology*, in prep.**

The project is led by JAN PHILIP OEYEN and comprises many individual research projects by various authors addressing aspects of the genome biology of *Athalia rosae* and *Orussus abietinus*.

Authors' contributions to the following original article part:  
Analyses, figures, manuscript design, and writing: JW.



---

## Introduction

---

**A** BASIC STEP IN COMPARATIVE GENOMICS is the analysis of genomic and genetic structural features like genome size, protein-coding gene length, and exon count. The description of genome and gene structure allows exploring the relationships between these structural properties and other parameters with the aim of identifying drivers of these structural features. The descriptive and comparative approach is crucial to facilitate research on the basic genome parameters of a species, to range in measured magnitudes with known parameters from the same taxon or clade, and to build hypotheses regarding genome dynamics and evolution.

A seminal study in this direction was published by ELLIOTT and GREGORY (2015), which was a meta-analysis of genome and gene summary statistics across more than 520 species. Considering this large number of genomes and the phylogenetic relationships of the respective species allowed the authors the robust detection of statistical trends and of genomic differences between organismic kingdoms; the study shows, *i.a.*, that genome size is positively correlated to gene number and negatively correlated to the proportion of coding sequences. However, the reliance on published results for their study could

have influenced the statistical power of tests due to missing data and/or inconsistent term usage. Furthermore, ELLIOTT and GREGORY (2015) did not analyze genomic differences within the organismic kingdoms, thus leaving open whether there are variations of genome structure relationships between the taxonomic orders of analyzed animals, for example.

For insects, there are few in-depth studies of comparative genomics. Most studies focus on functional gene annotation and compare functional repertoires, while those concerned with structural differences are restricted to a subset of genome properties. For example, it has been shown that the genomes of twelve insect species of several orders are more diverse than genomes of vertebrates (covering a comparable range divergence times) in terms of protein-coding gene arrangements and orthologous protein sequence identity (ZDOBNOV and BORK, 2007). A comparison of base composition revealed that the honey bee (*Apis mellifera*), the solitary parasitoid jewel wasp *Nasonia vitripennis*, and the ants *Pogonomyrmex barbatus* and *Linepithema humile* show a tendency for genes to occur in GC-poor regions of the genome, while no such tendency appears in the leaf cutter ant *Atta cephalotes* (SUEN *et al.*, 2011). The overall genome composition is considered to be quite conserved within twelve closely related *Drosophila* species, judging from genome size and the amount of coding and intronic sequences (CLARK *et al.*, 2007). Overall, little is known about the structure of protein-coding genes in insects, what can be considered to be 'within a usual range' and which measured parameters of genomic and gene structure are notable exceptions.

Here, two newly sequenced genomes of non-apocritan "symphytans", namely the parasitoid wood wasp *Orussus abietinus* and the turnip sawfly *Athalia rosae*, are examined with respect to the details of gene structure within their repertoires of protein-coding genes. This study is embedded in a large-scale project (OEYEN *et al.*, in prep.), which presents the two genomes and examines them from various angles (*e.g.*, diversity of transposable elements, presence of key protein families) with regard to the major transition from phytophagous to parasitoid Hymenoptera. Parasitoid Hymenoptera are highly diverse, while "symphytan" lineages are rather species-poor. Here, *O. abietinus* and *A. rosae* are highly interesting species: the former belongs to the parasitoid family Orussidae, the sister lineage to the also primary parasitoid Apocrita. *A. rosae* represents the phytophagous Eusymphyta, a lineage that diverged ca. 40 million years earlier from the ancestor of Apocrita and Orussidae.

I implemented analyses to identify patterns of changes in gene structure, protein family composition, and other aspects that might have played a role in facilitating or inhibiting diversification of the scrutinized lineages. Consequently, basic differences of gene structure between the early-divergent hymenopterans *A. rosae* and *O. abietinus* and the more derived, partially eusocial, Apocrita are elucidated and discussed with respect to the evolutionary relationships.

To describe and compare the genome and protein-coding gene structure of the parasitoid wood-wasp *O. abietinus* and the turnip sawfly *A. rosae*, I extracted and analyzed numerous (standard) parameters from their structural protein-coding gene annotations and compared them to our analysis of published annotations of ten apocritan hymenopterans and one outgroup insect species.





---

## Methods

---

### 2.1 Species

**T**HE PROTEIN-CODING GENE ANNOTATIONS (gffs) of 13 species were analyzed (see also Tab. IV.1). These species comprise twelve hymenopterans (two species of “Symphyta”, the following ten are apocritan species): *Orussus abietinus* (parasitoid wood-wasp), *Athalia rosae* (turnip sawfly), *Apis mellifera* (honey bee), *Bombus terrestris* (large earth bumblebee), *Megachile rotundata* (leafcutter bee), *Dufourea novaeangliae* (sweat bee), *Lasioglossum albipes* (white-footed sweat bee), *Acromyrmex echinator* (Panamanian leafcutter ant), *Camponotus floridanus* (Florida carpenter ant), *Harpegnathos saltator* (Jerdon’s jumping ant), *Polistes dominula* (European paper wasp), and *Nasonia vitripennis* (jewel wasp); and one outgroup species: *Tribolium castaneum* (red flour beetle). An overview of the used files and respective references can be found in Supplementary table D.1.1 a.

The genomes of *A. rosae* and *O. abietinus* have been sequenced for the first time by the i5k initiative (15K CONSORTIUM, 2013). Analysis of the assemblies was

Species	Assembly size	L75	L90pcG	Gene count
<i>Apis mellifera</i>	234.1	152	373	15,314
<i>Bombus terrestris</i>	248.7	14	1,708	10,400
<i>Megachile rotundata</i>	272.7	106	161	12,770
<i>Dufourea novaeangliae</i>	291.0	88	150	12,453
<i>Lasioglossum albipes</i>	336.5	352	991	13,421
<i>Acromyrmex echinator</i>	297.5	175	322	17,271
<i>Camponotus floridanus</i>	234.9	365	1,884	17,013
<i>Harpegnathos saltator</i>	296.8	344	1,505	18,518
<i>Polistes dominula</i>	208.0	86	130	11,815
<i>Nasonia vitripennis</i>	295.8	215	184	13,185
<i>Orussus abietinus</i>	201.2	68	126	10,959
<i>Athalia rosae</i>	163.8	79	127	11,894
<i>Tribolium castaneum</i>	165.9	8	10	12,863

**Table IV.1 – Species sample.** The species are ordered according to phylogenetic relationships, see Fig. IV.1; this figure also includes a visualization of assembly sizes. Assembly size (including Ns/gaps) is given in Mbp (megabasepairs).

committed to our working group; thus, evaluating the quality of assemblies and annotations focuses on these two species, taking the remaining ones as given.

## 2.2 Analysis

I counted and calculated more than 280 metrics of the gene annotations using COGNATE v1.0 (WILBRANDT *et al.*, 2017) with default parameters, *i.e.*, analyzing only the longest transcript for each gene. For this, also the genomic fasta file for each species was used. For an overview of the measured metrics, I refer to WILBRANDT *et al.* (2017). An overview of summary values, medians, and component sizes is given in Supplementary table D.1.1. The complete COGNATE result sets for all species can be found in Supplementary file D.1.2.

Component sizes, *i.e.*, the combined length (in Mbp) of all coding and intron sequences within one genome, obtained by COGNATE, were compared with component sizes of transposable elements and other repeats (values obtained from TE annotation by MALTE PETERSEN, pers. comm., March 2018).

The features of one, the longest, annotated transcript were analyzed for all predicted protein-coding genes of the 13 species. These features are, among others, GC contents and lengths of transcripts themselves, exons and introns as well as the number of introns and exons per transcript and the coverage and density of exons and introns on transcripts. It was suggested by YANDELL *et al.* (2006) to measure intron lengths, exon lengths and intron density as a proxy for gene structure. Thus, a special focus was put on these metrics, although differing from the suggested approach by calculating intron density as the count of introns of a transcript per length of the same transcript in base pairs (instead of protein length).

Features of transcripts, CDSs, exons, and introns were evaluated in density plots, hereafter called distributions. Furthermore, several parameters were plotted against each other (*e.g.*, GC content versus length, see below). Plots were obtained by custom R scripts using the ggplot2 library (WICKHAM, 2009) and the wesanderson library<sup>IV.1</sup>.

---

<sup>IV.1</sup> GitHub: KARTHIK *et al.*. 2018. R package wesanderson, <https://github.com/karthik/wesanderson>. Last accessed 30 March 2018.



---

## Results

---

### 3.1 Assembly features

**T**HE ASSEMBLIES of the *A. rosae* and *O. abietinus* genomes have low values of L75, L90 and L90pcG compared to the median of all species for these metrics. This indicates a high assembly quality. L75 and L90 represent the number of scaffolds, beginning to count at the longest, that are required to encompass at least 75 % and 90 % of the total assembly length, respectively. The smaller this number is, the less fragmented is the assembly. Similarly, L90pcG is the number of scaffolds, beginning to count at the longest, that bear at least 90% of all annotated protein-coding genes on them.

The annotation of protein-coding genes for both genomes does not have flaws that could be identified via intron length distribution skews. Introns are expected to occur not only in multiples of three bases (three bases are one codon triplet;  $3n$  introns), but also with a similar probability in multiples of three plus one or two bases; finding more or less  $3n$  introns than expected hints to systematic errors in mistaking exons for introns or introns for exons,

respectively (ROY and PENNY, 2007). The respective data can be found in Supplementary table D.1.1 b.

## 3.2 Component sizes

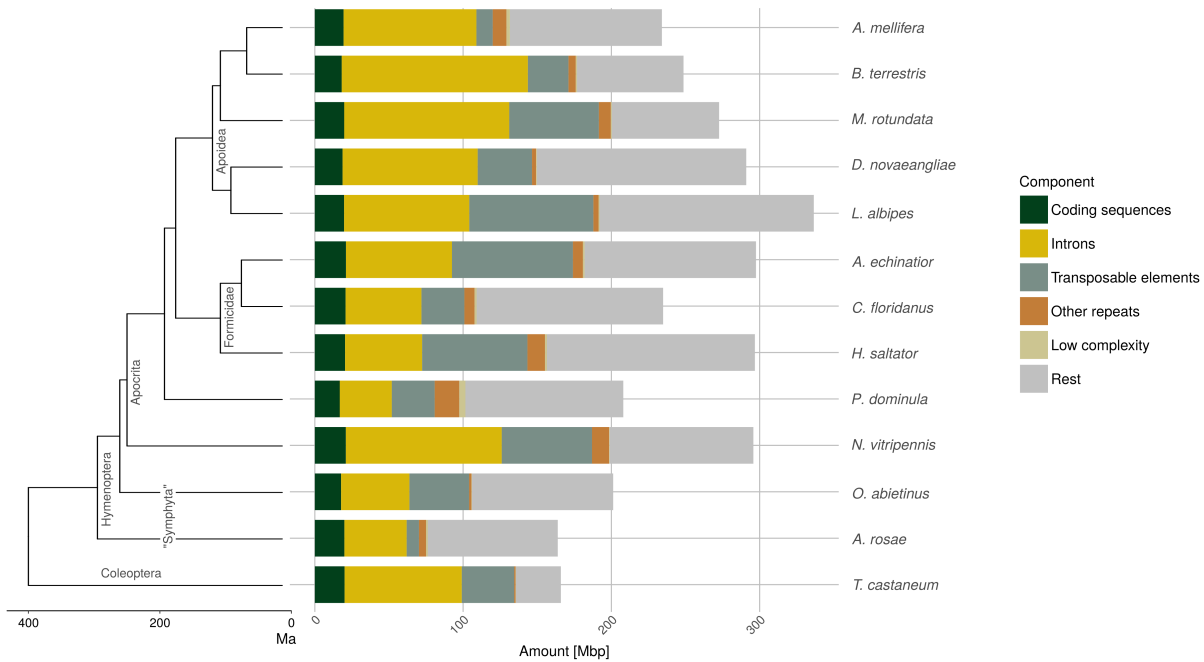
The genomes of *A. rosae* and *O. abietinus* are smallest among the compared Hymenoptera and harbor less intronic and repetitive sequences in absolute numbers (Fig. IV.1). Conversely, the relative amount of coding sequences is large in both genomes compared to the remaining species and the relative amount of introns is close to the overall median (Supplementary table D.1.1 c). The protein-coding amount is in all species in the comparison very similar (between 16 and 20 Mbp). Further correlations to assembly size (excluding gaps/Ns) are illustrated in Fig IV.2, where values of *P. dominula*, *B. terrestris*, and the Formicidae are often outliers. In some cases *O. abietinus* closely resembles the outgroup (*T. castaneum*), while *A. rosae* appears to be relatively extreme in some instances (e.g., the absolute amount of transposable elements being the smallest of all considered species, 10 % of that of *A. echinator*; Fig. IV.1).

Apparently, introns as well as transposable and other repetitive elements contributed to an increase in genome size within the Apocrita.

## 3.3 Structural features

### 3.3.1 Transcripts

The two non-apocritan Hymenoptera *O. abietinus* and *A. rosae* have decidedly longer transcripts (genomic transcript length, includes introns and exons) than Apocrita: the median genomic transcript length is almost doubled compared to the apocritan median (Fig. IV.3). Looking at the distributions of transcript lengths (Fig. IV.4, Supplementary figure D.1.3, p.1), Halictidae (*D. novaeangliae*, *L. albipes*) and Formicidae (*A. echinator*, *C. floridanus*, *H. saltator*) are distinctive because of their pronounced bimodal distribution, peaking around 250 and 2500 bp. The transcript lengths of *A. mellifera* are similarly distributed, although with considerably less transcripts of ca. 250 bp length. The distribution of transcript

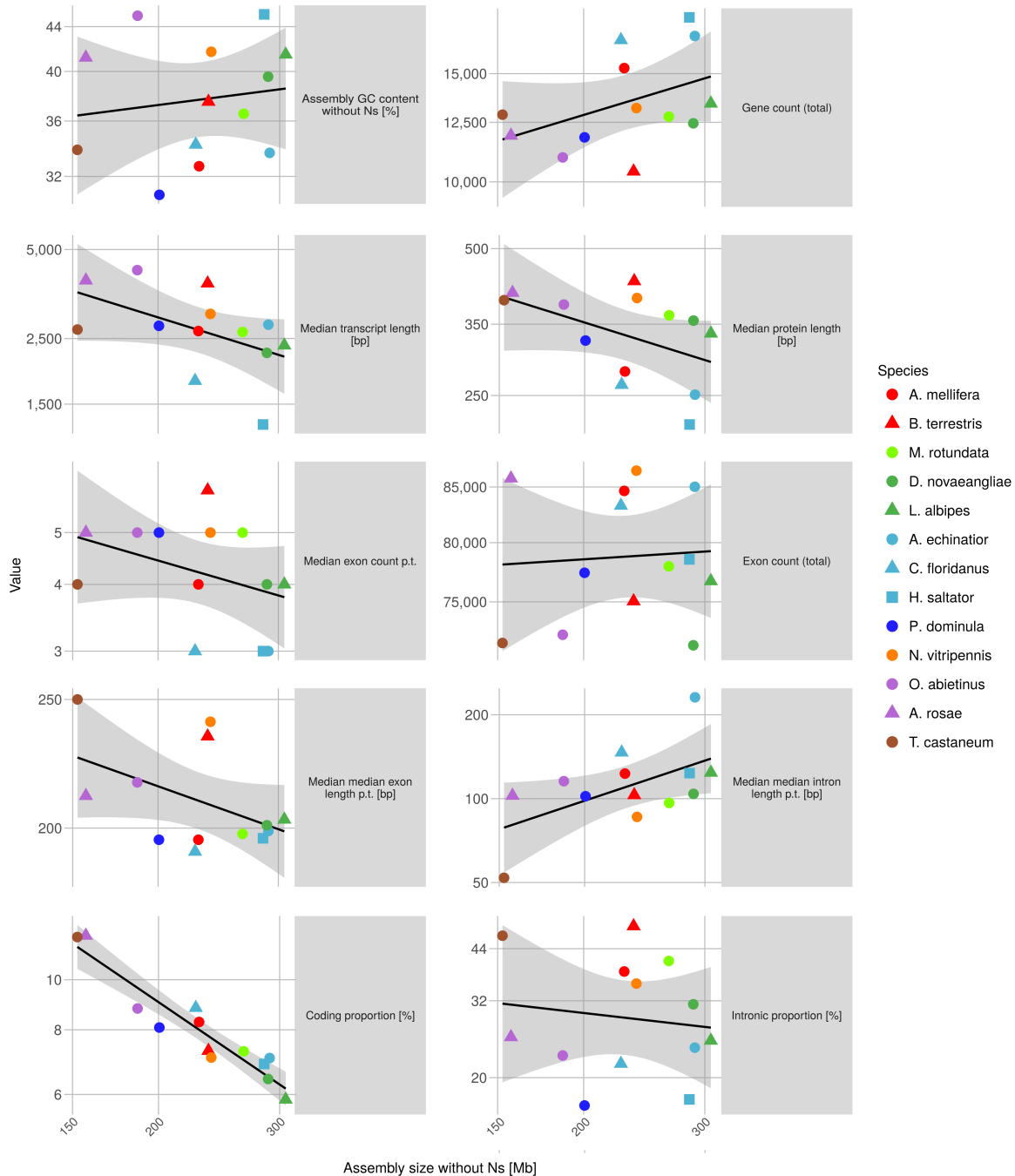


**Figure IV.1 – Genome and component sizes.**

A summary of the annotations of protein-coding genes and transposable elements. Amounts [Mbp] refer to the cumulative sequence length of the genomic regions annotated as one of the components. The coding component (amount of CDSs per species) is surprisingly constant across species, while the amounts of introns, transposable elements, and the rest contribute most to genome size.

The phylogenetic tree illustrates the relationships between the analyzed species, branch lengths correspond to divergence time. A few taxonomic labels are indicated for reference from the text.

Ma: Million years ago; Mbp: megabasepairs



**Figure IV.2 – Trends of assembly size and genomic and gene structure parameter correlations**

Several genome parameters and gene structure parameters are compared in relation to assembly size (size here excludes gaps/Ns, to the effect that, for example, *T. castaneum* appears to be smaller than *A. rosae*).

Coding and intronic proportion refer to the component sizes indicated in Fig. IV.1. Black line and grey are indicate smoothed conditional mean. The scales are transformed to log-scales, the values are not. Species coloration follows phylogenetic relationships, see Fig. IV.1

Mbp: megabasepairs



lengths of *P. dominula* and *N. vitripennis* are both unimodal and very similar with most transcripts being ca. 2510 bp long. *A. rosae* and *O. abietinus* appear to have more long transcripts (around 5600 bp) than any other species (Fig. IV.4). Thus, a decrease of gene length is observed within Apocrita, where *B. terrestris* is an exception. This pattern is reflected in protein length distributions, which are also right-skewed, but more narrow and show a less pronounced bimodality (Supplementary figure D.1.3, p.2). The median protein length of *A. rosae* (406 aa) and *O. abietinus* (384 aa) is considerably longer than the hymenopteran (345.5 aa) and apocritan (329.5 aa) as well as overall (356 aa) median; actually, *A. rosae* has the second largest median protein length in our sample (largest: *B. terrestris*, 429 aa).

### 3.3.2 Exons and CDSs

The distribution of median exon length per transcript (Supplementary figure D.1.3, p.10) is very similar across all species (max around 180 bp); only *A. echinator* deviates slightly with an increased number of transcripts with longer median exon length (Fig. IV.4). Median exon and CDS length, both individual and the median per transcript, is in *A. rosae* and *O. abietinus* very close to all species-summarizing medians (apocritan, hymenopteran, and overall), while the median coverage and density of exons and CDSs per transcript are lower compared to the other species. The median exon count is between 3 and 6 in all species. All considered ants have a median exon count of 3, while no clear pattern of count and phylogenetic relationship can be discerned for the remaining species. Note, however, that the five earlier-divergent species have very few transcripts with only one exon, opposing all later-diverged species (except *B. terrestris*) (indicated by solid-line rectangles in Fig. IV.4).

Most transcripts (averaged Q1 and Q3, respectively) have an added exon length of 695–2440 bp, a median exon length of 159–306 bp (both ranges are slightly lower for CDSs), harbor 2–7 exons, and have an exon density of 0.0009–0.003 (Supplementary file D.1.1 e). *A. rosae* has a mean coding gene size (introns and exons) of 6.5 kbp (median 3.9 kbp), and a mean intron size of 0.6 kbp (median 0.1 kbp). *O. abietinus* has a mean coding gene size (introns and exons) of 7.3 kbp (median 4.2 kbp), and a mean intron size of 0.8 kbp (median 0.1 kbp).

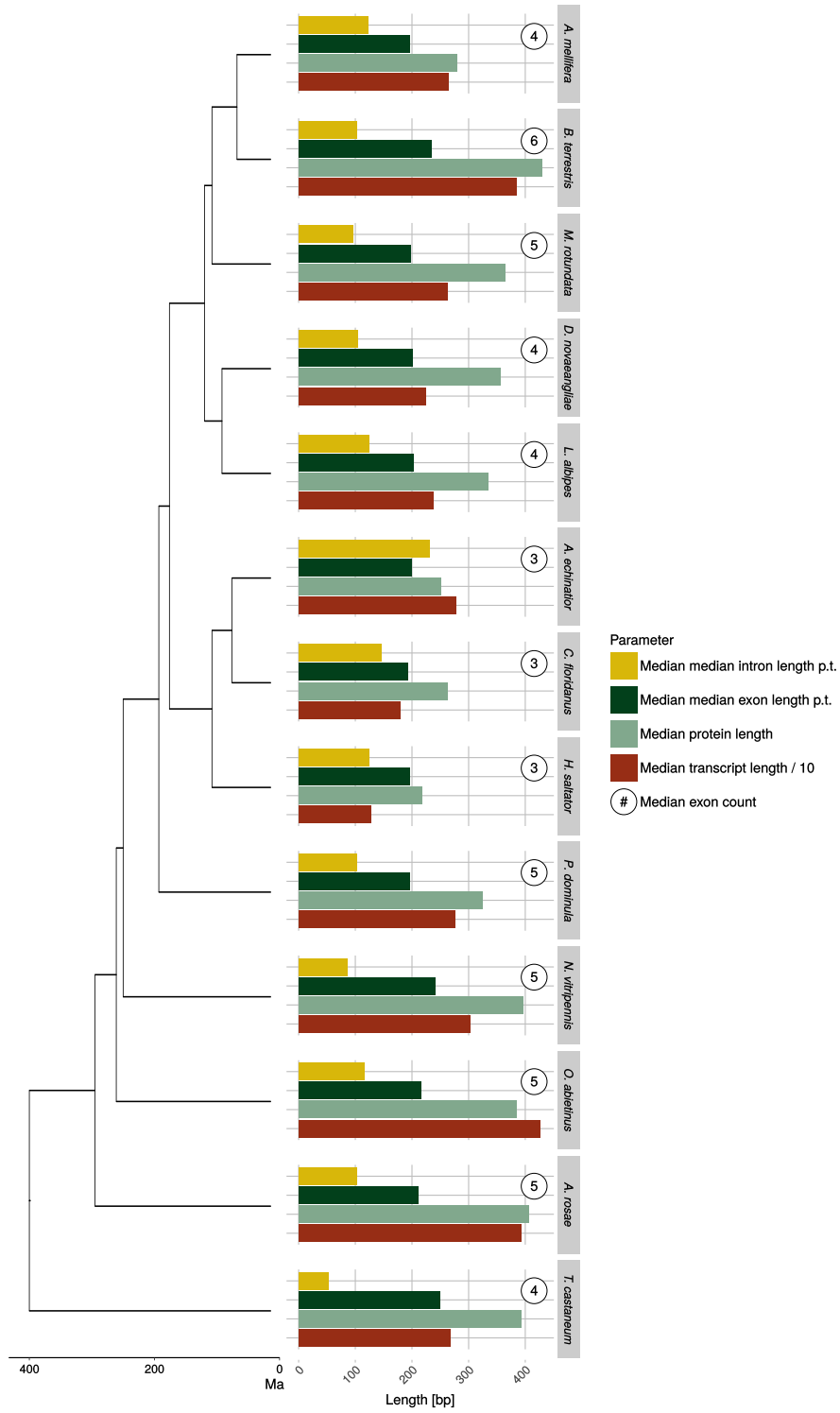


Figure IV.3 – Comparison of median lengths and counts of gene elements.  
(Continued on next page)

**Figure IV.3 – Comparison of median lengths and counts of gene elements.**

(Continued)

For each species and for each of their genes (as analyzed by COGNATE, longest transcript per gene), the median of each of the four length parameters and the exon count was recorded; then, the median for all transcript was calculated and plotted here.

The phylogenetic tree illustrates the relationships between the analyzed species, branch lengths correspond to divergence time.

bp: basepairs; Ma: Million years ago

**3.3.3 Introns**

*A. rosae*'s and *O. abietinus*'s median intron lengths are small compared to those of Apocrita, while *T. castaneum* has very short introns (in a notable bimodal distribution of intron lengths with much less long introns than any hymenopteran) (Fig. IV.3). It is also noticeable that there are more very short introns in the studied Formicidae (*A. echinator*, *C. floridanus*, *H. saltator*) than in any other species (Supplementary figure D.1.3, p.17), resulting in a bimodal distribution which peaks are very far apart considering the other distributions (indicated by a rectangle in Fig. IV.4). Most transcripts (average Q1–Q3, Supplementary table D.1.1 e) have an added intron length of 486–5108 bp, a median intron length of 81–328 bp, and have an intron coverage of 0.3–0.7 and an intron density of 0.0006–0.02.

**3.4 Distributions and correlations of GC content**

Concerning the distribution of GC content per transcripts, *A. mellifera* stands out with a left-shifted distribution (max at 25 %), followed by *P. dominula* (peak at ca. 28 %) and *B. terrestris* (max at ca. 29 %); for *M. rotundata*, *A. echinator*, and *C. floridanus*, the distributions are nearly congruent (max around 30 %); the rest of the species falls into a third group of similar distribution with a maximum close to 40 % (Fig. IV.5, Supplementary figure D.1.3, p.3). The distributions of median exon and intron GC content per transcript follow a very similar motif. Hence, a pattern of continuous GC content decrease can be postulated, from high GC contents in species that diverge early in the hymenopteran phylogeny to lower GC contents in later-diverging social and eusocial species.

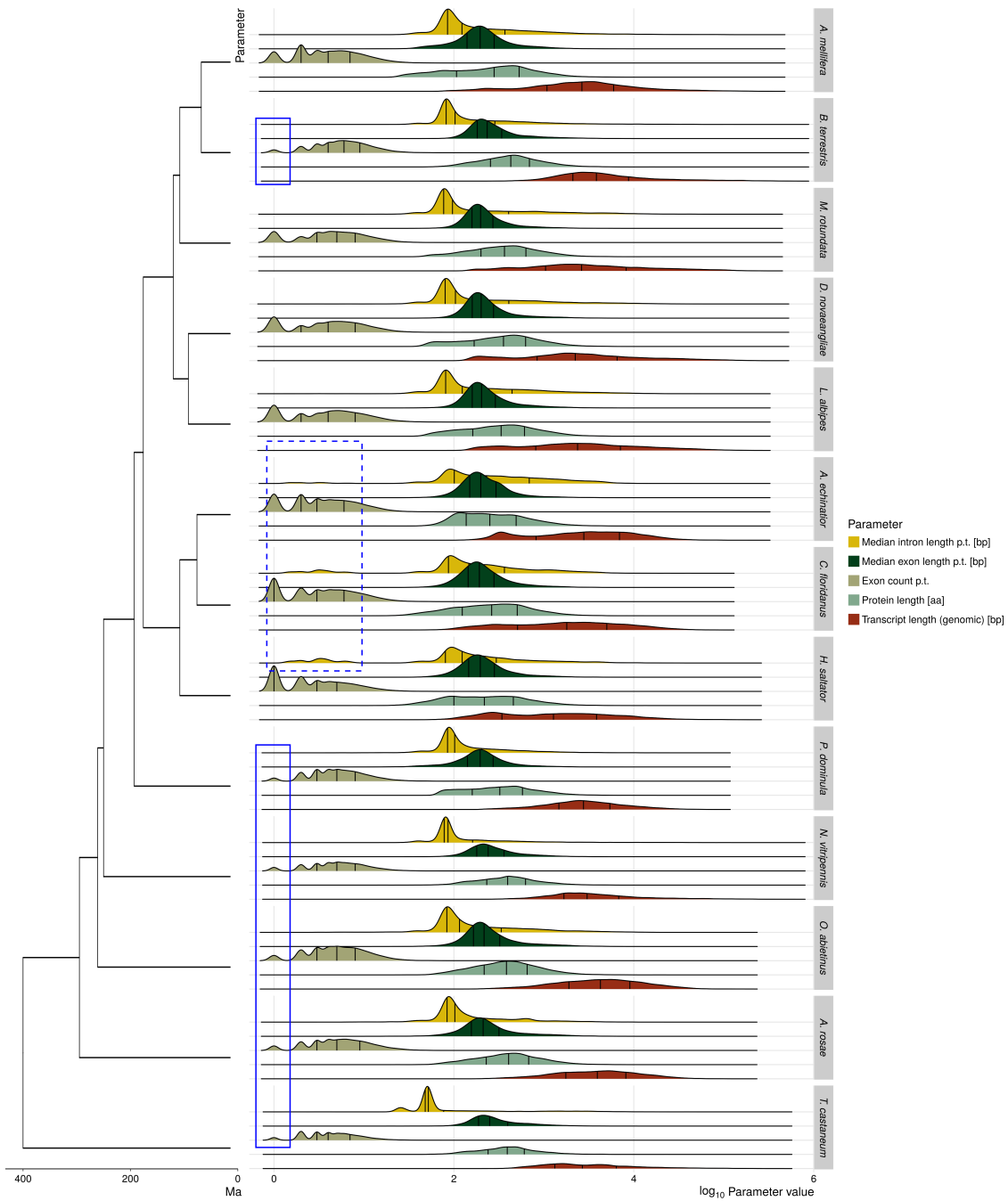


Figure IV.4 – Comparison of lengths and counts of gene elements for all gens.  
(Continued on next page)

**Figure IV.4 – Comparison of lengths and counts of gene elements for all genes.**

(Continued)

For each species, the five gene structure parameters were recorded for each of their genes (as analyzed by COGNATE, longest transcript per gene). Vertical lines indicate the 25 %, 50 % (median), and 75 % quantile of the respective parameter distribution.

Note the blue rectangles that indicate conspicuous distribution features: the solid rectangles mark the low number of transcripts with only one exon; the dashed rectangle highlights the peaks of transcripts with very short introns only found in Formicidae.

The phylogenetic tree illustrates the relationships between the analyzed species, branch lengths correspond to divergence time.

bp: basepairs; Ma: Million years ago

When comparing the GC content of exons or introns with the count of exons or introns per transcript, respectively, there appears to be a normal distributions of GC content over exon count; for introns, however, distributions are right-skewed, which means that the introns of transcripts with more introns have a lower GC content (Fig. IV.6). Note that the centers of data-gravity (*i.e.*, the highest concentration of data points) lie at small introns with rather low GC content. These findings resemble those mentioned above (introns having generally a lower GC content than exons or transcripts as a whole). Again, the observed trend increases within Apocrita, while a higher median intron GC content appears at the base of the considered phylogeny.

Additional data (*e.g.*, more parameters like coverages and densities of gene elements, correlative plots, and median overviews) can be found in the Appendix (Supplementary file D.1.1, Supplementary figure D.1.3).

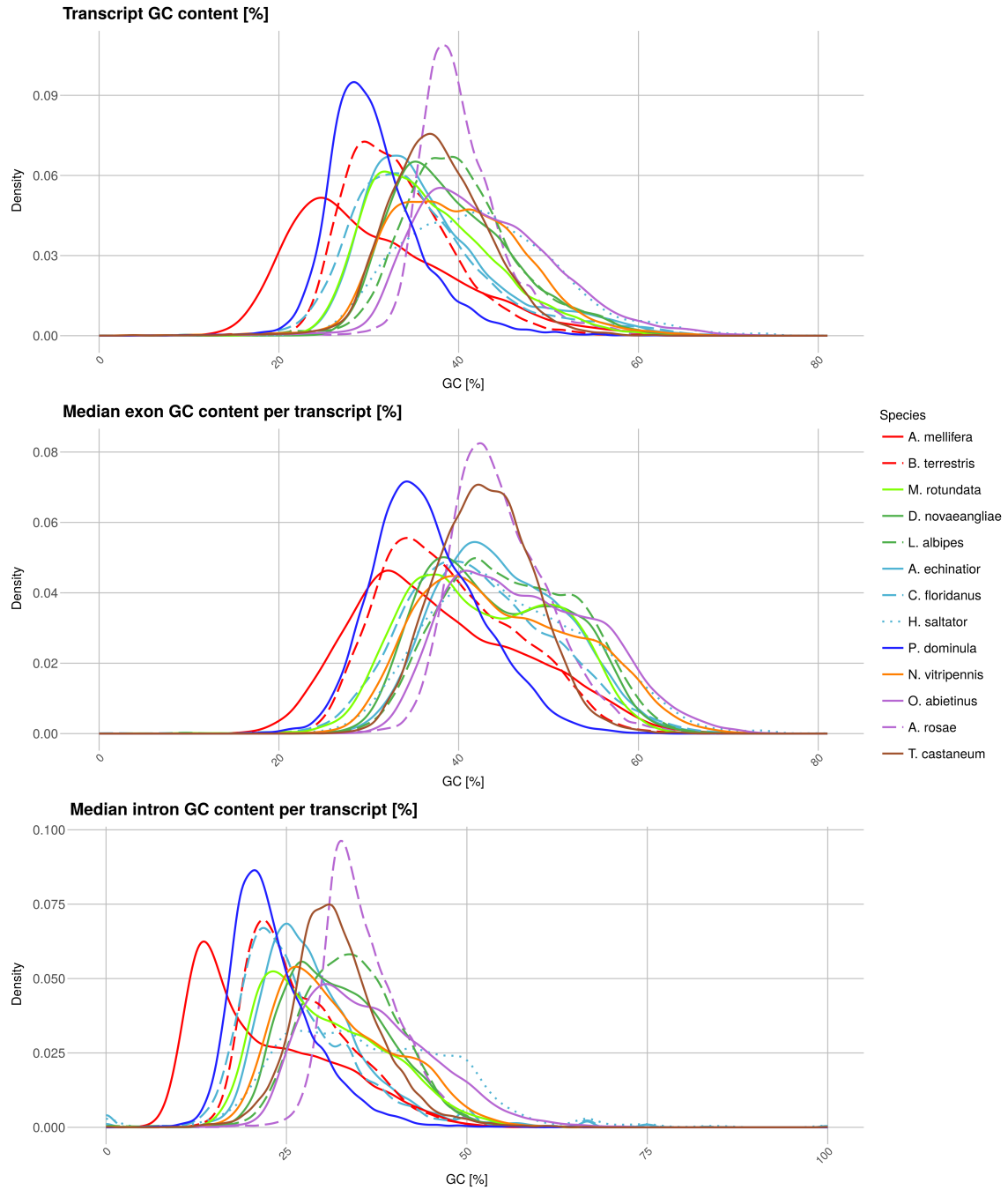


Figure IV.5 – Comparison of GC content across species. (Continued on next page)

**Figure IV.5 – Comparison of GC content across species.**

(Continued)

For each species, the distribution of GC content per transcript is given as value for each transcript itself as well as for the median of all exons and introns per transcript, respectively.

The list of species is ordered according to phylogenetic relationships (see Fig. IV.1).

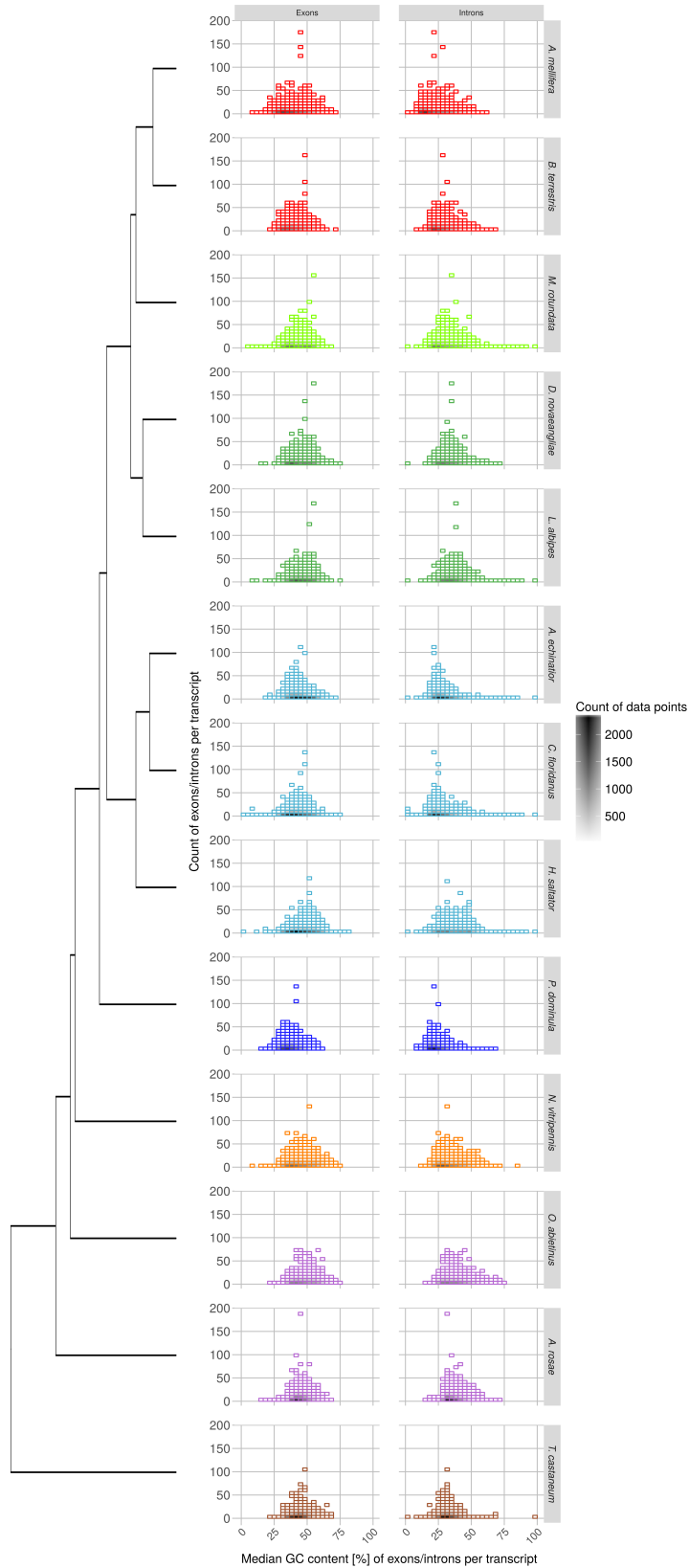


Figure IV.6 – Comparison of median GC content to count of gene element across species. (Continued on next page)



**Figure IV.6 – Comparison of median GC content to count of gene element across species.**

(Continued)

For each species, the count of the respective gene element (exon and intron) is recorded per transcript and plotted against the median GC content of the respective element per transcript. The count of data points (data gravity) is indicated by the shade of each data record cell, the darker, the more transcripts have the indicated characteristics.

The list of species is ordered according to phylogenetic relationships (see Fig. IV.1).



---

## Discussion

---

### 4.1 Component sizes

PREVIOUS GENOME PUBLICATIONS provided only a limited set of summary statistics to describe the protein-coding gene repertoire, like the mean gene, exon, and intron length. As outlined by WILBRANDT *et al.* (2017), gene structure parameters are rarely normally distributed, thus the mean is rarely an adequate summary statistic. Furthermore, it is important to not only provide the full dataset for future reference, but also to consider the whole distribution of data rather than a summary statistic alone. This study aimed to capture the diversity of gene structures within a repertoire by analyzing the whole parameter distributions in concordance with summary statistics.

When comparing the absolute and relative amounts of coding, non-coding and repetitive genome content and regarding this in context of genome size, it becomes apparent that the absolute amount of protein-coding sequences is quite stable across all compared species (median of ca. 19 Mbp). Conversely, genome size differences appear to be partly driven by content differences of introns, transposable elements and a varying amount of other sequences. It appears

that a small intron component of the genome might have been the ancestral state in hymenopterans, increasing towards Apoidea (with the exception of *N. vitripennis*, which diverged earlier but comprises more introns than the other considered Apocrita). It is easy to assume here that protein-coding gene content is thus overall conserved in Hymenoptera. However, constant numbers may mask a considerable gene turnover, as suggested by (HAHN *et al.*, 2007). Here, further studies, namely gene family annotation and turnover rate estimation, are worthwhile.

## 4.2 Tentative inference of gene structure evolution trends

It is difficult to establish hypotheses on ancestral gene configuration, because this would require to reconstruct the ancestral state of continuously varying traits. Thus, the proposed inferences are tentative.

Taking together the results of the COGNATE analysis comparisons, three trends appear to be present. Firstly, median exon length and transcript length seem to decrease while intron length increases along the phylogeny of the considered Hymenoptera, suggesting that the ancestral gene repertoire might have harbored more long transcripts, consisting of (in the median) five long exons and relatively short introns. However, whether this trend is statistically significant cannot be examined yet (it involves problems of phylogenetic interdependence and non-normal distribution). Secondly, within Formicidae, genes with very short introns have been established as discernible class, and the median exon count is consistently lower than in all other considered species. A detailed investigation of these short-intron genes in comparison to the remaining gene repertoire, including functional annotation appears to be rewarding. Thirdly, GC content is increasing along the considered hymenopteran phylogeny with high values (of, *e.g.*, assembly GC content) in *A. rosae* and *O. abietinus* of over 40 % to low values around 30 % in the remaining Apocrita. *A. mellifera* and *P. dominula* are especially noticeable with their strongly right-skewed distributions of GC contents across transcripts (as also discussed by STANDAGE *et al.*, 2016).

The strong bias towards a low GC content (as found in *A. mellifera* and *P. dominula*, and to a lesser extent in *B. terrestris*) has been suggested to be

driven by “a bias in DNA mismatch repair and other genome maintenance mechanisms, as well as the possibility of historically high levels of CpG methylation and cytosine deamination” (STANDAGE *et al.*, 2016). at the same time, *P. dominula* has been reported to have a reduced methylation system and low genomic levels of DNA methylation overall compared to other Hymenoptera (STANDAGE *et al.*, 2016). It has been speculated that the overall low GC content of *A. mellifera* is a result of a mutational bias towards a predominance of adenosine and thymidine in the DNA sequence, possibly part of a distinct mutational pattern, and accompanied by a codon usage bias (JØRGENSEN *et al.*, 2007). Another line of evidence indicates that regions of high GC content also show high recombination rates (KENT *et al.*, 2012; NIEHUIS *et al.*, 2010; ROSS *et al.*, 2015); recombination rates are overall very high in *A. mellifera* (BEYE *et al.*, 2006). The distribution of GC contents across transcripts is very broad, thus it is possible that genes with a high GC content (usually also located in high-GC DNA environments, JØRGENSEN *et al.*, 2007) lie in regions with exceptionally high recombination rates. It remains unknown how the overall low GC content of *A. mellifera* can be explained. Taken together, this indicates that the shift of GC distributions towards low GC contents as well as the high recombination rates found in *A. mellifera* and other apoid Hymenoptera are derived traits with a mechanistic correlation. The modes of evolution leading to these phenomena are yet open to further elucidation.



---

## Conclusion

---

**T**HE STRUCTURAL CHARACTERIZATION of eusocial Hymenoptera (representative samples of Apoidea, Formicidae, and a vespidae species) in comparison to non-social (*P. dominula*) and non-apocritan ("Symphyta") species revealed that the considered ants have distinct gene structures (few exons, long introns, resulting in short proteins). Furthermore, a distinctive class of genes with extremely short introns appears to have been established in the ant lineage. The ancestral gene structure of Hymenoptera appears to be rather complex and feature long exons as well as a high GC content.

This primary gene structure characterization lays the foundation for future research on the evolution of gene structure and gene repertoire composition. Additionally, it is an integral part of the description of Hymenopteran genomes in concordance with a detailed study of the evolution of olfactory receptors and other gene families important in the evolution of hymenopteran parasitoidism and eusociality.





---

## Bibliography IV

---

- BEYE, M, I GATTERMEIER, M HASSELMANN, T GEMPE, M SCHIOETT, JF BAINES, D SCHLIPALIUS, F MOUGEL, C EMORE, O RUEPELL, A SIRVIÖ, E GUZMÁN-NOVOA, G HUNT, M SOLIGNAC, and RE PAGE (2006). **Exceptionally high levels of recombination across the honey bee genome.** *Genome Research* 16.11, pp. 1339–1344. DOI: 10.1101/gr.5680406 (cit. on p. 149).
- CLARK, AG *et al.* (2007). **Evolution of genes and genomes on the *Drosophila* phylogeny.** *Nature* 450.7167, pp. 203–218. DOI: 10.1038/nature06341 (cit. on p. 126).
- ELLIOTT, TA and TR GREGORY (2015). **What's in a genome? The C-value enigma and the evolution of eukaryotic genome content.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678, p. 20140331. DOI: 10.1098/rstb.2014.0331 (cit. on pp. 125, 126).
- HAHN, MW, MV HAN, and SG HAN (2007). **Gene Family Evolution across 12 *Drosophila* Genomes.** *PLoS Genet* 3.11, e197. DOI: 10.1371/journal.pgen.0030197 (cit. on p. 148).
- i5K CONSORTIUM (2013). **The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.** *Journal of Heredity* 104.5, pp. 595–600. DOI: 10.1093/jhered/est050 (cit. on p. 129).
- JØRGENSEN, FG, MH SCHIERUP, and AG CLARK (2007). **Heterogeneity in Regional GC Content and Differential Usage of Codons and Amino Acids in GC-Poor and GC-Rich Regions of the Genome of *Apis mellifera*.** *Molecular Biology and Evolution* 24.2, pp. 611–619. DOI: 10.1093/molbev/msl190 (cit. on p. 149).
- KENT, CF, S MINAEI, BA HARPUR, and A ZAYED (2012). **Recombination is associated with the evolution of genome structure and worker behavior in honey bees.** *Proceedings of the National Academy of Sciences* 109.44, pp. 18012–18017. DOI: 10.1073/pnas.1208094109 (cit. on p. 149).

- NIEHUIS, O, JD GIBSON, MS ROSENBERG, BA PANNEBAKKER, T KOEVOETS, AK JUDSON, CA DESJARDINS, K KENNEDY, D DUGGAN, LW BEUKEBOOM, Lvd ZANDE, DM SHUKER, JH WERREN, and J GADAU (2010). **Recombination and Its Impact on the Genome of the Haplodiploid Parasitoid Wasp *Nasonia***. *PLOS ONE* 5.1, e8597. DOI: 10.1371/journal.pone.0008597 (cit. on p. 149).
- ROSS, CR, DS DEFELICE, GJ HUNT, KE IHLE, GV AMDAM, and O RUEPPELL (2015). **Genomic correlates of recombination rate and its variability across eight recombination maps in the western honey bee (*Apis mellifera* L.)** *BMC Genomics* 16.1, p. 107. DOI: 10.1186/s12864-015-1281-2 (cit. on p. 149).
- ROY, SW and D PENNY (2007). **Intron length distributions and gene prediction**. *Nucleic Acids Research* 35.14, pp. 4737–4742. DOI: 10.1093/nar/gkm281 (cit. on p. 134).
- STANDAGE, DS, AJ BERENS, KM GLASTAD, AJ SEVERIN, VP BRENDEL, and AL TOTH (2016). **Genome, transcriptome, and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect**. *Molecular Ecology*, n/a–n/a. DOI: 10.1111/mec.13578 (cit. on pp. 148, 149).
- SUEN, G, C TEILING, L LI, C HOLT, E ABOUHEIF, E BORNBERG-BAUER, P BOUFFARD, EJ CALDERA, E CASH, A CAVANAUGH, O DENAS, E ELHAIK, MJ FAVÉ, J GADAU, JD GIBSON, D GRAUR, KJ GRUBBS, DE HAGEN, TT HARKINS, M HELMKAMPF, H HU, BR JOHNSON, J KIM, SE MARSH, JA MOELLER, MC MUÑOZ-TORRES, MC MURPHY, MC NAUGHTON, S NIGAM, R OVERSON, R RAJAKUMAR, JT REESE, JJ SCOTT, CR SMITH, S TAO, ND TSUTSUI, L VILJAKAINEN, L WISLER, MD YANDELL, F ZIMMER, J TAYLOR, SC SLATER, SW CLIFTON, WC WARREN, CG ELSIK, CD SMITH, GM WEINSTOCK, NM GERARDO, and CR CURRIE (2011). **The Genome Sequence of the Leaf-Cutter Ant *Atta cephalotes* Reveals Insights into Its Obligate Symbiotic Lifestyle**. *PLoS Genet* 7.2, e1002007. DOI: 10.1371/journal.pgen.1002007 (cit. on p. 126).
- WICKHAM, H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York (cit. on p. 131).
- WILBRANDT, J, B MISOF, and O NIEHUIS (2017). **COGNATE: comparative gene annotation characterizer**. *BMC Genomics* 18.1, p. 535. DOI: 10.1186/s12864-017-3870-8 (cit. on pp. 130, 147).
- YANDELL, M, CJ MUNGALL, C SMITH, S PROCHNIK, J KAMINKER, G HARTZELL, S LEWIS, and GM RUBIN (2006). **Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes**. *PLoS Comput Biol* 2.3, e15. DOI: 10.1371/journal.pcbi.0020015 (cit. on p. 131).
- ZDOBNOV, EM and P BORK (2007). **Quantification of insect genome divergence**. *Trends in Genetics* 23.1, pp. 16–20. DOI: 10.1016/j.tig.2006.10.004 (cit. on p. 126).

---

# **CORE, SHELL, AND CLOUD – THE CONSERVATION CLASSES OF INSECT GENE REPERTOIRES DIFFER IN GENE STRUCTURE AND PROTEIN DOMAIN DIVERSITY**

---

The following text comprises preliminary work for a publication.

**WILBRANDT J, PETERSEN M, PROVATARIS P, WATERHOUSE RM, MISOF B,  
AND NIEHUIS O (2018) Core, Shell, and Cloud – the Conservation Classes of  
Insect Gene Repertoires Differ in Gene Structure and Protein Domain  
Diversity. In prep.**

Authors' and co-workers' contributions to this chapter:

Specimen collection: JW, MP, ZIESMANN T, OEYEN JP, JÄKEL H, ON, PETERS RS,  
SCHMITT T, WESER T, MCKENNA D, SCHÜTTE K, POHL H, NIEHUIS M,  
ALTENHOFER E, MEIER R, BERG A, WURDACK M, GREEN PWC, MARTIN S,  
LESNY P, JANSSEN H, GUNKEL S, SCHÄFER M, FOELKER C, RAJARATNAM G,  
JANŠTA P; sequencing (responsible): ON, BM; assembly (responsible): DONATH  
A, MP; annotation: JW (protein-coding genes), PP (protein domains); orthology  
prediction: RMW; manuscript design: JW, BM, ON; structural characterization,  
analyses, figures, manuscript writing: JW;



---

## Summary

---

**G**ENES HAVE DIFFERENT FATES — some are ancient and their orthologs can be found in distant taxa, while others are quickly gained and lost — and thus the gene repertoire of a species can be partitioned according to the genes' conservation in comparison to other repertoires/species. This can be summed up as 'universality' (how many species share a gene family/ortholog group) and 'duplicability' (how many family members does each species have).

**T**HE GENE REPERTOIRES of 26 newly sequenced species were obtained and analyzed to determine whether the conservation classification into core, shell, and cloud according to gene family universality produces a pattern similar to the proposed universal patterns (KOONIN, 2011; WATERHOUSE, 2015). In a second step, the analysis was extended to respect duplicability of the gene families. The gene sets of the different conservation classes and copy states were characterized using five central gene structure parameters as well as protein domain diversity to answer the question whether these repertoire subsets differed from one another in the distributions and medians of the characterized aspects.

**H**ERE IT IS SHOWN that core, shell, and cloud classes can be established as proposed. Furthermore, genes of distinct classes and copy states consistently differ in gene structure and domain diversity. Genes of the core are longer and more complex than shell and cloud genes (in line with previous evidence; CLARK *et al.*, 2007; LIPMAN *et al.*, 2002), and those core genes not being universal single-copy orthologs are astonishingly domain-rich. This indicates that the high costs associated with maintaining complex genes is balanced by a tremendous advantage that yet needs to be characterized.

**T**HESE RESULTS POINT OUT that there is considerable variation present within protein-coding gene repertoires and related to gene conservation as well as species evolution. Thus, the categorization according to universality and duplicability seems to be a relevant aspect to consider when studying protein-coding gene repertoire evolution.

---

## Introduction

---

**T**HE COMPARISON of whole gene repertoires across species has revealed considerable gene and gene family turnover (BEMM *et al.*, 2016; CARRETERO-PAULET *et al.*, 2015; HAHN *et al.*, 2007; KEITH *et al.*, 2015). On the other hand, minimal sets of housekeeping genes necessary for cell survival have been delineated (KOONIN, 2003; MUSHEGIAN and KOONIN, 1996), and sets of universally present genes have been described (HARRIS *et al.*, 2003; OUZOUNIS *et al.*, 2006; PARRA *et al.*, 2009; SIMÃO *et al.*, 2015). In other words, genes have different fates – some are ancient and their orthologs can be found in distant taxa, while others are quickly gained and lost in a “genomic ‘revolving door’ of gain and loss” leading to high turnover (DEMUTH *et al.*, 2006; PALMIERI *et al.*, 2014). Young, taxon-restricted genes likely play a role in the evolution of biological novelties (JOHNSON and TSUTSUI, 2011; KHALTURIN *et al.*, 2009; SIMOLA *et al.*, 2013; ZHAO *et al.*, 2015), while the loss of ancient orthologs appears to be tolerable (WYDER *et al.*, 2007).

As the knowledge on insect genomes culminates, patterns of duplicability and universality of ortholog groups have been postulated to be ubiquitous in gene repertoires (WATERHOUSE, 2015; WATERHOUSE *et al.*, 2011). Duplicability is the

tendency of a gene family (ortholog group) to comprise few (low duplicability) or many (high duplicability) members; universality reflects the conservation of a gene family, whether it is present in many (high universality) or few (low universality) species. The finding that most universal ortholog groups are either highly duplicable or not at all led to the coining of 'single-copy control' and 'multicopy license' to describe evolutionary modes governing gene conservation (WATERHOUSE *et al.*, 2011).

A closely related pattern was posited to be an omnipresent, even fractal (present at all levels of comparison regarding taxonomic membership and the number of species compared) feature of the "gene universe" (KOONIN, 2011). When analyzing the number of species and the number of gene families (and gene family sizes) shared by them, a characteristic bowl plot can be drawn, irrespective of how many or which species are compared (KOONIN, 2011). The bowl plot corresponds to the landscape plot put forward by WATERHOUSE (2015), namely when looking from the landscape's universality plane and adding all counts across the single-copyness axis; this is illustrated in Fig. V.4 using our data.

Basically, three conservation partitions or classes can be established according to the 'conservation distribution' of genes (KOONIN, 2011, p. 71): (1) the core, comprising universal and highly conserved genes shared by all species; (2) the shell, including moderately conserved genes shared by some species, representing a broad variety of the genome but not the majority of genes; (3) the cloud, a large number of poorly conserved genes shared by very few or no other species. In bacteriology, other names have been put forth for core and cloud, namely 'persistent genes' of the 'paleome' (core) and 'orphan genes' of the 'cenome' (cloud) (DANCHIN, 2009). There are also studies simply referring to young and old genes, *e.g.*, CAI and PETROV (2010).

It has been suggested that a large part of the minimal gene set a cell requires to live is conserved in the core (KOONIN, 2003). Essentiality, expression level, and the number of partners in protein-protein interaction networks also correlate with the tendency of a gene family to be small and conserved (KRYLOV *et al.*, 2003; WATERHOUSE *et al.*, 2011). Evidence has been collected that old orthologous genes (descendants from a gene present in the last common ancestor of a set of species) are longer than species-specific genes (CLARK *et al.*, 2007; LIPMAN *et al.*, 2002; WASMUTH *et al.*, 2008). However, these studies



---

were limited by small species sample sizes, they considered only the average of length parameters, and did not assess the full diversity of gene structures across the complete gene repertoire. Although the description of conservation patterns was accompanied by a detailed description of evolutionary traits like sequence divergence and essentiality, an intriguing route of investigation was left untouched: the structural characteristics of the genes found in either conservation class were not fully disclosed.

These studies raised my interest in the three following questions.

- ▷ First, can I reproduce the pattern of conservation shown by WATERHOUSE (2015)? This is interesting, because data was not disclosed for all insects of the set used by WATERHOUSE (2015), although it can be assumed to be representative. I would like to explore the variability.
- ▷ Second, what are the gene structural features of the repertoire partitions, split to conservation classes and copy status, considering not only averages but full parameter distributions? Since structural parameter distributions are often omitted (not always, see LIPMAN *et al.*, 2002), it is intriguing to examine them. Additionally, since previous studies were either limited in methods (*e.g.*, determining conservation by simple sequence similarity to any database entry, LIPMAN *et al.*, 2002) or sample size, and did not consider the repertoire partitions according to both conservation and copy status, this approach is interesting.
- ▷ Third, do the repertoire partitions differ in protein domain diversity? Although the protein domain signatures of ortholog groups have been investigated, this was a sophisticated assessment based on an inference of relationships among ortholog groups (WATERHOUSE *et al.*, 2011). Domains can be considered building blocks of genes, and it is expected that domain shuffling is the predominant route to expand protein space, rather than *de novo* domain creation (APIC and RUSSELL, 2010; MOORE and BORNBERG-BAUER, 2012). Thus, I am interested to explore whether core, shell, and cloud differ in domain diversity and domain arrangement diversity.

To approach the three posed questions, a sample of 26 newly sequenced arthropod species is analyzed. It comprises a dense sampling of Hymenoptera (sawflies, bees, and wasps) together with six species representing holometabolous (insects with an ontogenetic pupal stage) outgroup orders, one hemimetabolous (insects with direct ontogeny) outgroup species, and

one millipede as non-insect outgroup. The gene repertoires of these species are consistently annotated *de novo*, orthology relationships are inferred, and genes are structurally characterized and annotated with protein domains. The obtained parameter distributions as well as domain annotations are analyzed in detail with respect to the ten repertoire partitions put up between conservation classes and copy states. Also, a general description of these new gene repertoires is provided.

---

## Methods

---

**T**HE FOLLOWING STEPS of data compilation were conducted within the Leibniz Graduate School on Genomic Biodiversity Research of the Zoological Research Museum Alexander Koenig, Bonn, over the course of several years. The work was done not only by me but also by many co-workers (indicated in the respective sections) and the genome assemblies and parts of the meta-data are used in the dissertations of several other doctoral researchers as well.

### 3.1 Species sample

The species sample was selected under guidance of the phylogenetic backbone tree of insects (MISOF *et al.*, 2014) to cover the extant diversity of Holometabola with a focus on groups not well captured in previously published studies. The majority of holometabolan orders is covered by 24 species by including representatives of Diptera (two species), Mecoptera (one species), Megaloptera (one species), Coleoptera (one species), Strepsiptera (one species), Hymenoptera (18 species). Two outgroups were added: Psocodea (one species) and Glomerida (one species). The species sample is depicted in Table V.1 and Figure V.1.

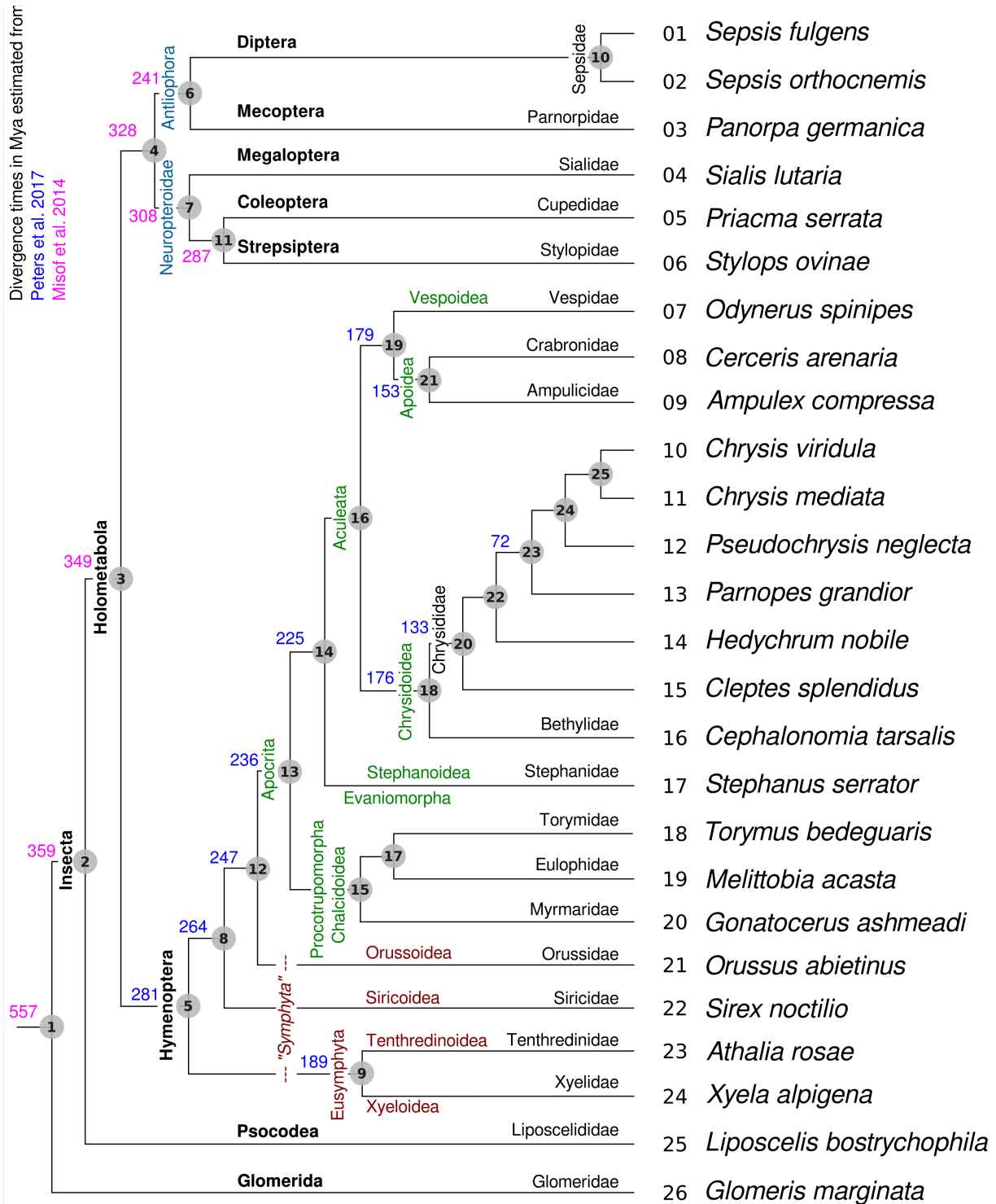


Figure V.1 – Phylogenetic relationships between species of the sample. (Continued on next page)

**Figure V.1 – Phylogenetic relationships between species of the sample.**

(Continued)

Numbers in grey circles are node IDs. Each species has a unique ID prepending its name.

Pink numbers indicate estimated approximate divergence times in Ma (Million years ago) referring to MISOF *et al.* (2014), while blue numbers mark estimated approximate divergence times from the publication by PETERS *et al.* (2017).

Coloring of taxonomic names follows the scheme red for “Symphyta” and the comprised lineages, green for Apocrita and descending lineages, and blue for other Holometabola than Hymenoptera. Branch lengths are arbitrary.

Mya: million years ago.

This taxonomic sampling covers ancient divergences (roughly 550 million years (My) between Glomeridera and the sampled insects) as well as rather recent radiations (less than 70 My between species of the jewel wasps (Chalcidoidea)), (MISOF *et al.*, 2014; PETERS *et al.*, 2017). Furthermore, the species set comprises a large range of genome sizes (63–1.2 Mbp).

## 3.2 Sequencing of genomes and transcriptomes

Library preparation and the actual sequencing steps were outsourced. Genome sequencing was done using Illumina HiSeq 2000 and HiSeq 4000 sequencing technology (Illumina, San Diego, CA, USA) to obtain a set of four libraries for each species: (1) 250 bp insert size paired-end library, minimum base coverage depth 40x, 150 bp read length; (2) 800 bp insert size paired-end library, minimum base coverage depth 10x, 150 bp read length; (3) 3 kbp insert size mate-pair library, minimum base coverage depth 10x, 100 bp read length; (4) 8 kbp insert size mate-pair library, minimum base coverage depth 10x, 100 bp read length.

Haploid genome sizes were estimated by OLIVER NIEHUIS using the program Jellyfish (MARÇAIS and KINGSFORD, 2011) and an in-house script with a 17-mer frequency distribution of reads in the 250-bp library.

Whole-body transcriptomes were sequenced on a Illumina HighSeq 2000 sequencing platform. Libraries contained after read quality trimming and adapter clipping at least 40 million 100 bp long paired-end read pairs. These transcriptomes (hereafter referred to as RNAseq) were used in the annotation of protein-coding genes.

ID	Abbrev.	Species	Genome size	Gene count
01	SEFU	<i>Sepsis fulgens</i>	211	16,146
02	SEOR	<i>Sepsis orthocnemis</i>	202	14,141
03	PAGE	<i>Panorpa germanica</i>	1,066	27,083
04	SILU	<i>Sialis lutaria</i>	564	18,165
05	PRSE	<i>Priacma serrata</i>	994	32,528
06	STOV	<i>Stylops ovinae</i>	63	9,592
07	ODSP	<i>Odynerus spinipes</i>	242	13,222
08	CEAR	<i>Cerceris arenaria</i>	482	27,352
09	AMCO	<i>Ampulex compressa</i>	374	19,080
10	CHVI	<i>Chrysis viridula</i>	209	17,474
11	CHME	<i>Chrysis mediata</i>	228	18,959
12	PSNE	<i>Pseudochrysis neglecta</i>	223	18,038
13	PAGR	<i>Parnopes grandior</i>	189	13,170
14	HENO	<i>Hedychrum nobile</i>	204	14,863
15	CLSP	<i>Cleptes splendidus</i>	571	14,580
16	CETA	<i>Cephalonomia tarsalis</i>	184	14,081
17	STSE	<i>Stephanus grandior</i>	1,173	30,725
18	TOBE	<i>Torymus bedeguaris</i>	918	31,685
19	MEAC	<i>Melittobia acasta</i>	246	20,111
20	GOAS	<i>Gonatocerus ashmeadi</i>	324	40,451
21	OABI	<i>Orussus abietinus</i>	247	14,049
22	SINO	<i>Sirex noctilio</i>	244	17,609
23	AROS	<i>Athalia rosae</i>	170	14,319
24	XYAL	<i>Xyela alpigena</i>	325	23,597
25	LIBO	<i>Liposcelis bostrychophila</i>	542	37,281
26	GLMA	<i>Glomeris marginata</i>	324	20,472

**Table V.1 – Species sample.** The species are ordered according to phylogenetic relationships, see V.1. Genome size (as estimated using Jellyfish) is given in Mbp.

### 3.3 Assembly and repeat masking

Genome assemblies were produced by ALEXANDER DONATH and LARS PODSIADLOWSKI. They also developed the method to select the best assembly, supported by MALTE PETERSEN, who also did the repeat masking steps.

In order to obtain the optimal assembly, an internal assemblathon was conducted. The assembler Platanus (version 1.2.4, KAJITANI *et al.*, 2014) was used with varying read library combinations for each of the three steps (contig assembly, scaffolding, gap closing) and compared with a single run of the assembler Allpaths-LG (version 52488, BUTLER *et al.*, 2008). The best assembly was chosen using a custom script based on a combination of standard assembly quality metrics (closeness to estimated genome size, scaffold N50), BUSCO values (SIMÃO *et al.*, 2015), and mapping of reads (using CLC Genomics Workbench, version 8.5<sup>V.1</sup>). The assembler calls and library combinations can be found in the Appendix (E.1.1), the custom script for quality assessment is available upon request from MALTE PETERSEN.

It is required for gene annotation with the BRAKER pipeline (HOFF *et al.*, 2016) that the assembled genome sequences are softmasked, *i.e.*, transposable elements and other repetitive or low complexity regions are 'hidden' from the searching algorithm. This is usually done by encoding softmasked sequences in lower-case letters, while unmasked sequences remain capitalized (done with fastasoftmask from Exonerate version 2.2, SLATER and BIRNEY, 2005). To acquire the information, which sequences should be softmasked, a custom pipeline based on RepeatModeler (versions 1.0.8-1.0.10, SMIT and HUBLEY, 2015) and RepeatMasker (versions 4.0.5-4.0.7, SMIT *et al.*, 2015) developed by MALTE PETERSEN was employed.

---

<sup>V.1</sup> CLC Genomics Workbench: <https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/>. Last accessed 30 Mar 2018.

### 3.4 MySQL database setup

To store the results specifically for the following analyses and to allow fast querying, I established a relational MySQL database. The following scheme was built using custom Perl scripts as well as direct MySQL commands (documented in the Appendix, E.2.1).

### 3.5 Protein-coding gene annotation and structural characterization

Protein-coding gene annotation and structural characterization was done by me.

The BRAKER pipeline (HOFF *et al.*, 2016) requires that intrinsic evidence, *i.e.*, species-specific RNAseq data, is provided with the information where it aligns to the genome (in a sorted `.bam` file). To obtain RNAseq alignments, the softmasked genome assemblies and raw RNA reads of each species were provided as input to HISAT2 version 2.1.0, KIM *et al.*, 2015). The HISAT output was modified using bamTools version 2.3.0 (BARNETT *et al.*, 2011) and samTools version 1.7 (LI *et al.*, 2009). HISAT was run (including output modification) locally with the following call scheme (SPECIES needed to be changed accordingly).

```
hisat2-build SPECIES.fasta.softmasked \
SPECIES_hisat_index

hisat2 -x SPECIES_hisat_index -1 RNASEQ_1.fq.gz \
-2 RNASEQ_2.fq.gz -S SPECIES_hisat2.sam

samtools view -bS SPECIES_hisat2.sam \
> SPECIES_hisat2.bam

bamtools sort -in SPECIES_hisat2.bam \
-out SPECIES_hisat2_sorted.bam
```



The softmasked genome assemblies and species-specific RNAseq alignments were used as input for the BRAKER2 pipeline (version 2.0.6, HOFF *et al.*, 2016; using GeneMark version 4.33, BESEMER and BORODOVSKY, 2005; Augustus version 3.3, STANKE *et al.*, 2004; bamTools version 2.3.0, BARNETT *et al.*, 2011; samTools version 1.7, LI *et al.*, 2009; and NCBI BLAST+ version 2.6.0, CAMACHO *et al.*, 2009) and run on the ZFMK HPC cluster with the following call scheme (enclosed in a submission script written in bash; SPECIES needed to be changed accordingly).

```
/BRAKER\_v2.1.0/braker.pl \  
  --species=SPECIES \  
  --genome=SPECIES.fasta-softmasked \  
  --bam=SPECIES_hisat2_sorted.bam \  
  --workingdir=SPECIES_DIR \  
  --UTR=off --gff3 --softmasking \  
  --overwrite \  
  --AUGUSTUS_CONFIG_PATH=/Augustus/3.3/config \  
  --AUGUSTUS_BIN_PATH=/Augustus/3.3/bin \  
  --AUGUSTUS_SCRIPTS_PATH=/Augustus/3.3/scripts \  
  --BAMTOOLS_PATH=/bamtools/2.5.1/bin \  
  --GENEMARK_PATH=/GeneMark/4.33/ \  
  --SAMTOOLS_PATH=/samtools/1.7/bin/ \  
  --BLAST_PATH=/blast/2.6.0+/
```

Note that the UTR option was toggled off, because it is according to the Augustus 3.3 documentation currently not working for annotating insect genomes.

The gene annotation output (hints.augustus.gff3 file) was, together with the softmasked genome assembly, given to COGNATE version 1.01 (WILBRANDT *et al.*, 2017) for analysis under default parameters (analyzing the longest transcript per gene and producing all possible files).

Since COGNATE analyzes the longest transcript per gene, these transcripts can be viewed as representing directly their respective gene. Thus, in this manuscript, the terms are used interchangeably for convenience.

To characterize gene structure, five central parameters were chosen to be further analyzed. These are the same as in my previous projects (see parts III and IV): (median) genomic transcript length (including introns and exons), (median) protein length, (median) exon and intron length, (median) exon count for each transcript or as median over all transcripts (of a set). The results were stored in the MySQL database (table Transcript\_COGNATE\_general).

### 3.6 Assessment of gene space coverage

To assess the gene space coverage with benchmark single-copy orthologs (BUSCOs, SIMÃO *et al.*, 2015) based on OrthoDB by ZDOBNOV *et al.* (2017), I employed the tool BUSCO version 3.0.2 (using NCBI BLAST+ 2.6.0, CAMACHO *et al.*, 2009, HMMER version 3.1b2<sup>V.2</sup>, and Augustus version 3.3, STANKE *et al.*, 2004), running it locally with the lineage dataset `insects_odb9` in protein mode (see command below). The lineage dataset contains 1,658 BUSCO genes.

```
python scripts/run_BUSCO.py \  
  -i ./SPECIES_PROTEINS.fa \  
  -o SPECIES \  
  -l insecta_odb9/ \  
  -m proteins \  
  -c 1
```

### 3.7 Protein domain annotation and analysis

Protein domain annotations were provided by PANAGIOTIS PROVATARIS, I analyzed the data.

Protein domains were annotated for all species based on the amino acid sequences of the longest transcript per gene as given by COGNATE. To this

---

<sup>V.2</sup> HMMER: [hmmerr.org](http://hmmerr.org). Last accessed 30 Mar 2018.

end, HMMER (version 3.1b2) was applied using the Pfam-A database (version 31.0<sup>V.3</sup>) in the following call (where `-domtblout` determines the output format and `-cut_ga` causes the use of gathering cutoffs set by Pfam curators to set all thresholding for domain identification).

```
hmmsearch --domtblout $OUTFILE --cut_ga Pfam-A.hmm $INPUTFILE
```

Since the annotation with HMMER predicts any hit of a domain to the query sequence, it is possible that hits overlap and actually constitute a single domain. To collapse these hits, `cath_resolve_hits` version 0.16.2<sup>V.4</sup> was applied with the following call.

```
cath_resolve_hits.untuntu14.04 \  
  --input-format hmmsearch_domtblout \  
  --hits-text-to-file $OUTPUT $INPUT
```

The output was parsed so that for each transcript, the information of annotated domains in order of appearance (in reading direction of the transcript) was available as tab-separated list using a custom awk script.

The protein domain distribution across conservation classes was assessed as follows. For each gene, the ordered domain annotations were recorded individually (counting all as well as only unique domains) and screened for arrangements of two (pairs), three (triplets), or four (quartets) domains in a row with a sliding window approach implemented in a custom perl script. For example, a gene annotated with five domains could at maximum harbor five individual domains, four pairs, three triplets, and two quartets (illustrated in Fig. V.14). Of each category there will be less unique combinations depending on the number of repeated domains.

Following this step of recording the arrangements per gene, individual domains and arrangements were counted according to the gene's belonging to either

---

<sup>V.3</sup> Pfam database: <https://pfam.xfam.org>. Last accessed 30 Mar 2018.

<sup>V.4</sup> GitHub: ORENGO *et al.* 2018. `cath_resolve_hits` from the `cath-tools` suite, <https://github.com/UCLOrengoGroup/cath-tools/releases/tag/v0.16.2>. Last accessed 30 March 2018.

of the three conservation classes. The resulting lists were stored in the MySQL database and evaluated using Venn diagrams drawn using a custom R script and the VennDiagram package<sup>V.5</sup> in combination with Inkscape version 0.91<sup>V.6</sup>.

### 3.8 Phylogenetic tree

I gathered the phylogenetic relationships between the 26 species of the sample from the publications by MISOF *et al.* (2014) and PETERS *et al.* (2017). Also, as many estimated divergence times as available for the present sample were extracted and mapped to the obtained tree from these publications. Note that not all nodes could be assigned an age. Thus, the tree topology shown here has arbitrary branch lengths (Fig. V.1).

For each node, I determined the subtree, *i.e.*, which species descended from it, and stored this information in the MySQL database (table 'Species\_at\_nodes'). For future quick reference, another MySQL table was laid out, containing the count of species descending from each node (table 'Count\_at\_nodes').

### 3.9 Orthology prediction and partitioning of repertoires

The orthology prediction was provided by ROBERT M WATERHOUSE. He employed the stand-alone pipeline of OrthoDB9 (ORTHOPIPE version 6.2.5; dependencies: BRHCLUS version 2.2.2\_debug<sup>V.7</sup>; NCBI BLAST version 2.2.24, SWIPE version 2.0.12, ROGNES, 2011; and cd-hit version 4.6.8-2017-1208, FU *et al.*, 2012). This was done to ensure that the unpublished genome data remained offline. As general parameters, the minimum overlap for pair-wise alignments

---

<sup>V.5</sup> GitHub: CHEN H. 2018. R package VennDiagram, <https://github.com/cran/VennDiagram>. Last accessed 30 March 2018.

<sup>V.6</sup> Inkscape: <https://inkscape.org/en/>. Last accessed 30 March 2018.

<sup>V.7</sup> ORTHOPIPE and BRHCLUS are available from [www.orthodb.org/?page=software](http://www.orthodb.org/?page=software). Last accessed 30 Mar 2018.

(in the BRHCLUS algorithm) was set to 30 basepairs, and the maximum *e-value* considered in clustering of best reciprocal hits was 1.0e-3. The minimum percentage ID (`select_ID`) as cutoff for clustering intra-species sequences was set to 97.

The partitioning of the repertoires according to orthology prediction followed loosely the core-shell-cloud terminology as used by KOONIN (2011). The core partition corresponds to node 1 (see Fig. V.1 for node IDs). The shell consists of all nodes above node 1 that comprise more than two species (nodes 2–8, 12–16, 18–20, 22–24). The cloud comprises all nodes combining exactly two species (five nodes: 9, 10, 17, 21, 25) and all nonOG genes. Furthermore, the following analyses also respect duplicability, hereafter given as copy status, as the scheme is different from the orthology types described by WATERHOUSE *et al.* (2011). Based on the count of gene family members contributed by each species to the gene family in comparison to all other contributing species (obtained using a custom Perl script, see Appendix, E.2.3), four states are discriminated:

1. All species descending from the considered node have exactly one copy of the considered gene family. These genes are universal single-copy orthologs, or short *USCs*.
2. The considered species has exactly one copy of the gene family, while other species descending from the considered node have either more or no copies. Thus, the gene family may be not present in all species. These genes are called species-specific single-copy orthologs (*sSCs*).
3. The opposite case, where the considered species has more than one copy, while other species descending from the considered node have either no or exactly one copy, causes genes to be termed species-specific multicopy (*sMC*).
4. Lastly, if no orthologous gene is found in any other species of the sample, the gene is assigned the status *nonOG*.

I partitioned the resulting orthology prediction data (`/Cluster/seqs.og`) according to the presence of ortholog groups at the nodes of the tree (it was taken care that each OG was only considered once, at the highest node it occurred at), according to the conservation class (core/shell/cloud) and according to the copy status (single-copy in all/single-copy in species/multicopy in species). This was done using custom scripts written in Perl (E.2.3). The partition

information was also stored in the MySQL database (table ‘Transcript\_at\_node’; MySQL calls and queries: see Appendix, E.2.2).

Note that the requirement of USCs to include members from all species from the node is very strict, not accounting for potential artifacts of incomplete assembly and/or annotation (as, for example, done by WATERHOUSE *et al.*, 2011). Such cases fall into the copy states sSC and sMC. The genes assigned to copy status sMC might be universal, *i.e.*, present and multi-copy in all species.

### 3.10 Plotting

All plots were generated using R (R CORE TEAM, 2017) with the following main packages: ggplot2 (WICKHAM, 2009), wesanderson<sup>V.8</sup>, RColorBrewer<sup>V.9</sup>, ggtree (YU GUANGCHUANG *et al.*, 2017), ggribes<sup>V.10</sup>. The respective R scripts can be found in the Appendix, E.2.3. Plot finalization was done using Inkscape.

---

<sup>V.8</sup> GitHub: KARTHIK *et al.*. 2018. R package wesanderson, <https://github.com/karthik/wesanderson>. Last accessed 30 March 2018.

<sup>V.9</sup> GitHub: NEUWIRTH E. 2014. R package RColorBrewer, <https://github.com/cran/RColorBrewer>. Last accessed 30 March 2018.

<sup>V.10</sup> GitHub: WILKE K. 2018. R package ggribes, <https://github.com/cran/ggribes>. Last accessed 30 March 2018.

---

## Results

---

**S**INCE THIS WORK FOCUSED on the characterization of gene repertoires, I will regard the assemblies as given and not discuss results of assembly or masking of transposable and repetitive elements steps in detail. In general, assembly quality of the assessed species is very good considering standard metrics. Details on assembly quality are available upon request from ALEXANDER DONATH.

In total (*i.e.*, summing up the gene counts of all repertoires), 535,983 genes were analyzed. The size of the conservation classes is illustrated in Fig. V.2 (rectangles). Of the total gene count, 206,019 are assigned to the core class, 122,475 are shell genes, and 207,444 are classified as cloud; the cloud includes 200,890 genes being lineage-specific without orthologs in any other species, *i.e.*, have the copy status nonOG. Thus, examining the gene conservation across the considered species yields the same pattern as described by KOONIN (2011) and WATERHOUSE (2015): a large (comprising many genes) core, a smaller shell, and a cloud of similar size as the core.

## 4.1 The complete repertoires

### 4.1.1 Gene counts and BUSCO assessment

The species differ tremendously in their gene counts; the contribution to gene count by genes of the three conservation classes and four copy states is also variable (Fig. V.2, left side). However, all Aculeata (species IDs 10–17) have similar repertoire sizes of less than 20,000 genes. *S. ovinae*, the species with the smallest genome of the sample size, also has the smallest annotated gene repertoire. However, genome size does not correlate with gene count, as becomes obvious when comparing the gene count and genome sizes of *G. ashmeadi* (40,451/324 Mbp) and *P. grandior* (30,725/1,173 Mbp).

The predicted gene models cover the expected gene space very well, especially within the studied Hymenoptera (right side of Fig. V.2). No gene repertoire of these Hymenoptera contains less than 90 % of the complete benchmark single-copy orthologs (BUSCOs).

### 4.1.2 Overall gene structure parameter distributions

The distributions of five general gene structure parameters were recorded for the whole gene repertoires per species as ridge plots (Fig. V.3). These serve as baseline for comparisons with the same parameter distributions split according to conservation class and copy status (section V.4.3.2).

Overall, no outstanding general pattern regarding phylogenetic relationships is obvious among the structural parameter distributions. Small-scale deviations occur, like the following example. The median intron lengths of the two *Sepsis* species are bimodally distributed in a relatively small range compared to all other species.

## 4.2 Universal landscapes of ortholog groups

The assessment of ortholog groups, their composition as well as distribution in the landscape defined by universality and duplicability is done using a two-



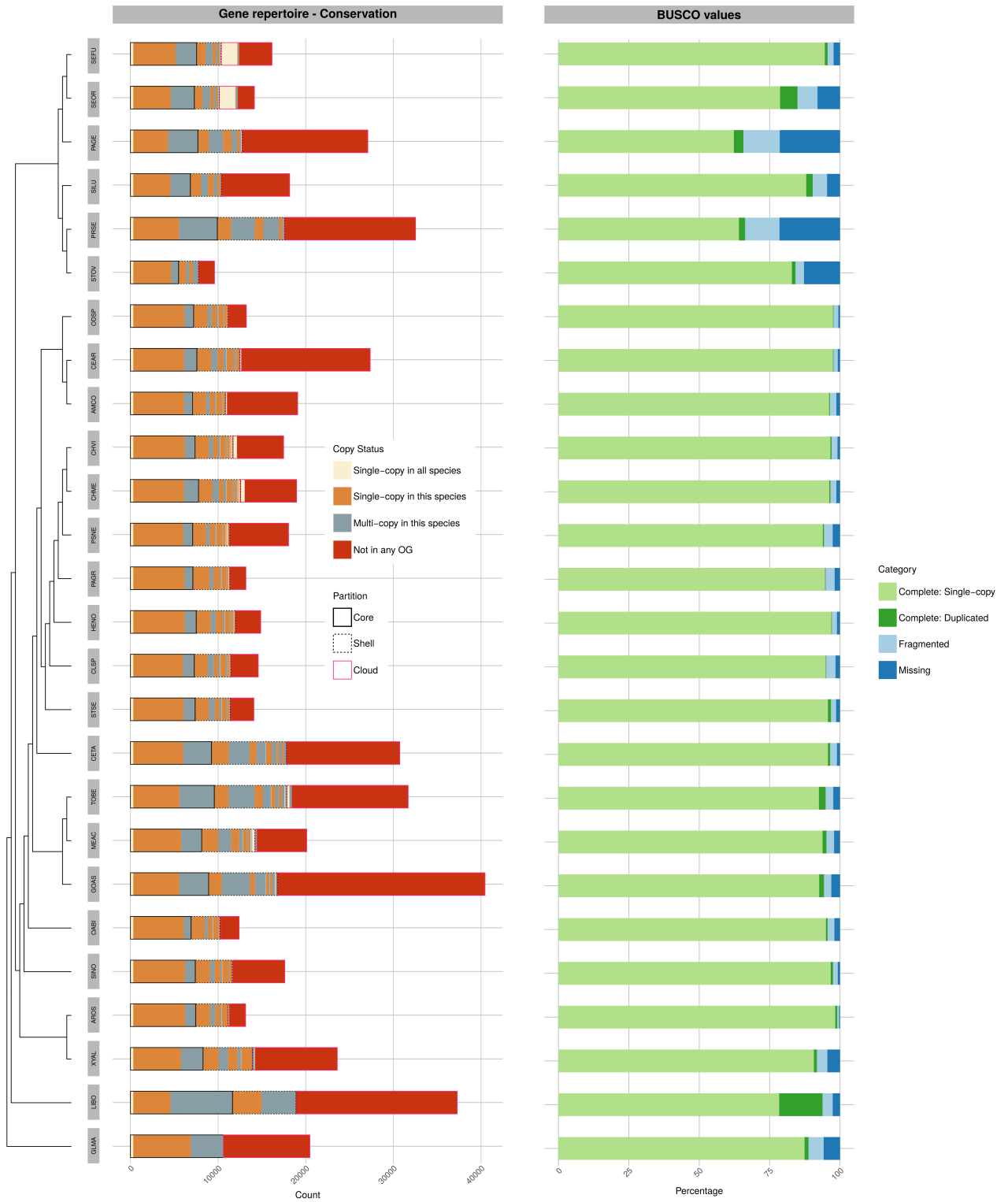


Figure V.2 – Gene repertoire and BUSCO values. (Continued on next page)

**Figure V.2 – Gene repertoire and BUSCO values.**

(Continued)

The phylogenetic tree is the same as given in Fig. V.1 and species abbreviations correspond to those given in Tab. V.1.

**Gene repertoire – Conservation:** This barplot shows the count of genes in each repertoire of the 26 species. Bar segments are colored according to their copy status (ivory: present and single-copy in all species sharing the respective ortholog group (OG), also termed USC, universal single-copy; orange: OG is present in some of the species of the respective node and single-copy in this species, also termed sSC, species-specific single-copy; grey: OG is present in some or all of the species of the respective node and multicopy in this species, also termed sMC, species-specific multicopy; red: not related to any OG, lineage specific genes).

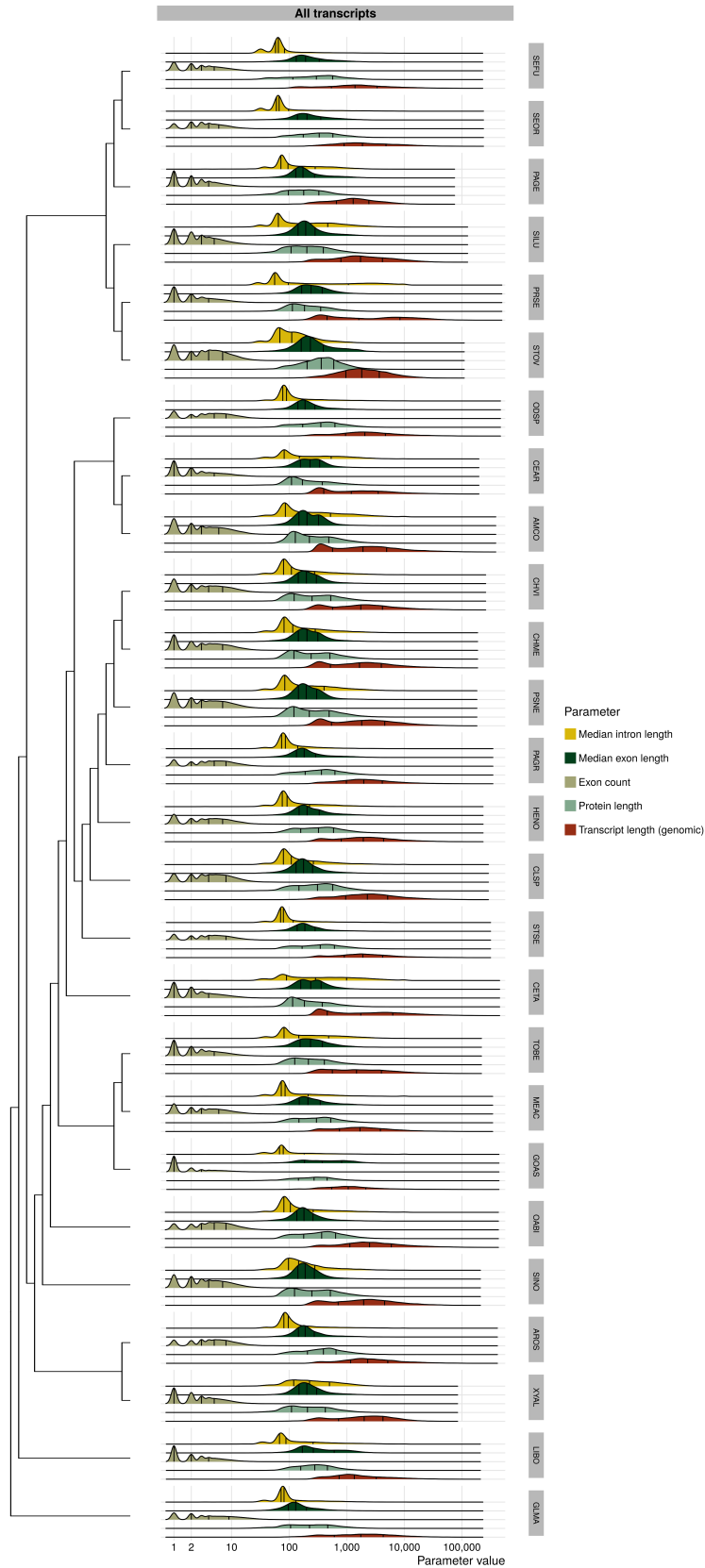
The copy status directly depends on the highest, *i.e.*, last common ancestral node at which the ortholog group is present. For example, a group found to have members in species 001 (*S. fulgens*) and species 024 (*X. alpigena*) is mapped to node 3, the last common ancestor of both species. Thus, the first three colors of the copy status repeat for each node at which orthologs have been found within the respective repertoire. NonOG genes (red) are not assigned to any node, but to the tips, *i.e.*, the species.

The gene repertoires were partitioned according to conservation. The core partition (solid rectangle) corresponds to node 1. The shell (dashed rectangle) consists of all nodes above node 1 that comprise more than two species. The cloud (pink rectangle) comprises all nodes combining exactly two species (five nodes) and all nonOG genes.

**BUSCO values:** The barplot illustrates the percentage of the benchmark universal single-copy orthologs (expected to be present in all species; lineage insecta\_odb9, n: 1658) retrieved in full length only once (complete single-copy) in the scrutinized species (light green), in full length but multiple times (complete duplicated, dark green), not in full length (fragmented, light blue), or not at all (missing, dark blue).

dimensional representation of the previously proposed (WATERHOUSE, 2015) landscape plots. To illustrate that these two graph types correspond closely, both were compared for the example of *C. viridula* (Fig. V.4).

Consistent patterns can be observed in all three heatmap types (A, B, C; Fig. V.5) across all species. In the original type (A; as proposed by WATERHOUSE, 2015), the pattern shown in the original publication for *Drosophila melanogaster* is recovered: many universal single-copies (top right corner) and many lineage-specific single-copies (bottom right corner) are present. This pattern fits the proposition of KOONIN (2011), the universal ‘bowl’: many core genes [top], few shell genes, and many cloud genes [bottom] can be distinguished. The two additional heatmap types (B, C) illustrate additional transformations of



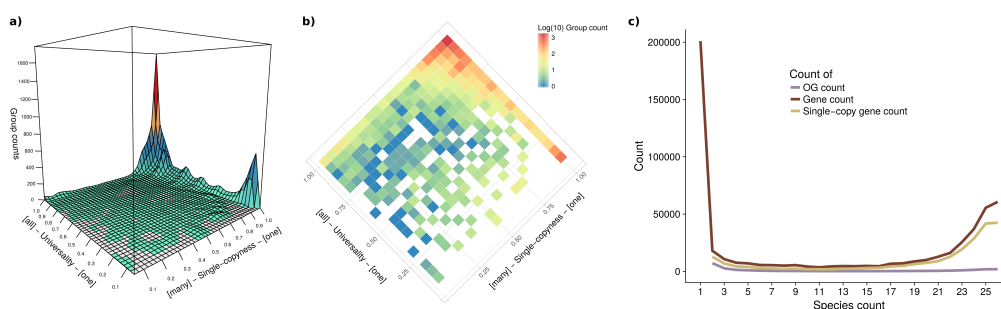
**Figure V.3 – Gene structural parameters for all transcripts.** (Continued on next page)

**Figure V.3 – Gene structural parameters for all transcripts.**

(Continued)

Five parameters (genomic transcript length [including exons and introns] in basepairs [bp], yellow; protein length in aa, dark green; exon count, khaki; median exon length in bp, turquoise green; median intron length in bp, red) compared for all transcripts of each species, ordered according to the phylogenetic tree (see also Fig. V.1).

Sample sizes (*i.e.*, the count of analyzed transcripts) correspond to the gene counts given in Tab. V.1.

**Figure V.4 – Comparison of two plot types illustrating the universality and duplicability of ortholog groups.**

Both plots depict the same data for the species *C. viridula*. The left graph (landscape plot) shows the group count on the z axis, while the right graph (heatmap) flattened the landscape to two dimensions while preserving the color code. This serves to illustrate how both types correspond.

Perl code for the landscape plot was kindly provided by Robert M Waterhouse (pers. comm. Mar 2018).

the analyzed data to get a more detailed understanding of the peculiarities of ortholog group phenomenology. Type B, which depicts copy counts instead of group counts by the color code, shows an expected pattern given the observation made in type A: there are higher counts of genes summed up for all ortholog groups present in the respective species in areas of lower single-copyyness (left side). The third type (C) shows in how many ortholog groups the respective species contributes most to the total gene counts in each bin defined by universality and copy-contribution. As would be expected, group counts are highest (red), where the species contributes copies to the count of an ortholog group that is shared by many species (top) but also highly duplicable (left).

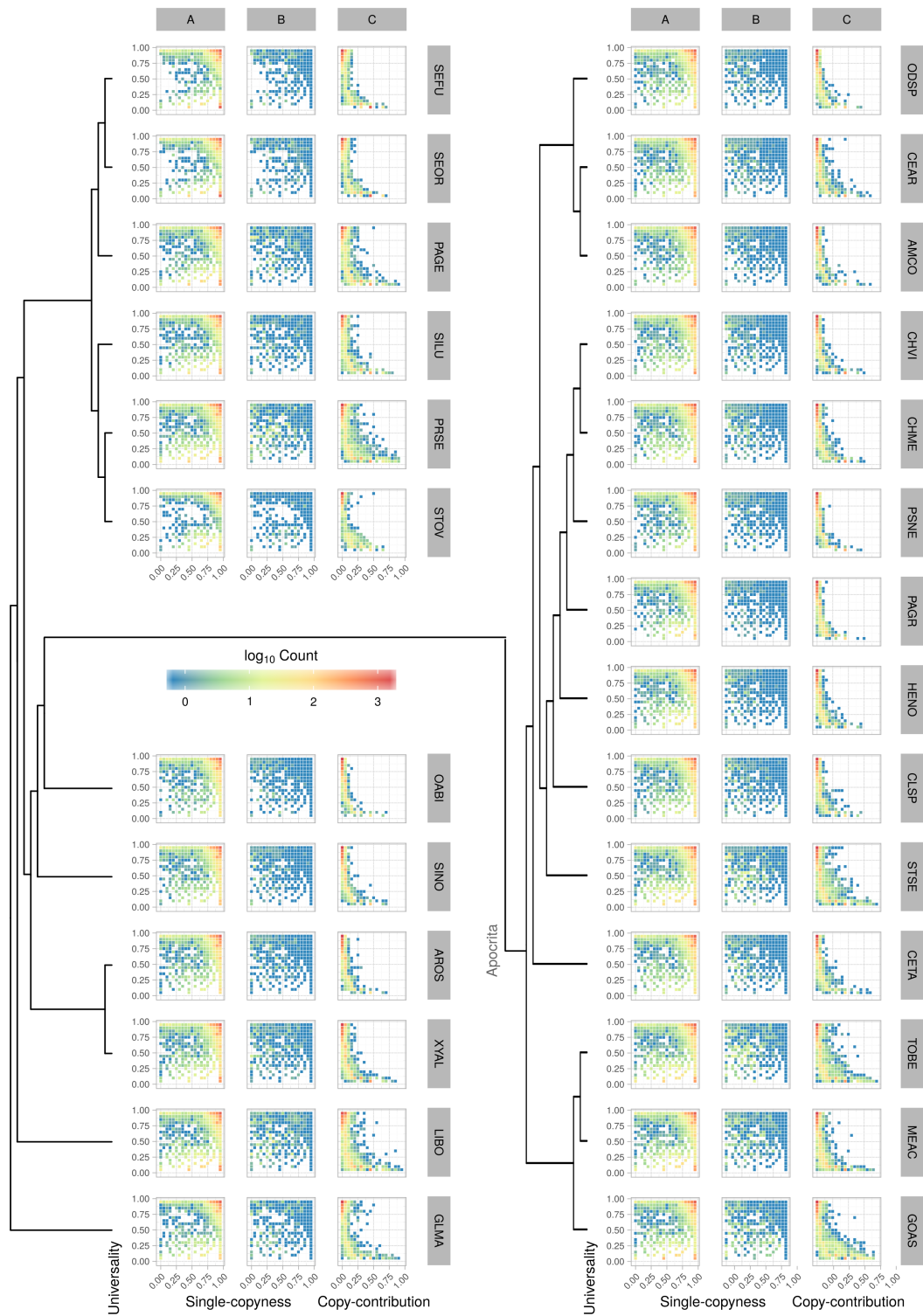


Figure V.5 – Ortholog groups show similar patterns of universality and single-copyness. (Continued on next page)

**Figure V.5 – Ortholog groups show similar patterns of universality and single-copyness.**

(Continued)

For each species (ordered according to the phylogeny given in Fig. V.1; note that the subtree of Apocrita (following node 13) is expanding on the right side), three heatmaps were generated: A, B, and C. B and C differ from A in one of two variables (Single-copyness/Copy-contribution, and Count). Each grid cell of the plot corresponds to a bin defined by the respective x and y values; counts include all ortholog groups (OGs) that meet the criteria defined by the bin.

**A – Original:** Following the landscape plot proposed by WATERHOUSE (2015), universality is calculated as the count of species sharing this OG divided by all species in the considered set (here, 26). Single-copyness (duplicability) is calculated as count of species that have one copy of this OG divided by the count of species sharing this OG (*i.e.*, have any number of copies). Count (depicted in log10 by the color scale) represents the count of OGs per bin, where the bin is defined by a value of universality between 0 (not shared between any species) and 1 (shared between all species) on the y axis, and a value of single-copyness between 0 (many species have many copies, this OG is highly duplicable) and 1 (all species sharing this OG have one copy, this OG is not duplicable) on the x axis.

**B – Copy counts:** The axes (universality, single-copyness) are the same as in A. Count now gives the number of genes in all OGs of a specific bin instead of the count of OGs.

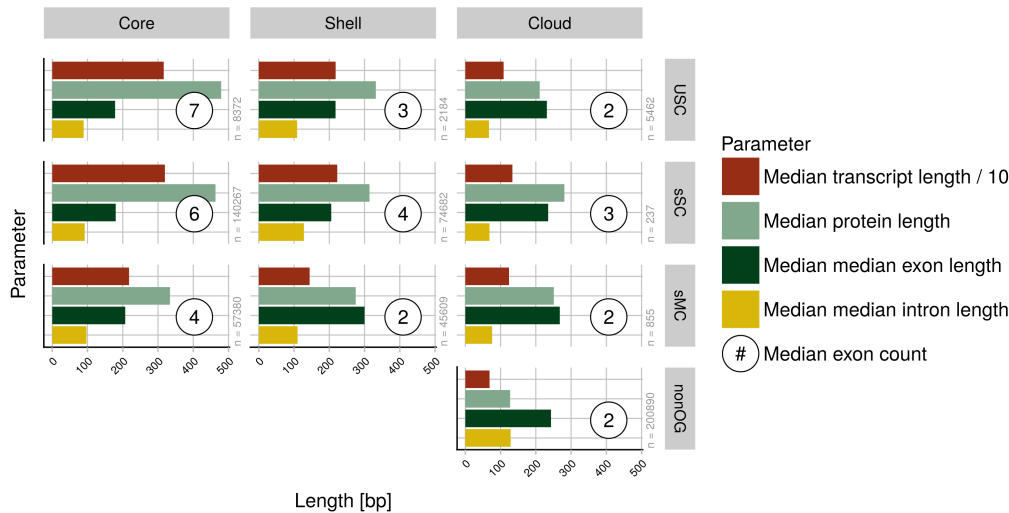
**C – Copy contribution:** The y axis (universality) and the count (OG count) are the same as in A. The x axis now depicts the copy contribution, which is calculated as the count of copies in this OG and this species divided by the count of copies in this OG (totaling all species' OGs). Values close to 0 mean that the species contributed few copies to the bin, 1 that all copies stem from this species.

For further interpretation, see text.

## 4.3 Comparing core, shell, and cloud

### 4.3.1 General gene structure medians

Using the medians over all transcripts, the five central gene structure parameters were compared between three conservation classes and the four copy states (Fig. V.6). Genes of the core have (in the median) longer transcripts and proteins than their copy status counterparts of the shell and core, but shorter exons and introns. This is the result of a higher median exon count. Thus, core genes seem to be complex and could in principle give rise to more splice variants. Comparing the three copy states within the core class, this pattern becomes apparent again: the universal single-copy orthologs, highly conserved



**Figure V.6 – Medians of five structural gene parameters compared across conservation classes and copy states.**

The same parameters as depicted in Fig. V.3, but given as median over all transcripts (*i.e.*, over all species) belonging to one of the three conservation classes (core, shell, cloud; for a definition refer to Fig. V.2) and one of the four copy states (USC: universal single-copy, sSC: species-specific single-copy, sMC: species-specific multicopy, nonOG: not in any ortholog group). The sample sizes are given for each set.

Note that median genomic transcript length (including introns and exons) is given as divided by 10 to allow a convenient, comparative overview.

genes present with only one copy in all of the 26 species in the sample, are longest and have the most introns.

Lineage-specific genes (cloud: nonOG) are (considering the median) the shortest compared to all other conservation class/copy status sets, and produce the shortest proteins, but have relatively long exons and introns (Fig. V.6).

### 4.3.2 Species-specific gene structure distributions

To elucidate differences between the three conservation classes (core, shell, and cloud) and between genes of different copy status (USC, sSC, sMC, nonOG), the five central gene structure parameters were analyzed in their full distribution

for each species' repertoire split into the according partitions. The following ridge plots allow to compare the distribution of one gene structure parameter for all species, conservation classes, and copy states at a time.

At the ridge plots, several interesting things can be noted. First, there is no clear pattern following taxonomy in all five parameters, although there are exceptions (for example very few genes with one exon in all core copy states in all Apocrita except *G. ashmeadi*, the species which diverged first from the remaining apocritans). Second, USC genes differ from sSC and sMC genes even in the core, and more so in the shell and cloud. Note that in the shell, sample sizes (transcript counts) of USC genes are rather small. For most species, there are only nonOG genes present in the cloud — except for those belonging to a node with two species at the tips.

A few further examples shall be highlighted. When comparing the distributions of transcript length (Fig. V.7) and protein length (Fig. V.8) of the core, *G. ashmeadi* is a striking case, in that a unimodal USC transcript length distribution results in a bimodal protein length distribution. This is most likely a result of marked exon count differences between the different transcripts balanced by exon and/or intron length effects.

It is also noticeable that most genes of the core class in all species have exons (Fig. V.9) and introns (Fig. V.11) of a very narrow length range (first to third quantile within 100–500 [exons]/80–200 [introns] base pairs) compared to the shell and cloud class genes (wider length distributions).

Another striking ridge deviation can be found when comparing exon counts (Fig. V.10) in the shell. Especially between USC genes of the different species, the distributions are highly dissimilar. Although the sSC and sMC exon count distributions among the species diverging from node 14 (Aculeata and *S. serrator*) can be considered a group by virtue of their similarity in comparison to other distributions, their USC distributions are astonishingly diverse. This is even more astonishing when considering that the largest USC exon count distribution difference can be found between the very recently diverged *Chrysis* species. However, it should be kept in mind that the sample size of USC genes is generally small in the shell class.

A final distinctive example are the differences between the cloud genes of the two recently diverged *Sepsis* species. The cloud genes of *S. fulgens* are in general



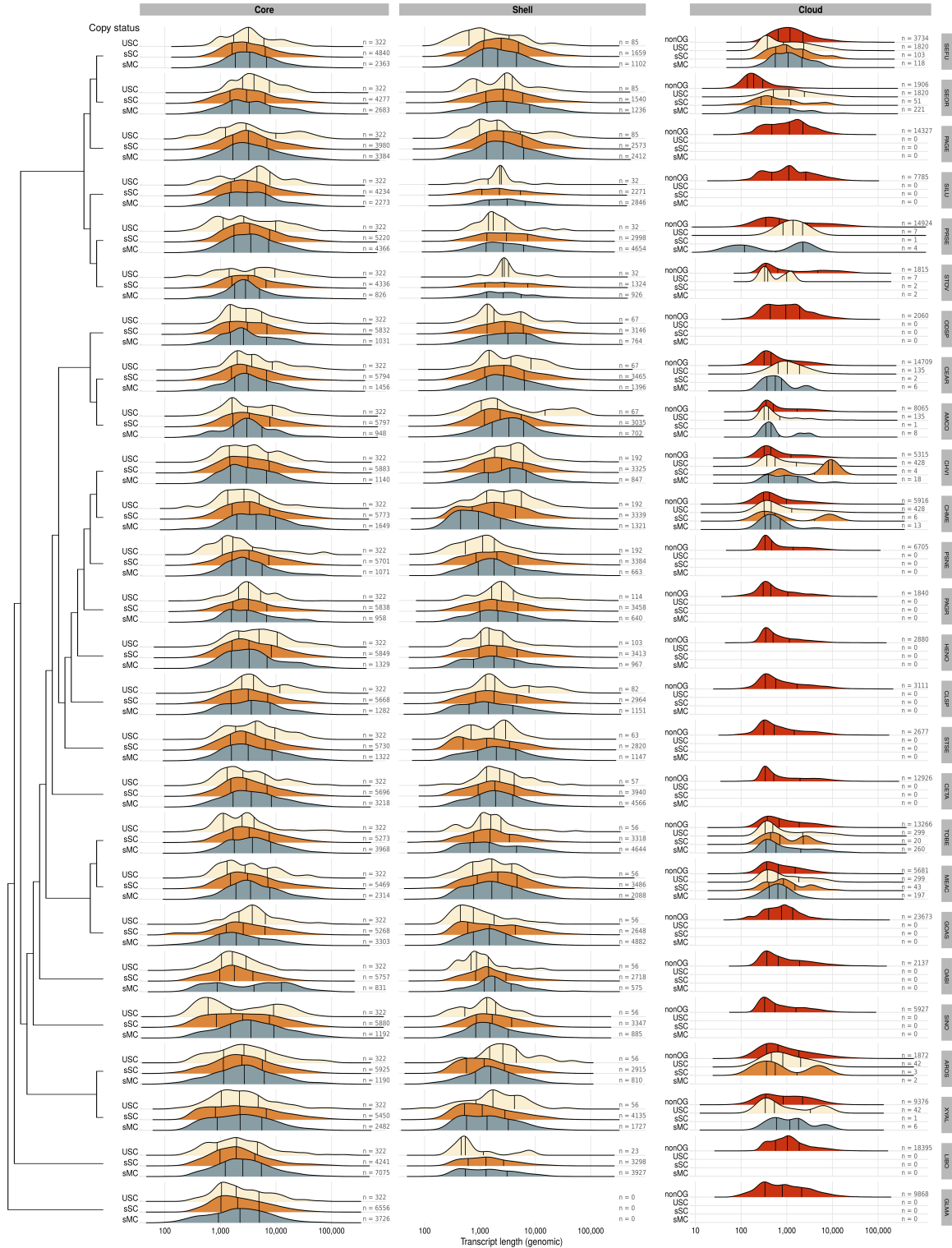


Figure V.7 – Transcript lengths of three conservation classes and three copy states. (Continued on next page)

**Figure V.7 – Transcript lengths of three conservation classes and three copy states.**

(Continued)

**Ridge plot** (also as caption for Figs. V.8, V.9, V.10, V.11): Compares the gene structure parameter for each species according to the transcript's assignment to one of the three conservation classes (*i.e.*, whether it belongs to an ortholog group [OG] shared at the core, shell, or cloud — for definitions of these, refer to Fig V.2), depicted in the facet columns, and its copy status (USC: universal single-copy, sSC: species-specific single-copy, sMC: species-specific multicopy, nonOG: lineage-specific genes without orthologs), depicted as rows within the species-specific row facets.

Vertical lines in each ridge indicate quantiles, the middle one is the ridge-specific median. For each ridge, the sample size is indicated.

Species are ordered according to the depicted phylogeny, see also Fig. V.1.

Note that 16 of the 26 species only have nonOG-genes in the cloud, since they do not belong to one of the five nodes ending in two tips. Note also that the outgroup species *G. marginata* does not possess any shell genes, since these are shared only among the ingroup species.

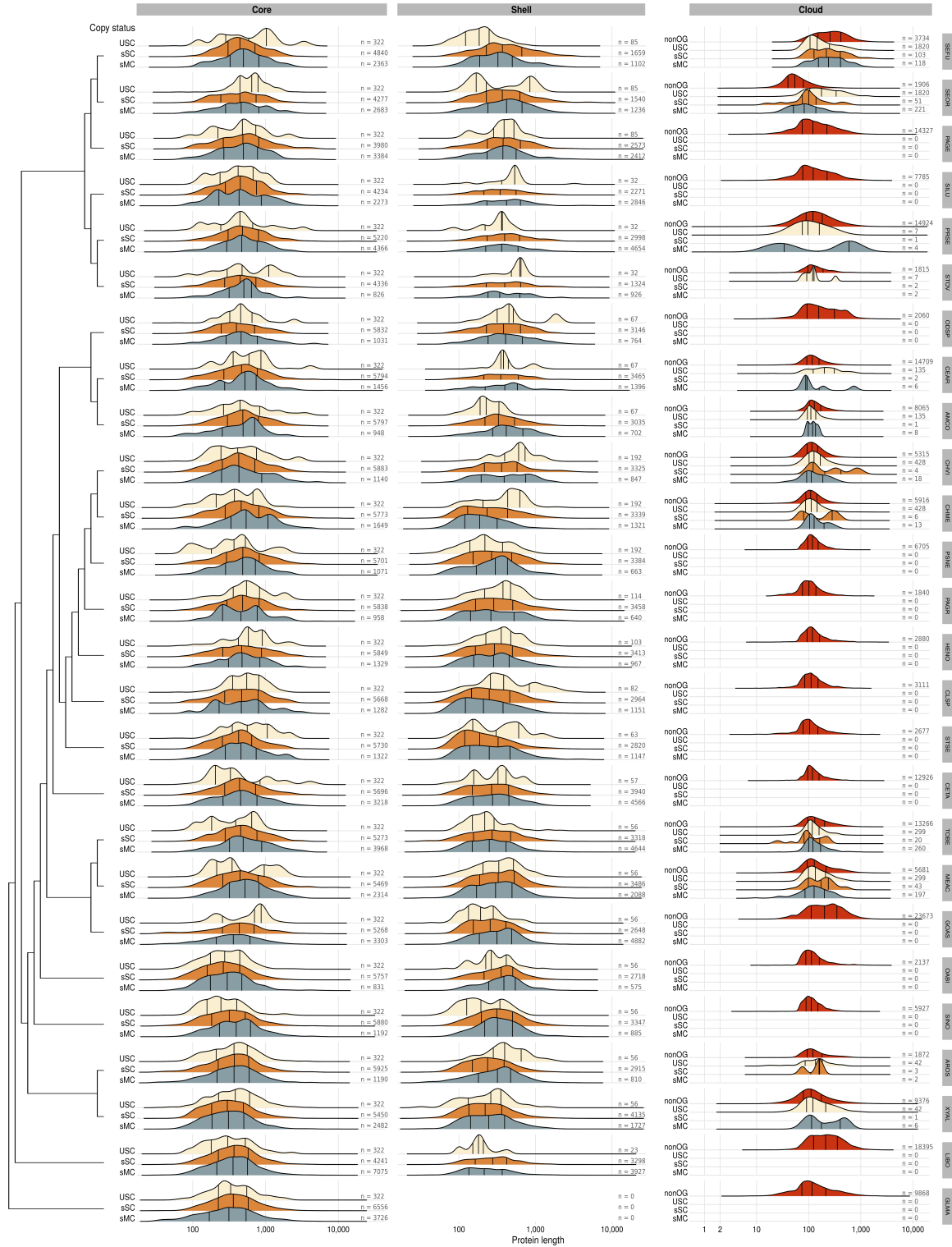
(also regarding the median) longer and produce longer transcripts, which is a result of a higher exon count compared to the cloud genes of *S. orthocnemis*. This is striking, because the difference in the distributions is larger than between any other species pair of the cloud.

### 4.3.3 Protein domain counts and arrangement diversity

In total (*i.e.*, across all species and ortholog groups), 144,766 genes (27 % of all genes) were annotated with one or more protein domains. Of these genes with domain annotation, 108,305 genes (74.8 % of genes with domains) are assigned to the core class, 26,421 (18.2 %) are shell genes, and 10,043 (6.9 %) are classified as cloud (9,515 [94.7 %] of these are nonOG genes).

The ratio of domains per gene is highest in the core: 1.83 domains/gene, while a ratio of 1.52 is found in the shell, and 1.28 in the cloud.

Figure V.12 illustrates the frequency of transcripts annotated with any number of domains, including consecutively repeated domains, between the three conservation classes and the four copy states. It becomes apparent that the universally conserved single-copy orthologs of the core do not have more than 14 domains in one transcript.



**Figure V.8 – Protein lengths of three conservation classes and three copy states.**  
 (Details see caption of Fig. V.7)

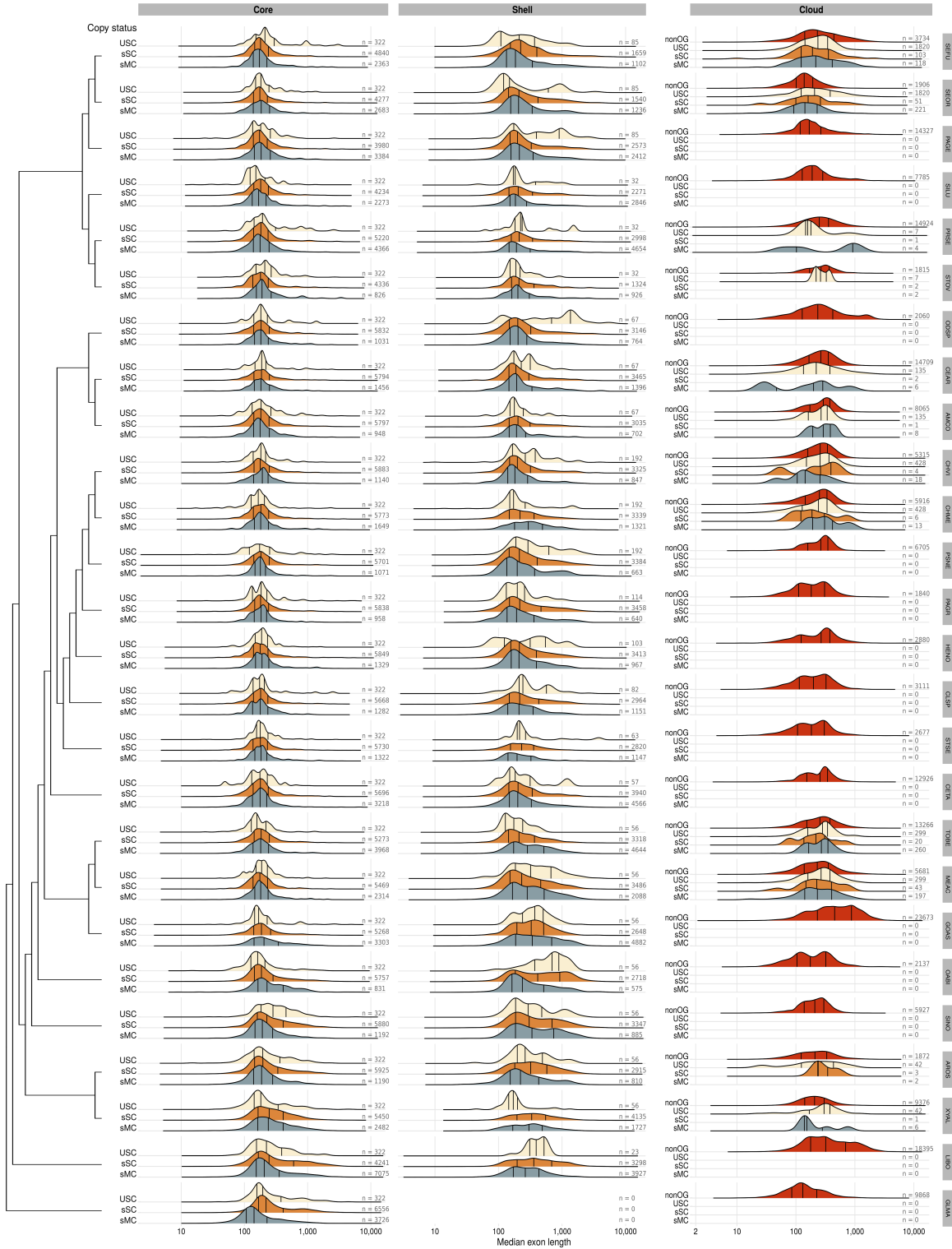


Figure V.9 – Exon lengths of three conservation classes and three copy states. (Details see caption of Fig. V.7)

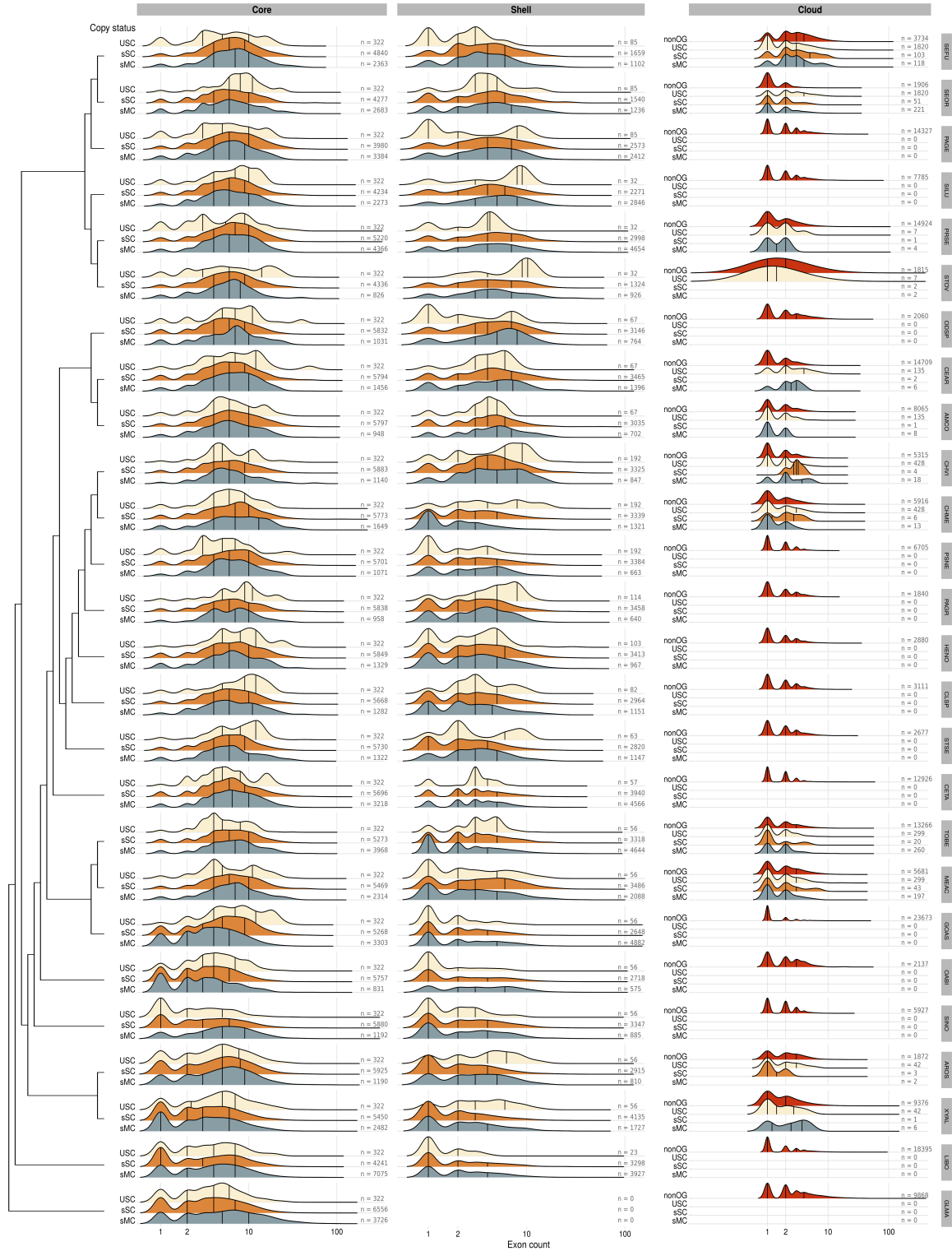


Figure V.10 – Exon counts of three conservation classes and three copy states.  
(Details see caption of Fig. V.7)

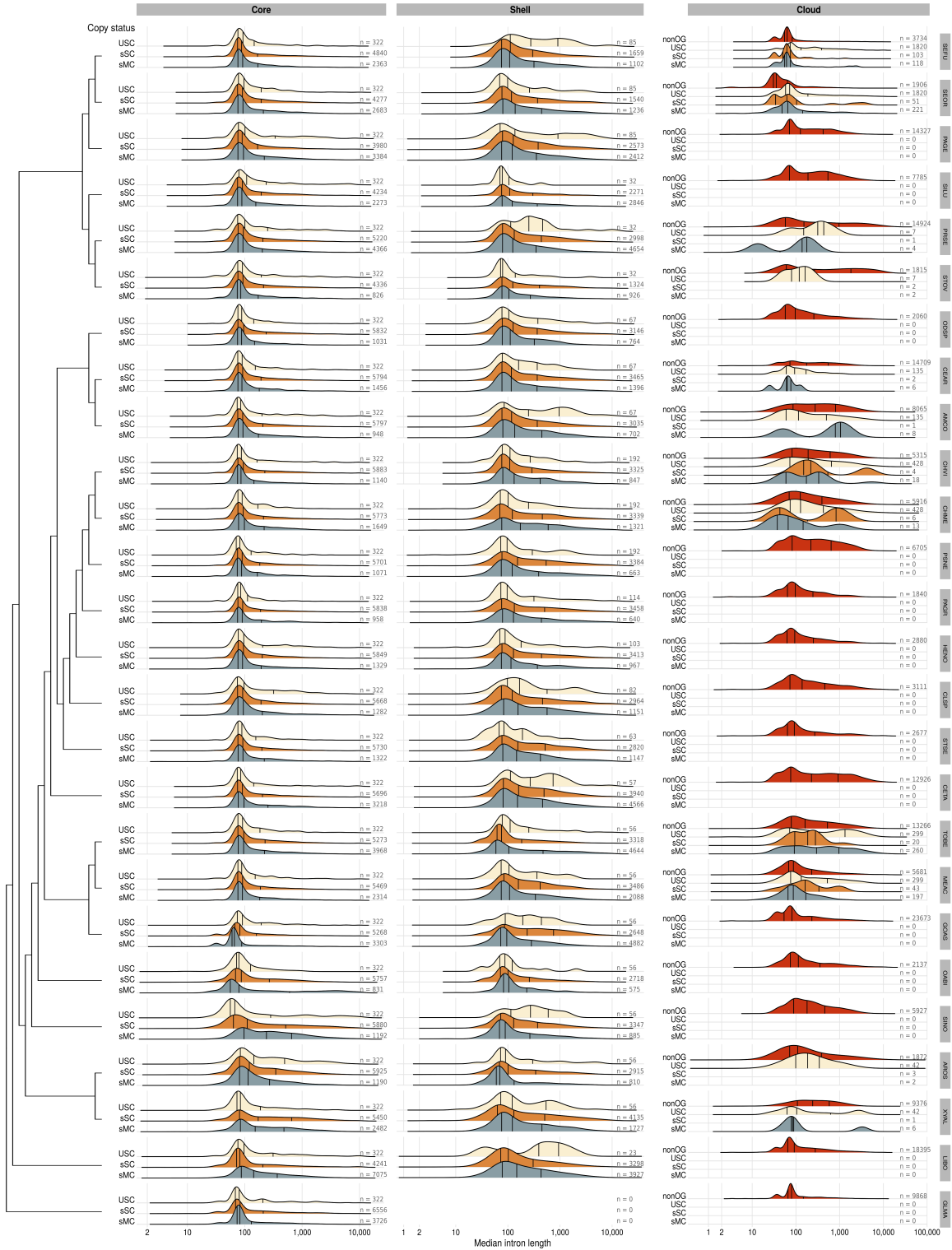
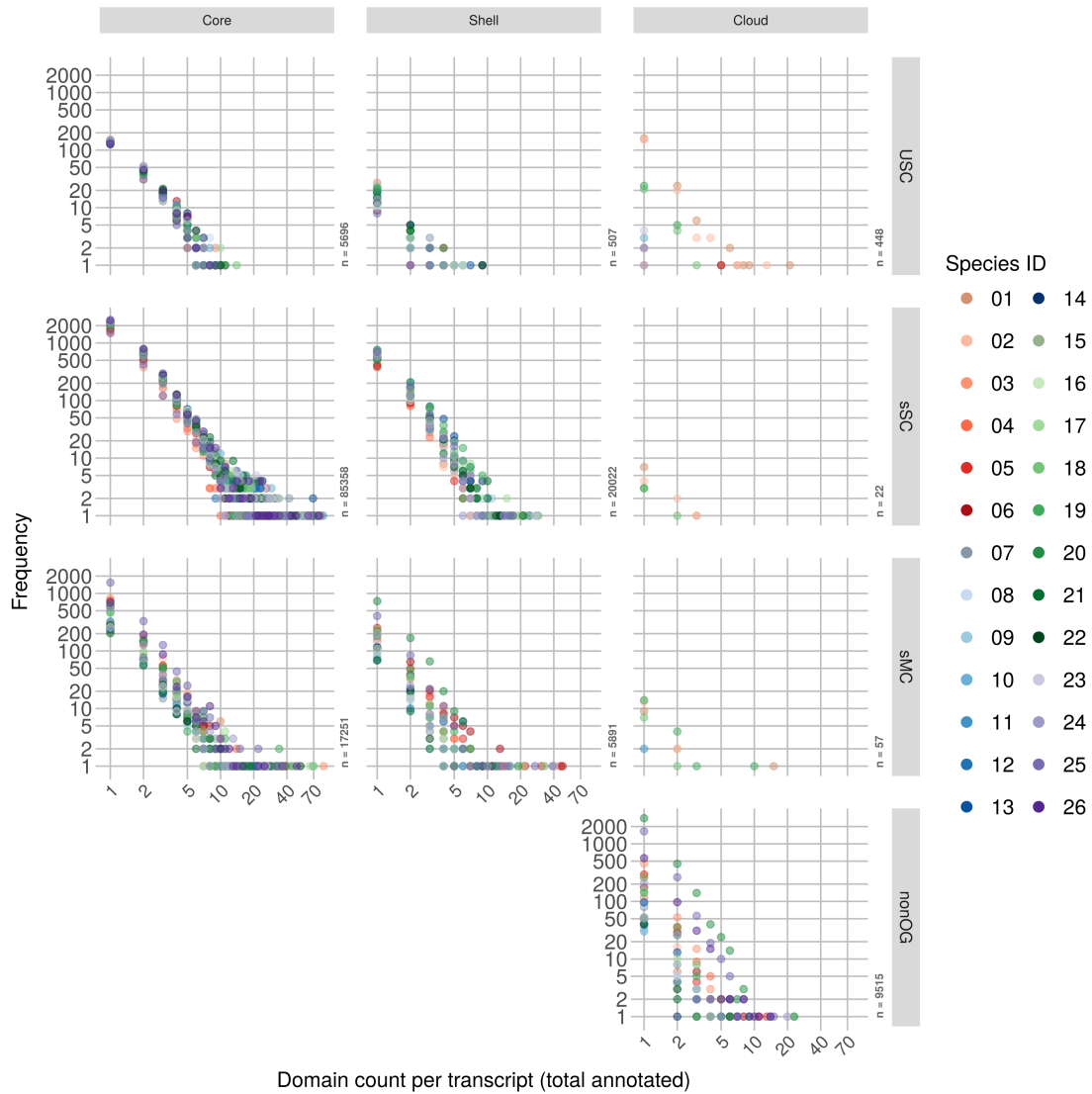


Figure V.11 – Intron lengths of three conservation classes and three copy states. (Details see caption of Fig. V.7)



**Figure V.12 – Protein domain count per transcript.**

The frequency (y axis) of transcripts with x domains (x axis) is presented for each copy status (rows; USC: universal single-copy, sSC: species-specific single-copy, sMC: species-specific multicopy, nonOG: lineage-specific genes without orthologs) and each conservation class (columns; core, shell, and cloud). The axes are logarithmically transformed, the values are not. Note the sample size information given for each facet.

Contrarily, the less strongly conserved subsets of the core, genes that can be absent in ingroup species and are either single- or multicopy in the specific species, are the ones with the highest domain count per transcript (up to 86 domains in sSC) in the whole comparison. The largest share of core genes with high domain count can be found in genes belonging to ortholog groups shared by 24–26 species (Fig. V.13).

Shell and cloud harbor much less transcripts with many domains, and the maximum domain count is clearly lower (15–25). It is interesting to note, though, that this maximum domain count is higher than in core USC genes.

To assess the diversity of protein domains and domain arrangements between the three conservation classes, unique individual domains and three kinds of arrangements (pairs, triplets, quartets) were analyzed for their presence in all transcripts assigned to either core, shell, or cloud. The ratio of unique domains per gene is highest in the cloud (0.54), while both shell (0.13) and core (0.05) have strikingly smaller ratios. This indicates a higher repetition of domains in core genes.

Regarding individual protein domains, a high diversity can be observed in all three conservation classes, although much of the shell's diversity is shared, while most of the diversity found in core and cloud is specific to these classes (Fig. V.14). Thus, the lineage- and species-specific genes of the cloud harbor many domains not found in the remaining species of the analyzed species sample and can be considered novel. Conversely, domains found in the core can be considered old, their origin most likely dates back to at least 557 million years ago. Almost two thirds of these old domains are also found in the cloud (core-cloud-overlap). Matching the line of evidence that shell and cloud genes generally have less domains (Fig. V.12), the diversity in arrangements is much smaller in shell and cloud than in the core. Thus, the potential arrangement diversity is more constrained by the count of domains per gene than by the uniqueness of novel domains. Note that repetitive domain arrangements decrease the actually observable count of unique arrangements in comparison to the theoretically possible given the count of individual domains.



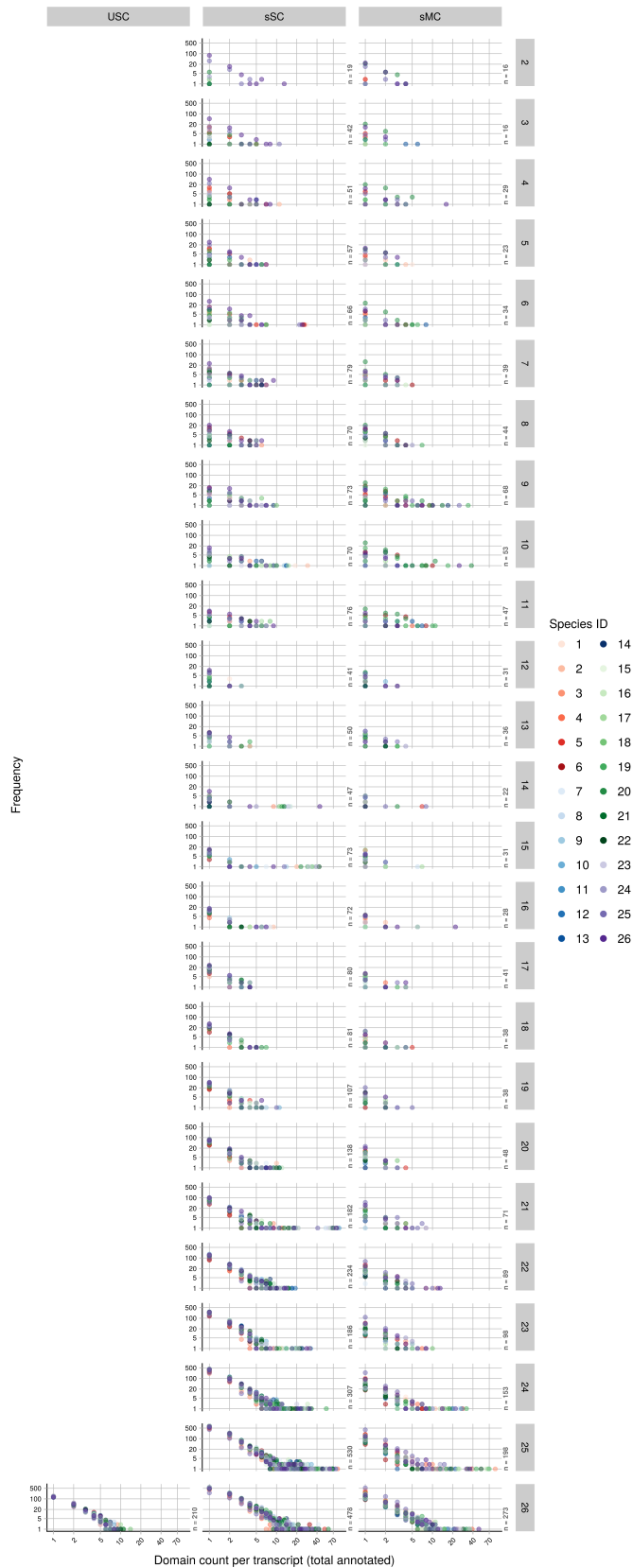


Figure V.13 – Core-specific protein domain count per transcript, split by number of species sharing the respective ortholog group. (Continued on next page)

**Figure V.13 – Core-specific protein domain count per transcript, split by number of species sharing the respective ortholog group.**

(Continued)

The same data that is also displayed in the first row of Fig. V.12 (*i.e.*, comprising only genes of the core class) is split according to the count of species sharing the respective ortholog group each transcript is affiliated to (facet rows; 2 species at the top, 26 at the bottom).

The frequency (y axis) of transcripts with x domains (x axis) is presented for each copy status (columns; USC: universal single-copy, sSC: species-specific single-copy, sMC: species-specific multicopy). Per definition, the USC copy state is only present in the last facet row, as these genes are shared by all 26 species of the sample.

The axes are logarithmically transformed, the values are not. Note the sample size information given for each facet.

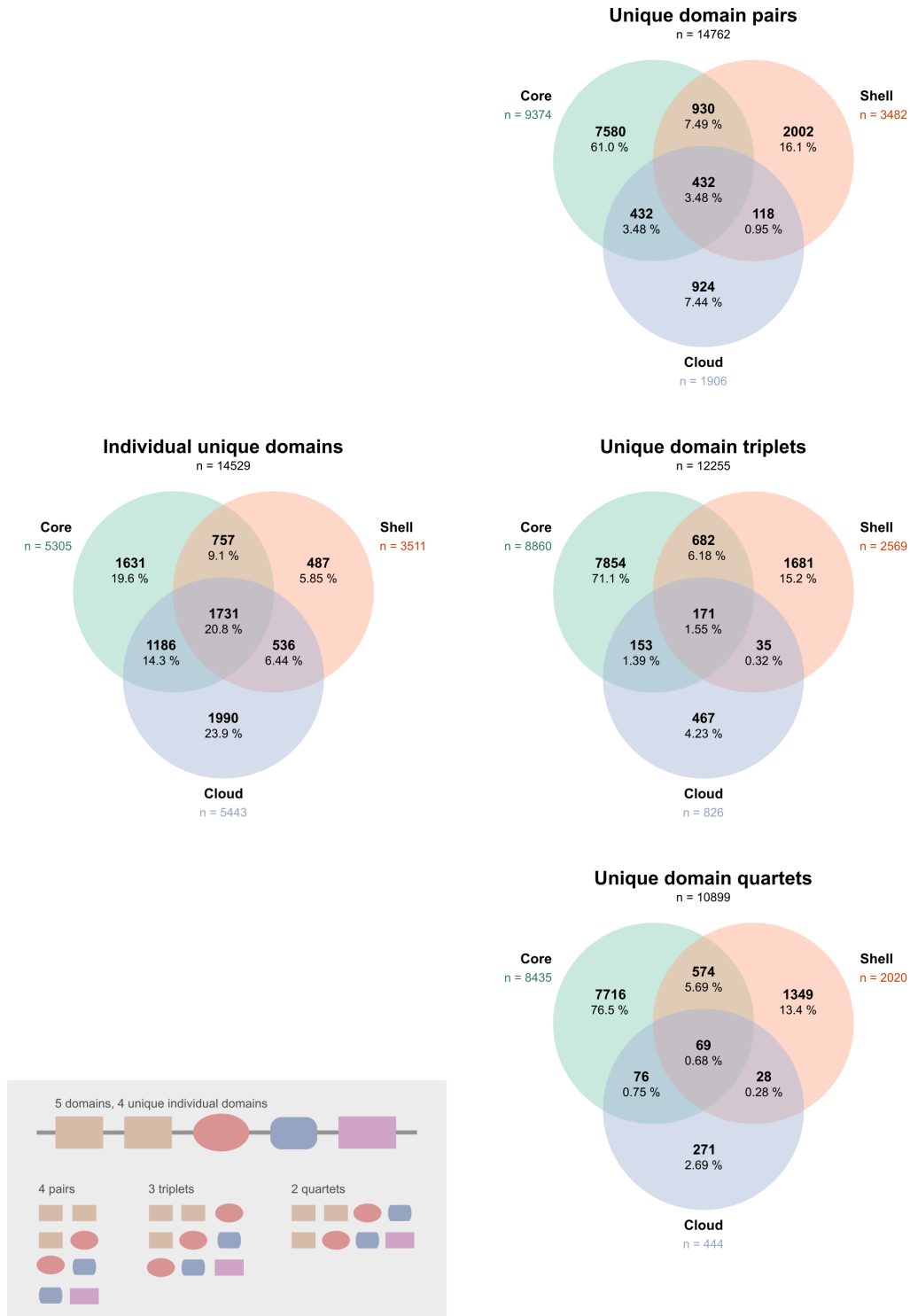


Figure V.14 – Protein domain and arrangement diversity. (Continued on next page)

**Figure V.14 – Protein domain and arrangement diversity.**

(Continued)

The Venn diagrams portray the diversity of unique domains (left) and arrangements (right; top: pairs, middle: triplets, bottom: quartets) at each conservation class (green: core, red: shell, blue: cloud). Additionally, an example of domain composition and the resulting arrangements is given (bottom left box).

For each Venn diagram, the total count of unique domains/arrangements is given below the title. The total count of these belonging to either of the conservation classes is given in the class color below the class name. Percentages refer to the all-embracing total count (not the class-specific).

Note that domains can be repeated along a gene, and depending on the amount of repetition, the amount of unique domains and arrangements found per gene decreases.

---

## Discussion

---

**I**N THE QUEST TO UNDERSTAND the generation of novelty from the viewpoint of gene repertoire changes, one specific route was followed in this study. The protein-coding gene repertoires of 26 newly sequenced species in a Hymenoptera-centric sample were analyzed with respect to their conservation across the phylogeny. This required the prediction of orthology relationships among the gene repertoires, whereafter genes could be classified according to their affiliation to an ortholog group, being a member of the core, shell, or cloud and being present in one or more copies in the considered species.

Potential sources of bias in this analysis are (1) sequencing and assembly; (2) repeat masking; (3) protein-coding gene and protein domain annotation; and (4) orthology prediction. Sequencing and assembly might influence the here presented results, because the fragmentation/completeness of a genome assembly directly affects the annotation of genes: for example, genes split across scaffolds falsely increase the count of shorter genes (see also Fig. III.1). RNAseq sequencing also affects the annotation of protein-coding genes: lowly expressed or highly tissue-specific genes might be less well covered and thus missing from the annotation. Repeat masking also might induce a falsely increased

gene count if repetitive sequences were left unmasked. Gene and domain annotations themselves can introduce bias depending on cutoff thresholds (see also Fig. III.4). Orthology prediction is a source of potential bias in that the count and size of ortholog groups depends on clustering settings: how coarse or fine-grained is the setting to constitute separate groups.

Applying the same methods for all species in the set ensures that all errors are systematic and the data is comparable. Thus, confidence is high that the observed patterns are consistent even if bias is present, although individual genes are likely missing or split.

## 5.1 Universal patterns of conservation

My analyses corroborate the proposed universality of distribution patterns of ortholog groups across a landscape of duplicability and universality (WATERHOUSE, 2015). Even when analyzed in more detail (*i.e.*, using the copy count of orthologs instead of the group count, and using the contribution of copies instead of the single-copyness) the observed patterns hold for all analyzed species. These findings vindicate the assumption that underlying laws of gene repertoire evolution and conservation are universal. Formulating these laws will be a future challenge.

## 5.2 Characteristics of conservation classes

Orthology is always established relative to the considered node and its identification depends on the available data basis, *i.e.*, genes from different species to compare. Note that a gene with the copy status nonOG is not necessarily a young gene. Many very old lineages are represented only by one species in our Hymenoptera-centric sample, for example, nonOG-genes of *S. noctilio* may be as old as the lineage, 264 million years (MISOFF *et al.*, 2014). It is likely that within the Siricoidea, orthologs for many of the 'orphan' genes in my analysis can be found when using a denser species sample for comparison, and thus are actually 'taxon-restricted genes' (KHALTURIN *et al.*, 2009). Nonetheless, the nonOG genes certainly comprise also very young genes, as becomes obvious

when regarding the count of nonOG genes in the *Chrysis* species (5,315 in *C. viridula*, 5,916 in *C. mediata*), which diverged less than 70 million years ago.

### 5.2.1 Gene structures

I find that genes classified as core are generally longer and produce longer proteins due to a higher complexity (more, but relatively short, exons and introns) than genes of the cloud. This is in line with previous evidence (CLARK *et al.*, 2007; LIPMAN *et al.*, 2002; WASMUTH *et al.*, 2008; WOLF *et al.*, 2009; YANG *et al.*, 2013), but note that previous research also revealed deviations from this ‘universal length difference’ (TATARINOVA *et al.*, 2016).

New, compared to previous studies, are the following findings: (1) the differences in (almost all) structure distributions are considerable between copy states; (2) there is astonishing variation in the distributions of transcript and protein lengths as well as in exon count, but not in exon and intron length in the core, when comparing species and copy states; (3) the variation of distributions of transcript and protein lengths is large in the cloud, while exon count variation is smaller compared to shell and core, and the breadth of exon and intron length distributions is wider than in the core.

The correlation of protein length and conservation has been vaguely accredited to functional relevance (LIPMAN *et al.*, 2002). No clear hypothesis explaining the rise and maintenance of the observed patterns has been proposed. Hence, I will outline some general thoughts regarding possible explanations of the relationship of length and complexity to conservation of protein-coding genes.

It can be expected that aging genes (*i.e.*, genes that are retained after their origination) become longer over time. This is based on the assumption that new genes are less likely to be long and/or complex when originating (TAUTZ and DOMAZET-LOŠO, 2011; WISSELER *et al.*, 2013). However, it can also be assumed that there are natural limits to gene growth, imposed by physical limits like cell space and by cost limitations related to transcription, replication, or product toxicity (*e.g.*, DRUMMOND and WILKE, 2008). It has been suggested that intron gain is adaptive (CARMEL *et al.*, 2007), thus playing a role in the elongation of retained genes.

Another naive hypothesis is that core genes are (functionally) important, otherwise they would not be (detectably) conserved over very large time scales as observed (*e.g.*, JORDAN *et al.*, 2002). The high complexity and length of these apparently important genes has benefits and drawbacks. One advantage is the possibility to produce more alternative splice variants, thereby improving the ratio of proteins that can be encoded to the required nucleotide sequence length (similar to a case study in *Volvox*, KIANIANMOMENI *et al.*, 2014). It has been shown that long genes tend to have more splice forms (in combination with high expression and low duplication rates; GRISHKEVICH and YANAI, 2014; ROUX and ROBINSON-RECHAVI, 2011). A disadvantage of high complexity and length is the theoretically higher probability of deleterious mutations and insertions of transposable elements. Furthermore, the sheer length of a gene theoretically also impacts replication and transcription speed and accuracy (here, reduced intron size is advantageous; CASTILLO-DAVIS *et al.*, 2002). Additionally, it has been found that the most conserved orthologs show also the highest DNA methylation levels in insects, potentially playing a role in reducing transcriptional noise (PROVATARIS *et al.*, 2018). This is in line with the suggestion that DNA methylation and gene expression regulation are interconnected and (partially) drive gene length evolution (ZENG and YI, 2010).

Intuitively, it seems that the costs of maintaining the integrity of long and complex genes (so that orthology can be established even after 550 million years of species divergence) are high and require considerable selective pressure. This implies that the advantages must outweigh the costs. Future work will have to elucidate both sides of the trade-off in detail to formulate a hypothesis explaining the maintenance of complex orthologs over long time spans.

### 5.2.2 Protein domains

In my analysis, a large share of unique domains is only found in the cloud and can be considered novel in the respective lineage. Note, however, that this might change with a denser species sample in which each lineage is represented by more than one species (KHALTURIN *et al.*, 2009; WASMUTH *et al.*, 2008); currently, there are single species representing very old lineages, *e.g.*, *L. bostrychophila* represents Psocodea, the order split 359 million years ago from



Holometabola (MISOF *et al.*, 2014). Contrastingly, I find few unique domain arrangements in the cloud; the highest arrangement diversity is confined to the core. This seems to contradict the pattern observed by GABALDÓN (2005), stating that most protein domains are ancient, while most combinations are lineage-specific. However, domain arrangements (just like genes) become longer over evolutionary time scales, as well as more diverse (BJÖRKLUND *et al.*, 2005; EKMAN *et al.*, 2005; ITOH *et al.*, 2007; M WANG and CAETANO-ANOLLÉS, 2009; Z WANG *et al.*, 2012). My finding reflects the larger ratio of annotated domains per gene in the core, although universal single-copy orthologs appear to be restricted to comparatively few domains per gene (less than most cloud genes).

The circumstance that the maximum domain count is higher in shell genes than in core-USC genes is surprising, as it is expected that novel genes are short. It is conceivable, though, that the novel genes with many domains originated *de novo* from fusion of other genes (KUMMERFELD and TEICHMANN, 2005). Alternatively, the genes with many domains might actually be core genes of, for example, Diptera (in the case of the two compared *Sepsis* species); note that the genes with most domains in the cloud-USC set are sepsid genes. Future work including a denser sample of sister species will allow to disentangle core, shell, and cloud genes at a higher resolution.

The here described differences of domain diversity in core, shell, and cloud provide not only an indication that variability in gene structure and domain content is related to a gene's conservation and to species evolution, but provide also starting points to study these relationships in detail.

### 5.3 Future directions

It will be interesting to study the diversity of protein domains and domain arrangements with respect to domain identity to see, for example, whether the domains of the universal single-copy orthologs of the core belong to a specific functional class or how the restriction to the comparatively few domains per gene in this subset could be explained.

Another line of research will be the incorporation of gene family size variation along the phylogeny – gene turnover – and the reconstruction of ancestral gene

(family) content to allow tracing of gene repertoire change along the phylogeny. This requires a fully dated tree, which will be available for this species set very soon.

The estimation of gene turnover rates along the phylogeny also allows to correlate turnover to other genomic phenomena, which will permit to elucidate their role in facilitating repertoire change. Namely large segmental duplications and direct and indirect effects of transposable element activity are promising potential drivers of change in protein-coding gene repertoires.

---

## Conclusion

---

**T**HE COMPARISON OF THE PREVIOUSLY PROPOSED conservation classes (core, shell, and cloud) in the aspects of gene structure and protein domain diversity provides first footholds for a detailed investigation of gene repertoire changes with respect to conservation/variation mechanisms.

Here it is shown that the conservation classes differ in gene structure and protein domain diversity, and that differences regarding these characteristics even extend to the copy status of gene family members. The underlying mechanisms and driving forces are yet to be discovered. When analyzing gene repertoires in evolutionary context, universality and duplicability of gene families should be taken into account.

Further research on gene family evolution based on turnover estimation and considering protein functions will help to elucidate mechanisms and drivers of evolutionary change in protein-coding gene repertoires.



---

## Bibliography V

---

- APIC, G and RB RUSSELL (2010). **Domain Recombination: A Workhorse for Evolutionary Innovation**. *Sci. Signal.* 3.139, pe30–pe30. DOI: 10.1126/scisignal.3139pe30 (cit. on p. 161).
- BARNETT, DW, EK GARRISON, AR QUINLAN, MP STRÖMBERG, and GT MARTH (2011). **BamTools: a C++ API and toolkit for analyzing and managing BAM files**. *Bioinformatics* 27.12, pp. 1691–1692. DOI: 10.1093/bioinformatics/btr174 (cit. on pp. 168, 169).
- BEMM, F, D BECKER, C LARISCH, I KREUZER, M ESCALANTE-PEREZ, WX SCHULZE, M ANKENBRAND, ALVd WEYER, E KROL, KA AL-RASHEID, A MITHÖFER, AP WEBER, J SCHULTZ, and R HEDRICH (2016). **Venus flytrap carnivorous lifestyle builds on herbivore defense strategies**. *Genome Research*. DOI: 10.1101/gr.202200.115 (cit. on p. 159).
- BESEMER, J and M BORODOVSKY (2005). **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses**. *Nucleic Acids Research* 33 (suppl 2), W451–W454. DOI: 10.1093/nar/gki487 (cit. on p. 169).
- BJÖRKLUND, ÅK, D EKMAN, S LIGHT, J FREY-SKÖTT, and A ELOFSSON (2005). **Domain Rearrangements in Protein Evolution**. *Journal of Molecular Biology* 353.4, pp. 911–923. DOI: 10.1016/j.jmb.2005.08.067 (cit. on p. 201).
- BUTLER, J, I MACCALLUM, M KLEBER, IA SHLYAKHTER, MK BELMONTE, ES LANDER, C NUSBAUM, and DB JAFFE (2008). **ALLPATHS: De novo assembly of whole-genome shotgun microreads**. *Genome Research* 18.5, pp. 810–820. DOI: 10.1101/gr.7337908 (cit. on p. 167).
- CAI, JJ and DA PETROV (2010). **Relaxed Purifying Selection and Possibly High Rate of Adaptation in Primate Lineage-Specific Genes**. *Genome Biology and Evolution* 2, pp. 393–409. DOI: 10.1093/gbe/evq019 (cit. on p. 160).
- CAMACHO, C, G COULOURIS, V AVAGYAN, N MA, J PAPADOPOULOS, K BEALER, and TL MADDEN (2009). **BLAST+: architecture and applications**. *BMC Bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421 (cit. on pp. 169, 170).

- CARMEL, L, IB ROGOZIN, YI WOLF, and EV KOONIN (2007). **Evolutionarily conserved genes preferentially accumulate introns.** *Genome Research* 17.7, pp. 1045–1050. DOI: 10.1101/gr.5978207 (cit. on p. 199).
- CARRETERO-PAULET, L, P LIBRADO, TH CHANG, E IBARRA-LACLETTE, L HERRERA-ESTRELLA, J ROZAS, and VA ALBERT (2015). **High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant *Utricularia gibba*.** *Molecular Biology and Evolution* 32.5, pp. 1284–1295. DOI: 10.1093/molbev/msv020 (cit. on p. 159).
- CASTILLO-DAVIS, CI, SL MEKHEDOV, DL HARTL, EV KOONIN, and FA KONDRASHOV (2002). **Selection for short introns in highly expressed genes.** *Nature Genetics* 31.4, pp. 415–418. DOI: 10.1038/ng940 (cit. on p. 200).
- CLARK, AG *et al.* (2007). **Evolution of genes and genomes on the *Drosophila* phylogeny.** *Nature* 450.7167, pp. 203–218. DOI: 10.1038/nature06341 (cit. on pp. 158, 160, 199).
- DANCHIN, A (2009). **Bacteria as computers making computers.** *FEMS Microbiology Reviews* 33.1, pp. 3–26. DOI: 10.1111/j.1574-6976.2008.00137.x (cit. on p. 160).
- DEMUTH, JP, TD BIE, JE STAJICH, N CRISTIANINI, and MW HAHN (2006). **The Evolution of Mammalian Gene Families.** *PLoS ONE* 1.1, e85. DOI: 10.1371/journal.pone.0000085 (cit. on p. 159).
- DRUMMOND, DA and CO WILKE (2008). **Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution.** *Cell* 134.2, pp. 341–352. DOI: 10.1016/j.cell.2008.05.042 (cit. on p. 199).
- EKMAN, D, ÅK BJÖRKLUND, J FREY-SKÖTT, and A ELOFSSON (2005). **Multi-domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions.** *Journal of Molecular Biology* 348.1, pp. 231–243. DOI: 10.1016/j.jmb.2005.02.007 (cit. on p. 201).
- FU, L, B NIU, Z ZHU, S WU, and W LI (2012). **CD-HIT: accelerated for clustering the next-generation sequencing data.** *Bioinformatics* 28.23, pp. 3150–3152. DOI: 10.1093/bioinformatics/bts565 (cit. on p. 172).
- GABALDÓN, T (2005). **Evolution of proteins and proteomes: a phylogenetics approach.** *Evolutionary Bioinformatics Online* 1, pp. 51–61 (cit. on p. 201).
- GRISHKEVICH, V and I YANAI (2014). **Gene length and expression level shape genomic novelties.** *Genome Research* 24.9, pp. 1497–1503. DOI: 10.1101/gr.169722.113 (cit. on p. 200).
- HAHN, MW, MV HAN, and SG HAN (2007). **Gene Family Evolution across 12 *Drosophila* Genomes.** *PLoS Genet* 3.11, e197. DOI: 10.1371/journal.pgen.0030197 (cit. on p. 159).
- HARRIS, JK, ST KELLEY, GB SPIEGELMAN, and NR PACE (2003). **The Genetic Core of the Universal Ancestor.** *Genome Research* 13.3, pp. 407–412. DOI: 10.1101/gr.652803 (cit. on p. 159).
- HOFF, KJ, S LANGE, A LOMSADZE, M BORODOVSKY, and M STANKE (2016). **BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and**

- AUGUSTUS**. *Bioinformatics* 32.5, pp. 767–769. DOI: 10.1093/bioinformatics/btv661 (cit. on pp. 167–169).
- ITOH, M, JC NACHER, Ki KUMA, S GOTO, and M KANEHISA (2007). **Evolutionary history and functional implications of protein domains and their combinations in eukaryotes**. *Genome Biology* 8, R121. DOI: 10.1186/gb-2007-8-6-r121 (cit. on p. 201).
- JOHNSON, BR and ND TSUTSUI (2011). **Taxonomically restricted genes are associated with the evolution of sociality in the honey bee**. *BMC Genomics* 12, p. 164. DOI: 10.1186/1471-2164-12-164 (cit. on p. 159).
- JORDAN, IK, IB ROGOZIN, YI WOLF, and EV KOONIN (2002). **Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria**. *Genome Research* 12.6, pp. 962–968. DOI: 10.1101/gr.87702 (cit. on p. 200).
- KAJITANI, R, K TOSHIMOTO, H NOGUCHI, A TOYODA, Y OGURA, M OKUNO, M YABANA, M HARADA, E NAGAYASU, H MARUYAMA, Y KOHARA, A FUJIYAMA, T HAYASHI, and T ITOH (2014). **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads**. *Genome Research* 24.8, pp. 1384–1395. DOI: 10.1101/gr.170720.113 (cit. on p. 167).
- KEITH, N, AE TUCKER, CE JACKSON, W SUNG, JIL LLEDÓ, DR SCHRIDER, S SCHAACK, JL DUDYCHA, M ACKERMAN, AJ YOUNGE, JR SHAW, and M LYNCH (2015). **High mutational rates of large-scale duplication and deletion in *Daphnia pulex***. *Genome Research*. DOI: 10.1101/gr.191338.115 (cit. on p. 159).
- KHALTURIN, K, G HEMMRICH, S FRAUNE, R AUGUSTIN, and TC BOSCH (2009). **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends in Genetics* 25.9, pp. 404–413. DOI: 10.1016/j.tig.2009.07.006 (cit. on pp. 159, 198, 200).
- KIANIANMOMENI, A, CS ONG, G RÄTSCH, and A HALLMANN (2014). **Genome-wide analysis of alternative splicing in *Volvox carteri***. *BMC Genomics* 15, p. 1117. DOI: 10.1186/1471-2164-15-1117 (cit. on p. 200).
- KIM, D, B LANGMEAD, and SL SALZBERG (2015). **HISAT: a fast spliced aligner with low memory requirements**. *Nature Methods* 12.4, pp. 357–360. DOI: 10.1038/nmeth.3317 (cit. on p. 168).
- KOONIN, EV (2003). **Comparative genomics, minimal gene-sets and the last universal common ancestor**. *Nature Reviews Microbiology* 1.2, pp. 127–136. DOI: 10.1038/nrmicro751 (cit. on pp. 159, 160).
- (2011). *The Logic of Chance: The Nature and Origin of Biological Evolution*. Google-Books-ID: fvmv2kU6PrYC. FT Press. 530 pp. (cit. on pp. 157, 160, 173, 175, 178).
- KRYLOV, DM, YI WOLF, IB ROGOZIN, and EV KOONIN (2003). **Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution**. *Genome Research* 13.10, pp. 2229–2235. DOI: 10.1101/gr.1589103 (cit. on p. 160).
- KUMMERFELD, SK and SA TEICHMANN (2005). **Relative rates of gene fusion and fission in multi-domain proteins**. *Trends in Genetics* 21.1, pp. 25–30. DOI: 10.1016/j.tig.2004.11.007 (cit. on p. 201).

- LI, H, B HANDSAKER, A WYSOKER, T FENNELL, J RUAN, N HOMER, G MARTH, G ABECASIS, and R DURBIN (2009). **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352 (cit. on pp. 168, 169).
- LIPMAN, DJ, A SOUVOROV, EV KOONIN, AR PANCHENKO, and TA TATUSOVA (2002). **The relationship of protein conservation and sequence length.** *BMC Evolutionary Biology* 2, p. 20. DOI: 10.1186/1471-2148-2-20 (cit. on pp. 158, 160, 161, 199).
- MARÇAIS, G and C KINGSFORD (2011). **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 27.6, pp. 764–770. DOI: 10.1093/bioinformatics/btr011 (cit. on p. 165).
- MISOF, B *et al.* (2014). **Phylogenomics resolves the timing and pattern of insect evolution.** *Science* 346.6210, pp. 763–767. DOI: 10.1126/science.1257570 (cit. on pp. 163, 165, 172, 198, 201).
- MOORE, AD and E BORNBERG-BAUER (2012). **The Dynamics and Evolutionary Potential of Domain Loss and Emergence.** *Molecular Biology and Evolution* 29.2, pp. 787–796. DOI: 10.1093/molbev/msr250 (cit. on p. 161).
- MUSHEGIAN, AR and EV KOONIN (1996). **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proceedings of the National Academy of Sciences* 93.19, pp. 10268–10273 (cit. on p. 159).
- OUZOUNIS, CA, V KUNIN, N DARZENTAS, and L GOLDOVSKY (2006). **A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective.** *Research in Microbiology. Space Microbiology* 157.1, pp. 57–68. DOI: 10.1016/j.resmic.2005.06.015 (cit. on p. 159).
- PALMIERI, N, C KOSIOL, and C SCHLÖTTERER (2014). **The life cycle of Drosophila orphan genes.** *eLife* 3, e01311. DOI: 10.7554/eLife.01311 (cit. on p. 159).
- PARRA, G, K BRADNAM, Z NING, T KEANE, and I KORF (2009). **Assessing the gene space in draft genomes.** *Nucleic Acids Research* 37.1, pp. 289–297. DOI: 10.1093/nar/gkn916 (cit. on p. 159).
- PETERS, RS, L KROGMANN, C MAYER, A DONATH, S GUNKEL, K MEUSEMANN, A KOZLOV, L PODSIADLOWSKI, M PETERSEN, R LANFEAR, PA DIEZ, J HERATY, KM KJER, S KLOPFSTEIN, R MEIER, C POLIDORI, T SCHMITT, S LIU, X ZHOU, T WAPPLER, J RUST, B MISOF, and O NIEHUIS (2017). **Evolutionary History of the Hymenoptera.** *Current Biology* 27.7, pp. 1013–1018. DOI: 10.1016/j.cub.2017.01.027 (cit. on pp. 165, 172).
- PROVATARIS, P, K MEUSEMANN, O NIEHUIS, S GRATH, B MISOF, and G WAGNER (2018). **Signatures of DNA Methylation across Insects Suggest Reduced DNA Methylation Levels in Holometabola.** *Genome Biology and Evolution* 10.4, pp. 1185–1197. DOI: 10.1093/gbe/evy066 (cit. on p. 200).
- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing (cit. on p. 174).
- ROGNES, T (2011). **Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation.** *BMC Bioinformatics* 12, p. 221. DOI: 10.1186/1471-2105-12-221 (cit. on p. 172).



- ROUX, J and M ROBINSON-RECHAVI (2011). **Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication.** *Genome Research* 21.3, pp. 357–363. DOI: 10.1101/gr.113803.110 (cit. on p. 200).
- SIMÃO, FA, RM WATERHOUSE, P IOANNIDIS, EV KRIVENTSEVA, and EM ZDOBNOV (2015). **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 31.19, pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351 (cit. on pp. 159, 167, 170).
- SIMOLA, DF, L WISSLER, G DONAHUE, RM WATERHOUSE, M HELMKAMPF, J ROUX, S NYGAARD, KM GLASTAD, DE HAGEN, L VILJAKAINEN, JT REESE, BG HUNT, D GRAUR, E ELHAIK, EV KRIVENTSEVA, J WEN, BJ PARKER, E CASH, E PRIVMAN, CP CHILDERS, MC MUÑOZ-TORRES, JJ BOOMSMA, E BORNBERG-BAUER, CR CURRIE, CG ELSIK, G SUEN, MAD GOODISMAN, L KELLER, J LIEBIG, A RAWLS, D REINBERG, CD SMITH, CR SMITH, N TSUTSUI, Y WURM, EM ZDOBNOV, SL BERGER, and J GADAU (2013). **Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality.** *Genome Research* 23.8, pp. 1235–1247. DOI: 10.1101/gr.155408.113 (cit. on p. 159).
- SLATER, GS and E BIRNEY (2005). **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 6.1, p. 31. DOI: 10.1186/1471-2105-6-31 (cit. on p. 167).
- SMIT, A and R HUBLEY (2015). *RepeatModeler Open-1.0* (cit. on p. 167).
- SMIT, A, R HUBLEY, and P GREEN (2015). *RepeatMasker Open-4.0* (cit. on p. 167).
- STANKE, M, R STEINKAMP, S WAACK, and B MORGENSTERN (2004). **AUGUSTUS: a web server for gene finding in eukaryotes.** *Nucleic Acids Research* 32 (suppl 2), W309–W312. DOI: 10.1093/nar/gkh379 (cit. on pp. 169, 170).
- TATARINOVA, TV, I LYSNYANSKY, YV NIKOLSKY, and A BOLSHOY (2016). **The mysterious orphans of Mycoplasmataceae.** *Biology Direct* 11, p. 2. DOI: 10.1186/s13062-015-0104-3 (cit. on p. 199).
- TAUTZ, D and T DOMAZET-LOŠO (2011). **The evolutionary origin of orphan genes.** *Nature Reviews Genetics* 12.10, pp. 692–702. DOI: 10.1038/nrg3053 (cit. on p. 199).
- WANG, M and G CAETANO-ANOLLÉS (2009). **The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World.** *Structure* 17.1, pp. 66–78. DOI: 10.1016/j.str.2008.11.008 (cit. on p. 201).
- WANG, Z, D ZARLENGA, J MARTIN, S ABUBUCKER, and M MITREVA (2012). **Exploring metazoan evolution through dynamic and holistic changes in protein families and domains.** *BMC Evolutionary Biology* 12, p. 138. DOI: 10.1186/1471-2148-12-138 (cit. on p. 201).
- WASMUTH, J, R SCHMID, A HEDLEY, and M BLAXTER (2008). **On the Extent and Origins of Genic Novelty in the Phylum Nematoda.** *PLOS Neglected Tropical Diseases* 2.7, e258. DOI: 10.1371/journal.pntd.0000258 (cit. on pp. 160, 199, 200).

- WATERHOUSE, RM (2015). **A maturing understanding of the composition of the insect gene repertoire.** *Current Opinion in Insect Science* 7. DOI: 10.1016/j.cois.2015.01.004 (cit. on pp. 157, 159–161, 175, 178, 182, 198).
- WATERHOUSE, RM, EM ZDOBNOV, and EV KRIVENTSEVA (2011). **Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi.** *Genome Biology and Evolution* 3, pp. 75–86. DOI: 10.1093/gbe/evq083 (cit. on pp. 159–161, 173, 174).
- WICKHAM, H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York (cit. on p. 174).
- WILBRANDT, J, B MISOF, and O NIEHUIS (2017). **COGNATE: comparative gene annotation characterizer.** *BMC Genomics* 18.1, p. 535. DOI: 10.1186/s12864-017-3870-8 (cit. on p. 169).
- WISSLER, L, J GADAU, DF SIMOLA, M HELMKAMPF, and E BORNBERG-BAUER (2013). **Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes.** *Genome Biology and Evolution* 5.2, pp. 439–455. DOI: 10.1093/gbe/evt009 (cit. on p. 199).
- WOLF, YI, PS NOVICHKOV, GP KAREV, EV KOONIN, and DJ LIPMAN (2009). **The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages.** *Proceedings of the National Academy of Sciences* 106.18, pp. 7273–7280. DOI: 10.1073/pnas.0901808106 (cit. on p. 199).
- WYDER, S, E KRIVENTSEVA, R SCHRÖDER, T KADOWAKI, and E ZDOBNOV (2007). **Quantification of ortholog losses in insects and vertebrates.** *Genome Biology* 8.11, R242. DOI: 10.1186/gb-2007-8-11-r242 (cit. on p. 159).
- YANG, L, M ZOU, B FU, and S HE (2013). **Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish.** *BMC Genomics* 14, p. 65. DOI: 10.1186/1471-2164-14-65 (cit. on p. 199).
- YU GUANGCHUANG, SMITH DAVID K, ZHU HUACHEN, YI GUAN, and TOMMY TSAN-YUK LAM (2017). **ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data.** *Methods in Ecology and Evolution* 8.1, pp. 28–36. DOI: 10.1111/2041-210X.12628 (cit. on p. 174).
- ZDOBNOV, EM, F TEGENFELDT, D KUZNETSOV, RM WATERHOUSE, FA SIMÃO, P IOANNIDIS, M SEPPEY, A LOETSCHER, and EV KRIVENTSEVA (2017). **OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.** *Nucleic Acids Research* 45 (Database issue), pp. D744–D749. DOI: 10.1093/nar/gkw1119 (cit. on p. 170).
- ZENG, J and SV YI (2010). **DNA Methylation and Genome Evolution in Honeybee: Gene Length, Expression, Functional Enrichment Covary with the Evolutionary Signature of DNA Methylation.** *Genome Biology and Evolution* 2, pp. 770–780. DOI: 10.1093/gbe/evq060 (cit. on p. 200).
- ZHAO, C, LN ESCALANTE, H CHEN, TR BENATTI, J QU, S CHELLAPILLA, RM WATERHOUSE, D WHEELER, MN ANDERSSON, R BAO, M BATTERTON, SK BEHURA, KP BLANKENBURG, D CARAGEA, JC CAROLAN, M COYLE, M EL-BOUHSSINI, L

FRANCISCO, M FRIEDRICH, N GILL, T GRACE, CJP GRIMMELIKHUIJZEN, Y HAN, F HAUSER, N HERNDON, M HOLDER, P IOANNIDIS, L JACKSON, M JAVAID, SN JHANGIANI, AJ JOHNSON, D KALRA, V KORCHINA, CL KOVAR, F LARA, SL LEE, X LIU, C LÖFSTEDT, R MATA, T MATHEW, DM MUZNY, S NAGAR, LV NAZARETH, G OKWUONU, F ONGERI, L PERALES, BF PETERSON, LL PU, HM ROBERTSON, BJ SCHEMERHORN, SE SCHERER, JT SHREVE, D SIMMONS, S SUBRAMANYAM, RL THORNTON, K XUE, GM WEISSENBERGER, CE WILLIAMS, KC WORLEY, D ZHU, Y ZHU, MO HARRIS, RH SHUKLE, JH WERREN, EM ZDOBNOV, MS CHEN, SJ BROWN, JJ STUART, and S RICHARDS (2015). **A Massive Expansion of Effector Genes Underlies Gall-Formation in the Wheat Pest *Mayetiola destructor*.** *Current Biology* 25.5, pp. 613–620. DOI: 10.1016/j.cub.2014.12.057 (cit. on p. 159).



---

# **GENERAL DISCUSSION & CONCLUSION**

---



---

## General discussion

---

I SET OUT TO ADDRESS several general questions: are major evolutionary transitions in Hymenoptera shadowed by structural changes in their gene repertoires? What is the magnitude of structural variation in hymenopteran gene repertoires? And finally, are there classes of genes more variable and thus prone to change than others?

Here, I shortly review the results of my research and discuss how they contribute to our understanding of gene repertoire evolution. Following this, future perspectives are highlighted and a general conclusion is drawn.

### 1.1 Prerequisites

To approach the posed questions, it was necessary to unlock gene structures of complete repertoires in an efficient and traceable manner. I developed the tool COGNATE (see part II) to facilitate comparative analyses of gene repertoire structural parameters, and it has been stable and useful throughout my further work. When asked how to describe a gene repertoire structurally, I advocate the

use of several metrics (outlined in section II.4.1), which can be summarized as GC content, cumulative sizes, counts, and lengths of exons, coding sequences (if discriminated during annotation from exons), and introns. Furthermore, I would like to highlight the importance of taking the full parameter distribution into account in addition to summary statistics. Frequently, distributions are omitted and statistical tests are applied that are not necessarily suitable for data in non-normal distributions, as I found them to be the norm in gene structure parameters (except in GC contents).

Before the focal biological questions could be explored, another prerequisite had to be met, namely knowing whether the available data basis, *i.e.*, structural parameter records of gene repertoires from unsupervised automated annotation, was suitable (part III). I surveyed the impact of manual curation on automatically predicted gene models. The results are reassuring: overall, gene structure is only marginally changed and general trends of gene structure correlations to other genomic parameters were not influenced. Note, however, that the sample size of manually curated genes was small and the effect on protein sequence was not investigated. Further research is required to fully benchmark not only annotation tools but also manual curation to obtain a guideline of the cases that require manual curation. Based on these findings, I decided that the investigation of full protein sequences to characterize the functions comprised by the different gene repertoires would not be reliable enough without in-depth vetting of gene models to ensure model correctness.

## 1.2 Gene repertoires of Hymenoptera

Although COGNATE delivers measurements of a plethora of parameters, I decided to focus on the analyses on five key parameters to characterize gene structures (the full datasets are provided in the Appendix, B.2.4, C.2.7, D.1.2, E.2.4).

### 1.2.1 Gene structure variation along the phylogenetic tree

The first opportunity to compare gene structure parameters across whole protein-coding gene repertoires was created in part III. In this study, I



analyzed seven insect species including the two “Symphyta” that are a constant component of my species samples. I observed a pronounced difference in the parameter distributions of the two beetle species (more short proteins, more long introns) compared to the remaining species (Fig. III.3); the sawfly distributions resembled each other, but the two hemipteran species differed from each other considerably in intron length. I took these observations as foothold: there is variation between insect orders, but also within orders; a larger species sample is required to elucidate the magnitude of these inter- and intra-ordinal variations.

In part IV, gene structure variation over the whole protein-coding repertoires of twelve selected Hymenoptera were investigated in an evolutionary context. The most striking observed trend in concordance with phylogenetic relationships was a pronounced overall reduction of gene complexity in the gene repertoires of the three compared ants. Additionally, these three gene repertoires featured a distinctive, albeit small, class of genes with extremely short introns.

Although ants are a highly interesting model system to study the genomic basis of sociality (reviewed by GADAU *et al.*, 2012; LIBBRECHT *et al.*, 2013), the focus on gene structure is rare and not addressed in the genome publications (*e.g.*, CR SMITH *et al.*, 2015, 2011; CD SMITH *et al.*, 2011; SUEN *et al.*, 2011; WURM *et al.*, 2011). Given previous observations of correlations, the genes of the short-intron class might reside in a region of very high recombination (COMERON and KREITMAN, 2000; NIEHUIS *et al.*, 2010), or are highly expressed (CASTILLO-DAVIS *et al.*, 2002). It has also been argued that the ‘minimal introns’ found in humans are under selective pressure and enhance the RNA-export rate from the nucleus (YU *et al.*, 2002). It will be very interesting see whether the same class can be found in other ant genomes, to annotate these genes functionally and test whether these hypotheses hold. Potentially, the short-intron genes play an important role in ant evolution.

Comparing the covariance of genome and gene structure parameter summary statistics with assembly size between the sample of four taxonomic orders (Fig. III.3 b and c) and the Hymenoptera-centric sample (Fig. IV.2), it is striking that the relationship between assembly size and assembly GC content is reversed: within Hymenoptera, the correlation is slightly positive, while variation is larger than in the four-order sample (which comprises less species). This relationship can be found in previous studies (*e.g.*, STANDAGE *et al.*, 2016),

but has to my knowledge not been addressed in its own right specifically in insects. However, GC content has been shown to be related to recombination rates in Hymenoptera (NIEHUIS *et al.*, 2010). GC content has been largely addressed in the context of vertebrate isochore structure and evolution (*e.g.*, BERNARDI, 2007; COSTANTINI *et al.*, 2009; EYRE-WALKER and HURST, 2001) and shown to influence DNA bendability (ORTIZ and PABLO, 2011; VINOGRADOV, 2001). In plants, the biological significance of the relationship of GC content and genome size emerges from the higher thermal stability of GC-rich DNA (YAKOVCHUK *et al.*, 2006), which might confer an advantage during cell freezing and desiccation (ŠMARDÁ *et al.*, 2014). An additional selective advantage of GC-rich DNA in plants might be that it plays a role in facilitating more complex gene regulation (ŠMARDÁ *et al.*, 2014). The present data does not allow to test whether GC content variation in Hymenoptera correlates with habitat temperature and humidity or with gene regulation and other epigenomic aspects (*e.g.*, chromosome packaging, as suggested for vertebrates COSTANTINI *et al.* (2009)).

I conclude here that the study of genome composition and gene structure can provide highly exciting investigative leads for future research addressing more specific questions like: what is the biological role of the miniature introns in ants? Is a high GC content an ancestral state of Hymenoptera and did it play a role in diversification?

### 1.2.2 Gene structure variation within repertoire partitions

From my previous analyses, it became apparent that there might be partitions or modules of the protein-coding gene repertoire that evolve in a partition-specific way and which can be identified not only via functional assessment or gene expression analysis but also by their structural configuration. Thus, my subsequent project (described in part V) aimed to explore the variation of repertoire partitions based on conservation of ortholog groups.

I included an assessment of the gene structure parameter distributions of the whole repertoires (Fig. V.3) to compare it to previous results (Fig. IV.4). It became apparent that there exist differences between the two studies in the gene structure distributions across the whole gene repertoires of *O. abietinus* and *A. rosae*. These might result from the different annotation procedures. In part IV,

protein-coding genes were annotated by the i5k community with a customized MAKER-workflow, while the data analyzed in part V results from BRAKER runs. As demonstrated in part III, the tools produce different gene annotations. To date it is not possible to quickly and reliably assess which of the annotations (as a whole or which individual models) are a more correct representation of the biological reality. Although MAKER appears to produce more conservative models, it might miss lineage-specific genes, which can be expected to be included by BRAKER (see section III.4 for a detailed discussion).

Assessing my results, what can I say about the variability of gene repertoires? It became apparent that there is high variability of gene complexity (in terms of exon count) and compactness (lengths of exons and introns) within the repertoires of protein-coding genes of insects, co-varying with the conservation and duplicability of gene families (*e.g.*, Fig. V.7). Furthermore, apparently close-to-universal constraints (*e.g.*, in median intron length, Fig. V.11) can be found in the core class that are relaxed in less conserved gene families of the shell and cloud classes, even within those gene families that are universal single-copy orthologs of a clade. Potentially, the best evolutionary and mechanistic models explaining these findings are related to gene expression; the most prominent of these are 'selection for economy', 'genome complexity', and 'mutational bias', reviewed by WOODY and SHOEMAKER (2011). Showing the variability among repertoire partitions in full detail without a precedent assessment of and relation to gene expression levels (as, for example, in studies of plants, CAMIOLO *et al.*, 2009; PINGAULT *et al.*, 2015; and chicken, RAO *et al.*, 2010) allows to form testable hypothesis and identify gene classes of interest independently. For example, we can now ask whether the genes shared by all species, evolving under multicopy license, and comprising numerous domains (core-sMC-26 in Fig. V.13) also represent a special class with regard to sequence composition, gene expression patterns, and functional roles. Thereby, another route to test the expression-related models of evolution opens.

My work is a contribution towards the unification of our knowledge on gene structure, gene expression, and essentiality with repertoire evolution. It includes the notion of single-copy control and multicopy license, similar to the example of studying single-copy control acting on housekeeping genes in plants (DE SMET *et al.*, 2013) and in human (ACHARYA *et al.*, 2015). This inclusion adds a relevant aspect to the view on gene repertoires. With this, I lay here the foundations for future work.

## 1.3 Future prospects

### 1.3.1 Methodological aspects

The presented work relied on a widely accepted method of orthology inference (ZDOBNOV *et al.*, 2017; the five publications describing the OrthoDB database and its developments of the last ten years have been cited at least 338 times<sup>VI.1</sup>; still, alternatives are being developed and potentially reward testing to assess the extent of artifact observations). However, it has been argued that the nature of protein architecture, being build from domains that may be added to ‘unrelated’ genes (by *e.g.*, processes of gene fission and fusion), might lead to actually reticulate orthology-relationships among genes (GABALDÓN and KOONIN, 2013; SONNHAMMER *et al.*, 2014). It will be interesting to investigate gene repertoire dynamics within a domain-aware framework of orthologous relationships.

### 1.3.2 Repertoire dynamics: the role of gene turnover

When thinking about the variability of protein-coding gene repertoires, it is easy to arrive at questions regarding the dynamics, *i.e.*, how changes over time occur and manifest. Along the same line, further questions are whether anything can be inferred about evolutionary dynamics of a repertoire from structural characteristics. In other words, what can we learn about genome evolution and gene repertoire evolution from analyzing gene structure and family content? Is there a connection between a gene’s structure, its conservation (similarity across species), and its evolutionary path? Does gene structure, function, or domain content influence its evolvability beyond selective effects? Do gene turnover rates change, why, and how? Are rate changes specific to organismic kingdoms or other clades?

Approaches to these questions have to use robust estimates of gene turnover rates on the lineages leading to the studied gene repertoires/genomes.

---

<sup>VI.1</sup> OrthoDB citation counts obtained from PubMed.  
<https://www.ncbi.nlm.nih.gov/pubmed>. Last accessed 30 March 2018.

These can only be obtained with some reliability when a sufficiently dense and representative sample of species is used — where sufficiency and representativeness have to be subject to further discussion and research. In addition to an adequate genome sample, a dated phylogeny (*i.e.*, known divergence times and phylogenetic relationships) as well as genome assemblies and annotations obtained from comparable and reliable methods, are required. The ever-filling and expanding treasure trove of sequenced genomes only now allows to tackle issues related to gene turnover and its rate(s).

It is possible to hypothesize on the drivers of gene turnover. The mechanisms that most likely influence changes in repertoire composition are large segmental or whole-genome duplications as well as direct and indirect effects of transposable element activity.

It has been demonstrated that whole genome duplication has been a factor in the evolution of eukaryotes (reviewed by LI *et al.*, 2018). The mechanisms and consequences of whole genome duplication have been addressed (*e.g.*, FORCE *et al.*, 1999; WATERHOUSE *et al.*, 2011), but not fully elucidated, for example it is open whether whole genome duplication leads to increased substitution rates within the whole genome (reviewed by RAES and VAN DE PEER, 2003; TAYLOR and RAES, 2004). The large variation in genome size in insects could be a result of ancient whole genome duplication (GREGORY and JOHNSTON, 2008; GREGORY and HEBERT, 2003; LI *et al.*, 2018). Segmental duplications have been identified in the genomes of ants (BONASIO *et al.*, 2010) and can be the source of arthropod taxon-restricted genes (WISSLER *et al.*, 2013). The impact of (ancient) whole genome duplication on gene turnover rates as a driving force remains uncertain, though, and would warrant further research.

In contrast to the vestigial evidence of whole genome duplication driving gene turnover, the body of evidence documenting that transposable elements can influence turnover directly (by transposition, ALBERTIN *et al.*, 2015; MAUMUS *et al.*, 2015) and indirectly (by constituting potential sites of ectopic recombination, ALBERTIN *et al.*, 2015; GRAY, 2000; LIM *et al.*, 2008) is growing (KAZAZIAN, 2004). In addition, transposable elements can be the source of *de novo*-recruited genes originally encoded by transposable elements that become integrated in the organism's genome and gain a 'relevant' function (called 'molecular domestication'), as has been shown to occur in insects (BIEDLER and TU, 2003; CASOLA *et al.*, 2007; KAPITONOV and JURKA, 2004). In ants, one eight (12.4 %) of genes are of this type.

of species-specific genes might have been domesticated from transposable elements (WISSLER *et al.*, 2013). It will be worthwhile to explore in how far abundance, diversity, and historical activity of transposable elements might drive gene turnover rates and thereby foster the evolution of diversity.

### 1.3.3 A window to the past leading to the future

If we can estimate historical gene turnover rates that governed gene repertoire composition during the evolution of Holometabola and insects, it will also be possible to reconstruct the gene repertoire of the ancestral species. One specific goal comes within reach: the characterization of HANS, the Holometabolan ANcestral Species<sup>VI.2</sup>.

The fully characterized HANS will allow to trace the genomic autapomorphies of Holometabola and serve as a guideline to explore and understand the evolution of insect biodiversity.

---

<sup>VI.2</sup> I invented the term HANS following the scheme used for LECA (last eukaryotic ancestor, MARGULIS *et al.*, 2006), and LUCA (last universal common ancestor, GLANSDORFF *et al.*, 2008).

---

## General conclusion

---

**T**HE HERE PRESENTED gene repertoire analyses are outstanding as they are based on three previously unavailable prerequisites: (1) a new tool (COGNATE) was developed and used that records all parameters instead of relying on summary metrics; (2) it was ensured that automatically generated gene models are suitable for gene structure analyses; and (3) a unique species sample was employed (in part V), which covers representatively the younger radiation of Chrysididae and simultaneously allows the comparison of holometabolous and hemimetabolous insects to a millipede outgroup.

Equipped with the means to describe a gene repertoire on the structural level, I explored whether the vast majority of currently available data — annotations that have been generated automatically and not been curated by human experts — are adequate to study gene repertoire dynamics with respect to structure. The comparison of around 1,000 manually curated genes with their overlapping uncurated counterparts in each of seven insect species showed that in the context of gene repertoire-wide structural assessments, automatically generated gene models are sufficiently reliable. COGNATE was used in two further pursuing analyses. The first was a characterization of the protein-coding

gene repertoires of twelve hymenopteran species, which bespoke ant-specific miniature introns. My study of structural characteristics of the gene repertoires partitioned by conservation and duplicability revealed an unexpected variation between and among conservation classes.

My work provides a solid baseline of expectations on insect gene structure variation as well as manifold investigative leads prompting further exciting studies.

This is only the beginning.



---

## Bibliography VI

---

- ACHARYA, D, D MUKHERJEE, S PODDER, and TC GHOSH (2015). **Investigating Different Duplication Pattern of Essential Genes in Mouse and Human.** *PLoS ONE* 10.3. DOI: 10.1371/journal.pone.0120784 (cit. on p. 219).
- ALBERTIN, CB, O SIMAKOV, T MITROS, ZY WANG, JR PUNGOR, E EDSINGER-GONZALES, S BRENNER, CW RAGSDALE, and DS ROKHSAR (2015). **The octopus genome and the evolution of cephalopod neural and morphological novelties.** *Nature* 524.7564, pp. 220–224. DOI: 10.1038/nature14668 (cit. on p. 221).
- BERNARDI, G (2007). **The neoselectionist theory of genome evolution.** *Proceedings of the National Academy of Sciences* 104.20, pp. 8385–8390. DOI: 10.1073/pnas.0701652104 (cit. on p. 218).
- BIEDLER, J and Z TU (2003). **Non-LTR Retrotransposons in the African Malaria Mosquito, *Anopheles gambiae*: Unprecedented Diversity and Evidence of Recent Activity.** *Molecular Biology and Evolution* 20.11, pp. 1811–1825. DOI: 10.1093/molbev/msg189 (cit. on p. 221).
- BONASIO, R, G ZHANG, C YE, NS MUTTI, X FANG, N QIN, G DONAHUE, P YANG, Q LI, C LI, P ZHANG, Z HUANG, SL BERGER, D REINBERG, J WANG, and J LIEBIG (2010). **Genomic Comparison of the Ants *Camponotus floridanus* and *Harpegnathos saltator*.** *Science* 329.5995, pp. 1068–1071. DOI: 10.1126/science.1192428 (cit. on p. 221).
- CAMIOLO, S, D RAU, and A PORCEDDU (2009). **Mutational Biases and Selective Forces Shaping the Structure of Arabidopsis Genes.** *PLoS ONE* 4.7. DOI: 10.1371/journal.pone.0006356 (cit. on p. 219).
- CASOLA, C, AM LAWING, E BETRÁN, and C FESCHOTTE (2007). **PIF-like Transposons are Common in *Drosophila* and Have Been Repeatedly Domesticated to Generate New Host Genes.** *Molecular Biology and Evolution* 24.8, pp. 1872–1888. DOI: 10.1093/molbev/msm116 (cit. on p. 221).

- CASTILLO-DAVIS, CI, SL MEKHEDOV, DL HARTL, EV KOONIN, and FA KONDRASHOV (2002). **Selection for short introns in highly expressed genes.** *Nature Genetics* 31.4, pp. 415–418. DOI: 10.1038/ng940 (cit. on p. 217).
- COMERON, JM and M KREITMAN (2000). **The Correlation Between Intron Length and Recombination in Drosophila: Dynamic Equilibrium Between Mutational and Selective Forces.** *Genetics* 156.3, pp. 1175–1190 (cit. on p. 217).
- COSTANTINI, M, R CAMMARANO, and G BERNARDI (2009). **The evolution of isochores patterns in vertebrate genomes.** *BMC Genomics* 10, p. 146. DOI: 10.1186/1471-2164-10-146 (cit. on p. 218).
- DE SMET, R, KL ADAMS, K VANDEPOELE, MCE VAN MONTAGU, S MAERE, and Y VAN DE PEER (2013). **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.** *Proceedings of the National Academy of Sciences of the United States of America* 110.8, pp. 2898–2903. DOI: 10.1073/pnas.1300127110 (cit. on p. 219).
- EYRE-WALKER, A and LD HURST (2001). **The evolution of isochores.** *Nature Reviews Genetics* 2.7, pp. 549–555. DOI: 10.1038/35080577 (cit. on p. 218).
- FORCE, A, M LYNCH, FB PICKETT, A AMORES, YI YAN, and J POSTLETHWAIT (1999). **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 151.4, pp. 1531–1545 (cit. on p. 221).
- GABALDÓN, T and EV KOONIN (2013). **Functional and evolutionary implications of gene orthology.** *Nature Reviews Genetics* 14.5, pp. 360–366. DOI: 10.1038/nrg3456 (cit. on p. 220).
- GADAU, J, M HELMKAMPF, S NYGAARD, J ROUX, DF SIMOLA, CR SMITH, G SUEN, Y WURM, and CD SMITH (2012). **The genomic impact of 100 million years of social evolution in seven ant species.** *Trends in Genetics* 28.1, pp. 14–21. DOI: 10.1016/j.tig.2011.08.005 (cit. on p. 217).
- GLANSDORFF, N, Y XU, and B LABEDAN (2008). **The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner.** *Biology Direct* 3.1, p. 29. DOI: 10.1186/1745-6150-3-29 (cit. on p. 222).
- GRAY, YHM (2000). **It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements.** *Trends in Genetics* 16.10, pp. 461–468. DOI: 10.1016/S0168-9525(00)02104-1 (cit. on p. 221).
- GREGORY, TR and JS JOHNSTON (2008). **Genome size diversity in the family Drosophilidae.** *Heredity* 101.3, pp. 228–238. DOI: 10.1038/hdy.2008.49 (cit. on p. 221).
- GREGORY, TR and PD HEBERT (2003). **Genome size variation in lepidopteran insects.** *Canadian Journal of Zoology* 81.8, pp. 1399–1405. DOI: 10.1139/z03-126 (cit. on p. 221).
- KAPITONOV, VV and J JURKA (2004). **Harbinger Transposons and an Ancient HARBI1 Gene Derived from a Transposase.** *DNA and Cell Biology* 23.5, pp. 311–324. DOI: 10.1089/104454904323090949 (cit. on p. 221).
- KAZAZIAN, HHJ (2004). **Mobile Elements: Drivers of Genome Evolution.** *Science* 303.5664. ArticleType: research-article / Full publication date: Mar. 12, 2004

- / Copyright © 2004 American Association for the Advancement of Science, pp. 1626–1632. DOI: 10.2307/3836446 (cit. on p. 221).
- LI, Z, G TILEY, S GALUSKA, C REARDON, T KIDDER, R RUNDELL, and MS BARKER (2018). **Multiple large-scale gene and genome duplications during the evolution of hexapods.** *bioRxiv*, p. 253609. DOI: 10.1101/253609 (cit. on p. 221).
- LIBBRECHT, R, PR OXLEY, DJ KRONAUER, and L KELLER (2013). **Ant genomics sheds light on the molecular regulation of social organization.** *Genome Biol* 14, p. 212 (cit. on p. 217).
- LIM, KY, DE SOLTIS, PS SOLTIS, J TATE, R MATYASEK, H SRUBAROVA, A KOVARIK, JC PIRES, Z XIONG, and AR LEITCH (2008). **Rapid Chromosome Evolution in Recently Formed Polyploids in Tragopogon (Asteraceae).** *PLoS ONE* 3.10, e3353. DOI: 10.1371/journal.pone.0003353 (cit. on p. 221).
- MARGULIS, L, M CHAPMAN, R GUERRERO, and J HALL (2006). **The last eukaryotic common ancestor (LECA): Acquisition of cytoskeletal motility from aerotolerant spirochetes in the Proterozoic Eon.** *Proceedings of the National Academy of Sciences* 103.35, pp. 13080–13085. DOI: 10.1073/pnas.0604985103 (cit. on p. 222).
- MAUMUS, F, AS FISTON-LAVIER, and H QUESNEVILLE (2015). **Impact of transposable elements on insect genomes and biology.** *Current Opinion in Insect Science*. DOI: 10.1016/j.cois.2015.01.001 (cit. on p. 221).
- NIEHUIS, O, JD GIBSON, MS ROSENBERG, BA PANNEBAKKER, T KOEVOETS, AK JUDSON, CA DESJARDINS, K KENNEDY, D DUGGAN, LW BEUKEBOOM, Lvd ZANDE, DM SHUKER, JH WERREN, and J GADAU (2010). **Recombination and Its Impact on the Genome of the Haplodiploid Parasitoid Wasp Nasonia.** *PLOS ONE* 5.1, e8597. DOI: 10.1371/journal.pone.0008597 (cit. on pp. 217, 218).
- ORTIZ, V and JJ de PABLO (2011). **Molecular Origins of DNA Flexibility: Sequence Effects on Conformational and Mechanical Properties.** *Physical Review Letters* 106.23, p. 238107. DOI: 10.1103/PhysRevLett.106.238107 (cit. on p. 218).
- PINGAULT, L, F CHOULET, A ALBERTI, N GLOVER, P WINCKER, C FEUILLET, and E PAUX (2015). **Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome.** *Genome Biology* 16.1. DOI: 10.1186/s13059-015-0601-9 (cit. on p. 219).
- RAES, J and Y VAN DE PEER (2003). **Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico.** *Applied bioinformatics* 2.2, pp. 91–101 (cit. on p. 221).
- RAO, YS, ZF WANG, XW CHAI, GZ WU, M ZHOU, QH NIE, and XQ ZHANG (2010). **Selection for the compactness of highly expressed genes in Gallus gallus.** *Biology Direct* 5, p. 35. DOI: 10.1186/1745-6150-5-35 (cit. on p. 219).
- ŠMARDA, P, P BUREŠ, L HOROVÁ, IJ LEITCH, L MUCINA, E PACINI, L TICHÝ, V GRULICH, and O ROTREKLOVÁ (2014). **Ecological and evolutionary significance of genomic GC content diversity in monocots.** *Proceedings of the National Academy of Sciences* 111.39, E4096–E4102. DOI: 10.1073/pnas.1321152111 (cit. on p. 218).
- SMITH, CR, SH CAHAN, C KEMENA, SG BRADY, W YANG, E BORNBERG-BAUER, T ERIKSSON, J GADAU, M HELMKAMPF, D GOTZEK, MO MIYAKAWA, AV SUAREZ, and

- A MIKHEYEV (2015). **How Do Genomes Create Novel Phenotypes? Insights from the Loss of the Worker Caste in Ant Social Parasites.** *Molecular Biology and Evolution* 32.11, pp. 2919–2931. DOI: 10.1093/molbev/msv165 (cit. on p. 217).
- SMITH, CR, CD SMITH, HM ROBERTSON, M HELMKAMPF, A ZIMIN, M YANDELL, C HOLT, H HU, E ABOUHEIF, R BENTON, E CASH, V CROSET, CR CURRIE, E ELHAIK, CG ELSIK, MJ FAVÉ, V FERNANDES, JD GIBSON, D GRAUR, W GRONENBERG, KJ GRUBBS, DE HAGEN, ASI VINIEGRA, BR JOHNSON, RM JOHNSON, A KHILA, JW KIM, KA MATHIS, MC MUNOZ-TORRES, MC MURPHY, JA MUSTARD, R NAKAMURA, O NIEHUIS, S NIGAM, RP OVERSON, JE PLACEK, R RAJAKUMAR, JT REESE, G SUEN, S TAO, CW TORRES, ND TSUTSUI, L VILJAKAINEN, F WOLSCHIN, and J GADAU (2011). **Draft genome of the red harvester ant *Pogonomyrmex barbatus*.** *Proceedings of the National Academy of Sciences* 108.14, pp. 5667–5672. DOI: 10.1073/pnas.1007901108 (cit. on p. 217).
- SMITH, CD, A ZIMIN, C HOLT, E ABOUHEIF, R BENTON, E CASH, V CROSET, CR CURRIE, E ELHAIK, CG ELSIK, MJ FAVE, V FERNANDES, J GADAU, JD GIBSON, D GRAUR, KJ GRUBBS, DE HAGEN, M HELMKAMPF, JA HOLLEY, H HU, ASI VINIEGRA, BR JOHNSON, RM JOHNSON, A KHILA, JW KIM, J LAIRD, KA MATHIS, JA MOELLER, MC MUÑOZ-TORRES, MC MURPHY, R NAKAMURA, S NIGAM, RP OVERSON, JE PLACEK, R RAJAKUMAR, JT REESE, HM ROBERTSON, CR SMITH, AV SUAREZ, G SUEN, EL SUHR, S TAO, CW TORRES, Ev WILGENBURG, L VILJAKAINEN, KKO WALDEN, AL WILD, M YANDELL, JA YORKE, and ND TSUTSUI (2011). **Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*).** *Proceedings of the National Academy of Sciences* 108.14, pp. 5673–5678. DOI: 10.1073/pnas.1008617108 (cit. on p. 217).
- SONNHAMMER, ELL, T GABALDÓN, AWSd SILVA, M MARTIN, M ROBINSON-RECHAVI, B BOECKMANN, PD THOMAS, and C DESSIMOZ (2014). **Big data and other challenges in the quest for orthologs.** *Bioinformatics* 30.21, pp. 2993–2998. DOI: 10.1093/bioinformatics/btu492 (cit. on p. 220).
- STANDAGE, DS, AJ BERENS, KM GLASTAD, AJ SEVERIN, VP BRENDEL, and AL TOTH (2016). **Genome, transcriptome, and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect.** *Molecular Ecology*, n/a–n/a. DOI: 10.1111/mec.13578 (cit. on p. 217).
- SUEN, G, C TEILING, L LI, C HOLT, E ABOUHEIF, E BORNBERG-BAUER, P BOUFFARD, EJ CALDERA, E CASH, A CAVANAUGH, O DENAS, E ELHAIK, MJ FAVÉ, J GADAU, JD GIBSON, D GRAUR, KJ GRUBBS, DE HAGEN, TT HARKINS, M HELMKAMPF, H HU, BR JOHNSON, J KIM, SE MARSH, JA MOELLER, MC MUÑOZ-TORRES, MC MURPHY, MC NAUGHTON, S NIGAM, R OVERSON, R RAJAKUMAR, JT REESE, JJ SCOTT, CR SMITH, S TAO, ND TSUTSUI, L VILJAKAINEN, L WISSELER, MD YANDELL, F ZIMMER, J TAYLOR, SC SLATER, SW CLIFTON, WC WARREN, CG ELSIK, CD SMITH, GM WEINSTOCK, NM GERARDO, and CR CURRIE (2011). **The Genome Sequence of the Leaf-Cutter Ant *Atta cephalotes* Reveals Insights into Its Obligate Symbiotic Lifestyle.** *PLoS Genet* 7.2, e1002007. DOI: 10.1371/journal.pgen.1002007 (cit. on p. 217).

- TAYLOR, JS and J RAES (2004). **DUPLICATION AND DIVERGENCE: The Evolution of New Genes and Old Ideas.** *Annual Review of Genetics* 38.1, pp. 615–643. DOI: 10.1146/annurev.genet.38.072902.092831 (cit. on p. 221).
- VINOGRADOV, AE (2001). **Bendable Genes of Warm-blooded Vertebrates.** *Molecular Biology and Evolution* 18.12, pp. 2195–2200. DOI: 10.1093/oxfordjournals.molbev.a003766 (cit. on p. 218).
- WATERHOUSE, RM, EM ZDOBNOV, and EV KRIVENTSEVA (2011). **Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi.** *Genome Biology and Evolution* 3, pp. 75–86. DOI: 10.1093/gbe/evq083 (cit. on p. 221).
- WISSLER, L, J GADAU, DF SIMOLA, M HELMKAMPF, and E BORNBERG-BAUER (2013). **Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes.** *Genome Biology and Evolution* 5.2, pp. 439–455. DOI: 10.1093/gbe/evt009 (cit. on pp. 221, 222).
- WOODY, JL and RC SHOEMAKER (2011). **Gene Expression: Sizing It All Up.** *Frontiers in Genetics* 2. DOI: 10.3389/fgene.2011.00070 (cit. on p. 219).
- WURM, Y, J WANG, O RIBA-GROGNOUZ, M CORONA, S NYGAARD, BG HUNT, KK INGRAM, L FALQUET, M NIPITWATTANAPHON, D GOTZEK, MB DIJKSTRA, J OETTLER, F COMTESSE, CJ SHIH, WJ WU, CC YANG, J THOMAS, E BEAUDOING, S PRADERVAND, V FLEGEL, ED COOK, R FABBRETTI, H STOCKINGER, L LONG, WG FARMERIE, J OAKLEY, JJ BOOMSMA, P PAMILO, SV YI, J HEINZE, MAD GOODISMAN, L FARINELLI, K HARSHMAN, N HULO, L CERUTTI, I XENARIOS, D SHOEMAKER, and L KELLER (2011). **The genome of the fire ant *Solenopsis invicta*.** *Proceedings of the National Academy of Sciences* 108.14, pp. 5679–5684. DOI: 10.1073/pnas.1009690108 (cit. on p. 217).
- YAKOVCHUK, P, E PROTOZANOVA, and MD FRANK-KAMENETSKII (2006). **Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.** *Nucleic Acids Research* 34.2, pp. 564–574. DOI: 10.1093/nar/gkj454 (cit. on p. 218).
- YU, J, Z YANG, M KIBUKAWA, M PADDOCK, DA PASSEY, and GKS WONG (2002). **Minimal Introns Are Not “Junk”.** *Genome Research* 12.8, pp. 1185–1189. DOI: 10.1101/gr.224602 (cit. on p. 217).
- ZDOBNOV, EM, F TEGENFELDT, D KUZNETSOV, RM WATERHOUSE, FA SIMÃO, P IOANNIDIS, M SEPPEY, A LOETSCHER, and EV KRIVENTSEVA (2017). **OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.** *Nucleic Acids Research* 45 (Database issue), pp. D744–D749. DOI: 10.1093/nar/gkw1119 (cit. on p. 220).



---

# **APPENDICES**

---





---

## Posters

---

Apart from the poster shown in the Introduction (Fig. I.1), I prepared two more posters to present current states of my work to a larger audience. One of them (Fig. VII.1) was displayed in September 2015 at the Third Leibniz PhD Symposium held at the Leibniz Institute for Zoo- and Wildlife Research in Berlin and provides an overview of early results and plans. The second poster (Fig. VII.2) was put up at the 110<sup>th</sup> Annual Conference of the German Research Society in Bielefeld (2017) as a teaser for the publication “How suitable are automatically inferred gene models for uncovering taxon-specific gene structural differences in gene repertoires?” (included here as part III).

Querschnittsthema – Overarching Theme  
Comparative Genomics, Bioinformatics



3rd Leibniz PhD Symposium — Leibniz Institute for Zoo and Wildlife Research, Berlin — September 24th and 25th, 2015

## Genome Annotation Comparison Towards a better understanding of gene repertoire alterations

Jeanne Wilbrandt<sup>1</sup>, Malte Petersen<sup>1</sup>, Christoph Mayer<sup>1</sup>, Bernhard Misof<sup>1</sup>, Oliver Niehuis<sup>1</sup>

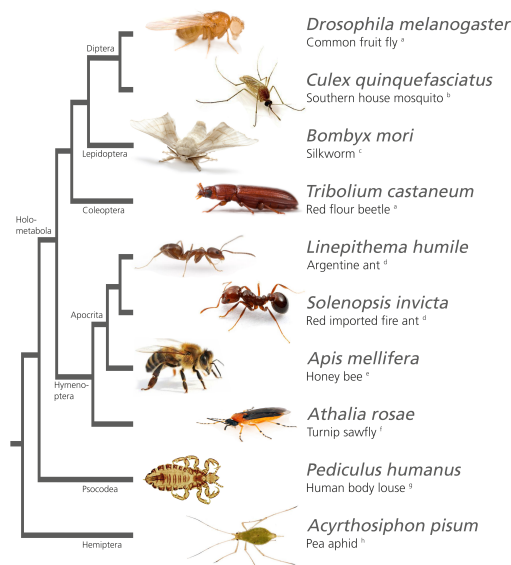


Surveying the characteristics of insect protein-coding genes is the first step to assess the differences between insect genomes and ultimately to explain biodiversity.

Insects, highly diverse given their phylogenetic age, are suspected to have genomes that differ in their organization and evolution from those of vertebrates and plants [1]. We aim to characterize their genomes and elucidate evolutionary processes potentially responsible for their success in terms of biodiversity. Basing our study on a comparative analysis of gene structure and genome composition parameters, we will comparatively characterize insect gene repertoire content. Here we present the first steps.

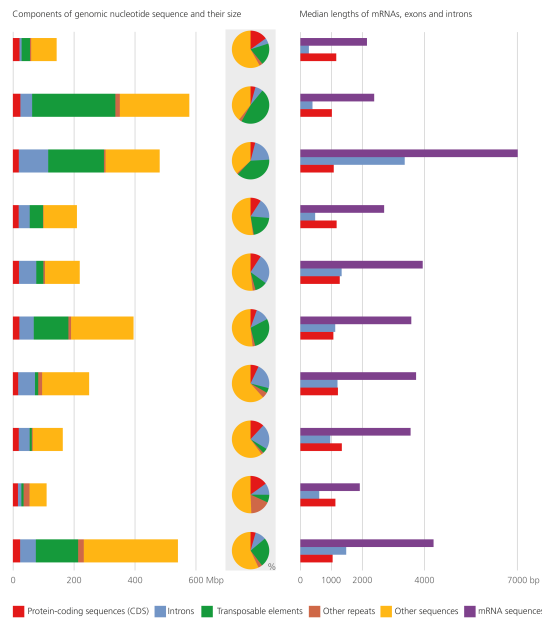
How do genomes evolve? Despite enormous genome size differences, we find a constant number of genes encoded therein. However, the amount of genes shared between genomes, the specific gene repertoire, differs [2]. Questions arise: which forces drive the evolution of genetic and genomic composition? How do gene repertoires change? Can we quantify the contribution of gene repertoire changes to genome and eventually organismic evolution? How is this related to the effect of changes in other genome components?

### Considering evolutionary relationships



Cladogram based on [3]. Species data are published in and taken from NCBI genome database. Image sources see below.

### Measuring genome and gene parameters



### Reviewing the methods

#### Structural annotation

Gene models were predicted by NCBI using their tool Gnomon. Other components (transposable elements, etc.) were annotated with RepeatMasker. Custom scripts were used for evaluation.

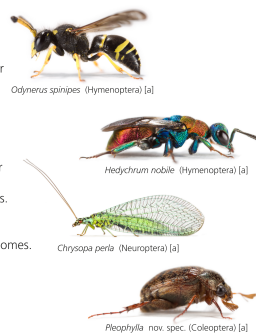
#### Sources of bias

All annotation tools require parameters and definitions to discern structure from random sequence. Automated approaches rely either on sequence similarity or model parameters to/of known genes. This means, finding truly new structures requires different approaches.

Previously, comparative genomics was limited by the few available genomes. There is still a bias towards smaller, easier to sequence genomes.

#### Extending the sample

Within GBR, we currently sequence 37 insect and outgroup species to foster phylogenetically informed comparisons. For example:



### Completing the picture

#### Families and orphans — repertoire content

The next steps require the grouping of genes into families. The comparison between species allows to identify genes related by speciation (orthologs, family seeds) and duplication (paralogs, family members).

The expansion and contraction of families — gene turnover — is what shapes gene repertoire content over time. We aim to identify and quantify the responsible processes and mechanisms.

#### Beyond protein-coding genes

We observe that genome size and the proportion of repetitive elements vary between species. Potential drivers of gene turnover are whole genome duplication (providing 'raw material' and possibilities of differential gene loss) and effects of transposable element activity (allowing duplication via copy and paste or disrupting genes by fostering non-homologous recombination). Analyzing these driving forces is our ongoing research.



<sup>1</sup> Leibniz Graduate School on Genomic Biodiversity Research<sup>1</sup>  
<sup>2</sup> Zentrum für molekulare Biodiversitätsforschung (zmb)  
Zoologisches Forschungsmuseum Alexander Koenig  
— Leibniz Institute of Animal Biodiversity —  
Adenauerallee 160, 53113 Bonn, Germany

Image Sources: [a] O. Niehuis, R.S. Peters,  
[b] www.images.onset.freedom.com, [c] www.  
seidenstrassen.biz, [d] A. Wild, [e] www.frontline.de,  
[f] M. Hatakeyama, [g] Insect Wiki, [h] B. Chabbert

Acknowledgements go to all GBR members.

References: [1] Hahn et al., *PLoS Genet* 3.11 (2007): e197.  
[2] Waterhouse, *COIS* 7 (2015): 15-23. [3] Misof et al.,  
*Science* 346.6210 (2014): 763-767.



Figure VII.1 – Towards a better understanding of gene repertoire alterations.  
(Continued on next page)

**Figure VII.1 – Towards a better understanding of gene repertoire alterations.**

(Continued)

This poster was presented at the Third Leibniz PhD Symposium held at the Leibniz Institute for Zoo- and Wildlife Research in Berlin.

The shown data and plans were not all used/realized, but the rationale described was followed throughout this thesis. Thus, this piece should be perceived as a fossil of my working progress.

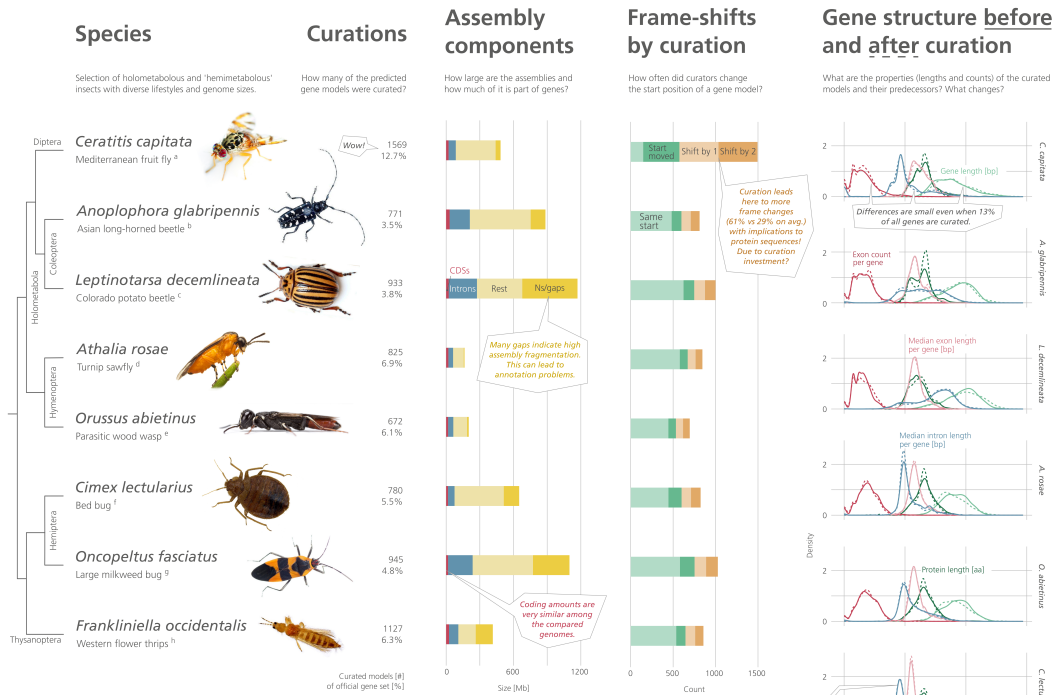
Wilbrandt et al. in prep. — September 2017

# Comparing automatically generated and manually curated annotations Data basis, tool choice, human review: influences on predicted protein-coding gene structure

Jeanne Wilbrandt<sup>1</sup>, Bernhard Misof<sup>1</sup>, Kristen Panfilio<sup>2,3</sup>, Oliver Niehuis<sup>4</sup>

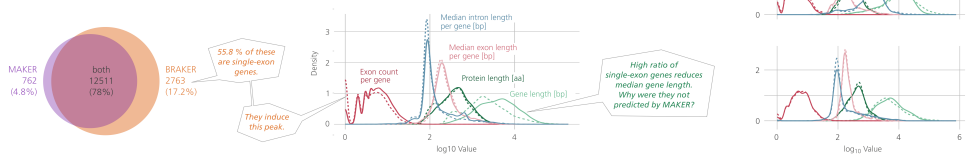


**Motivation:** where are the limits of automated structural gene annotation and of interpreting gene structure data across genomes? When is manual curation crucial?



## Gene sets predicted by MAKER and BRAKER

How large is the influence of the chosen tool for automated gene annotation on the generated data? In the example *A. rosae*, how many protein-coding gene models are predicted by MAKER (5k default) and BRAKER (low re-annotation)? How many overlap? Do they differ in their structure?



**Conclusion:** automated approaches appear to be biased in the prediction of single-exon genes. Are we missing a whole class of genes with MAKER? Also, correct reading frame prediction appears to be problematic. Curation is thus recommended for reliable protein structure analyses.



<sup>1</sup> Wilbrandt@leibniz-zfmk.de; <sup>2</sup> Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany; <sup>3</sup> School of Life Sciences, Gibbet Hill Campus, University of Warwick, Coventry, CV4 7AL, UK; <sup>4</sup> Institut für Zoologie, Albert-Ludwigs-Universität zu Köln, Zùlpicher Str. 47b, 50674 Köln, Germany; <sup>5</sup> Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert-Ludwigs-Universität Freiburg, Hertzstr. 1, 79104 Freiburg, Image and data sources: All data is part of the iG project (10.1093/ihered/056). [a] sborn.com.au; 10.1186/13059-016-1049-2; [b] entolip-uk.stn.edu; 10.1186/13059-016-1088-8; [c] pinterest.com; used with permission; [d] mattotaphology.co.uk; ZFMK; [e] O. Niehuis, ZFMK; [f] commons.wikimedia.org; 10.1038/ncomms10165; [g] flimimgflood.com; used with permission; [h] noselfies-agriculture.com; used with permission. Methods: iG: NAJ; annotation: used modified MAKER (10.1186/1471-2105-12-491); reannotation done with BRAKER (10.1093/bioinformatics/btt661); retrieved structural properties with COGNATE (10.1186/12864-017-3870-8). Acknowledgements go to all GBR and iG members.



Figure VII.2 – Data basis, tool choice, and human review: influences on predicted gene structures. (Continued on next page)

**Figure VII.2 – Data basis, tool choice, and human review: influences on predicted gene structures.**

(Continued)

This poster was presented at the 110<sup>th</sup> Annual Conference of the German Research Society in Bielefeld.

The shown data do not fully match with those used in this thesis: the species *Ceratitis capitata* was removed from the sample due to technical incompatibilities during analyses; and the analysis of frame-shifts by curation was not included in the final manuscript because it was only based on gene starts where an analysis of exon boundaries would have been more conclusive but were not feasible within reasonable time.

This piece should be perceived as a fossil of my working progress, not presenting the final results and conclusions as described in this thesis (part III).



# **B**

---

## **Appendix to part II**

---

### **B.1 Supplementary tables**

#### **B.1.1 Definitions**

Term (Structural entities)	COGNATE definition	Sequence ontology term <sup>a</sup>
Transcript	An RNA synthesized on a DNA or RNA template by an RNA polymerase [SO:0000673]. In case of protein-coding genes, the RNA includes all exons ( <i>i.e.</i> , the 5' and the 3' UTRs and all CDSs) and all introns of the gene; the transcript requires modifications (RNA maturation) to be ready for translation into a protein. The transcript thus represents pre-mRNA.	<b>SO:0000673 transcript:</b> An RNA synthesized on a DNA or RNA template by an RNA polymerase. <b>SO:0000185 primary_transcript:</b> A transcript that in its initial state requires modification to be functional.
mRNA (messenger RNA)	Transcript of a protein-coding gene that has been post-transcriptionally modified. In eukaryotes, the modification typically includes 5'-capping, polyadenylation, and the splicing (removal) of introns, which can result in alternative transcripts (recombination of all or some exons). Naturally occurring mature mRNA typically consists of a 5' cap, 5' UTR, concatenated CDSs, 3' UTR, and a poly-A tail. In structural annotations, 5' cap and poly-A-tail are not indicated. Only the concatenated CDS between start codon and stop codon is translated into amino acid sequence.	<b>SO:0000233 mature_transcript:</b> A transcript which has undergone the necessary modifications, if any, for its function. In eukaryotes this includes, for example, processing of introns, cleavage, base modification, and modifications to the 5' and/or the 3' ends, other than addition of bases. In bacteria functional mRNAs are usually not modified. <b>SO:0000234 mRNA:</b> Messenger RNA is the intermediate molecule between DNA and protein. It includes UTR and coding sequences. It does not contain introns.
Exon	Any part of a gene that becomes part of the mature mRNA. Exons can be classified by their position within the mRNA and contain UTRs and CDSs in various combinations [2]. Thus, neither all exons nor all parts of them are necessarily coding. In structural annotations, UTRs and exons may be separately annotated and not overlapping; in these cases, exons coincide with CDSs.	<b>SO:0000147 exon:</b> A region of the transcript sequence within a gene which is not removed from the primary RNA transcript by RNA splicing.

**Table VII.1 – Definitions part I.** Glossary and definitions used by COGNATE. This table contains the definitions used by COGNATE and in this manuscript for structural entities. Where available, we added matching Sequence Ontology terms.

<sup>a</sup> The Sequence Ontology Browser.  
<http://www.sequenceontology.org/browser/obob.cgi>. Accessed 15 November 2016.



Term (Structural entities)	COGNATE definition	Sequence ontology term <sup>a</sup>
CDS (CoDing Sequence)	Any part of a gene that becomes translated, <i>i.e.</i> , contains information for synthesizing an amino acid sequence <sup>b</sup> . All CDSs are exonic <sup>c</sup> .	<b>SO:0000195 coding_exon:</b> An exon whereby at least one base is part of a codon (here, 'codon' is inclusive of the stop_codon). <b>SO:0000316 CDS:</b> A contiguous sequence which begins with, and includes, a start codon and ends with, and includes, a stop codon.
Intron	Every transcribed non-coding (in respect of this gene) part of a gene that is removed by RNA splicing during RNA maturation.	<b>SO:0000188 intron:</b> A region of a primary transcript that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it.
UTR (UnTranslated Region)	mRNA sequence of a protein-coding gene that is non-coding ( <i>i.e.</i> , remains untranslated) and lies 5'- or 3'-adjacent to sequences in the same mRNA that are translated (CDSs). UTRs are parts of exons <sup>d</sup> .	<b>SO:0000203 UTR:</b> Messenger RNA sequences that are untranslated and lie five prime or three prime to sequences which are translated.

**Table VII.2 – Definitions part II.** Glossary and definitions used by COGNATE. This table contains the definitions used by COGNATE and in this manuscript for structural entities. Where available, we added matching Sequence Ontology terms.

- <sup>a</sup> The Sequence Ontology Browser.  
<http://www.sequenceontology.org/browser/obob.cgi>. Accessed 15 November 2016.
- <sup>b</sup> JM MUDGE and J HARROW (2016). **The state of play in higher eukaryote gene annotation.** *Nature Reviews Genetics* 17.12, pp. 758–772. DOI: 10.1038/nrg.2016.119
- <sup>c</sup> MQ ZHANG (2002). **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 3.9, pp. 698–709. DOI: 10.1038/nrg890
- <sup>d</sup> ZHANG, 2002

Term (Measured parameters)	Definition
GC content	Amount of guanine and cytosine in a given DNA sequence, in percent.
GC content without ambiguity	Amount of guanine, cytosine, and S (G or C IUPAC ambiguity base) in a given DNA sequence, excluding ambiguous bases (NRYKMBDHFV), in percent.
CpG o/e	CpG dinucleotide depletion, normalized by the GC content of the region under scrutiny. The CpGo/e for each sequence is defined as the frequency (count/total length) of CpG dinucleotides divided by the product of the frequencies of C nucleotides and G nucleotides in the sequence <sup>a</sup>
Length	Total count of nucleotide bases/amino acids in a DNA/protein sequence, respectively.
Count	Total count of features.
Coverage	Ratio of the length of a feature covered by another, length-wise. For example, 'exon coverage of a transcript' translates to the added length of all exons divided by the length of their corresponding annotated transcript.
Density	Ratio of the count of features found along another feature. <i>E.g.</i> , the count of exons divided by the length of their corresponding annotated transcript.

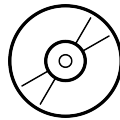
**Table VII.3 – Definitions part III.** Glossary and definitions used by COGNATE. This document contains the definitions used by COGNATE and in this manuscript for measured parameters.

<sup>a</sup> N ELANGO *et al.* (2009). **DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*.** *Proceedings of the National Academy of Sciences* 106.27, pp. 11206–11211. DOI: 10.1073/pnas.0900301106

## B.2 Electronic supplements

### B.2.1 Additional file 1: Parameter table

List of parameters recorded by COGNATE. The first sheet of this table contains all 296 parameters evaluated by COGNATE, including the output file in which to find them and explanatory comments. Sorting for parameters, the individual feature ('of') or the feature location ('per'), and files allows to quickly find a parameter of interest. The second sheet contains a comparison of the values recorded by COGNATE when analyzing the latest annotation of the *Apis mellifera* genome (genome version 4.58, annotation release 1039) to those values given in the publications of the official gene sets version 1 (MORIOKA *et al.*, 2006) and 3.2 (ELSIK *et al.*, 2014) and in the annotation report by NCBI<sup>VII.1</sup>. As an addition, we included the results of GenomeTools' `gt stat` command applied to the annotation release 103 GFF file for comparison. (XLSX, 43 kB)



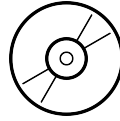
Please find this table on the attached CD at  
`./electronic_supplement/Chapter_II/Additional_file_1.xlsx`

### B.2.2 Additional file 2: Definition table

Glossary and definitions used by COGNATE (see also tables in section B.1.1). This document contains the definitions used by COGNATE and in this manuscript for structural entities and measured parameters. Where available, we added matching Sequence Ontology terms. (PDF, 110 kB)

---

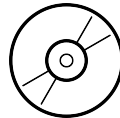
<sup>VII.1</sup> NCBI: NCBI *Apis mellifera* Annotation Release 103 report site. 2016. [https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Apis\\_mellifera/103](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Apis_mellifera/103). Last accessed 20 March 2017



Please find this document on the attached CD at  
`./electronic_supplement/Chapter_II/Additional_file_2.pdf`

### B.2.3 Additional file 3: Result table

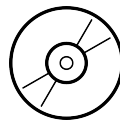
COGNATE results of analyzing exon and intron lengths of *Apis mellifera*. This data sheet contains the mean and median lengths of exons and introns, which are part of the 10,733 transcripts analyzed by COGNATE (default run, *i.e.*, using the longest of each gene's alternative transcripts). In total, 76,276 exons and 65,543 introns were taken into account. The data is visualized in Fig. II.2. (XLSX, 225 kB)



Please find this table on the attached CD at  
`./electronic_supplement/Chapter_II/Additional_file_3.xlsx`

### B.2.4 Additional file 4: The COGNATE package

This archive file contains the COGNATE package, including Perl scripts, Additional file 1: Parameter table, README, example data and output, and the GAL library. (ZIP, 566 kB)



Please find this file on the attached CD at  
`./electronic_supplement/Chapter_II/Additional_file_4.zip`

---

## Supplementary bibliography B

---

- ELANGO, N, BG HUNT, MAD GOODISMAN, and SV YI (2009). **DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*.** *Proceedings of the National Academy of Sciences* 106.27, pp. 11206–11211. DOI: 10.1073/pnas.0900301106 (cit. on p. 242).
- ELSIK, CG, KC WORLEY, AK BENNETT, M BEYE, F CAMARA, CP CHILDERS, DCd GRAAF, G DEBYSER, J DENG, B DEVREESE, E ELHAIK, JD EVANS, LJ FOSTER, D GRAUR, R GUIGO, \f \\\\$AUTHOR.LASTNAME, KJ HOFF, ME HOLDER, ME HUDSON, GJ HUNT, H JIANG, V JOSHI, RS KHETANI, P KOSAREV, CL KOVAR, J MA, R MALESZKA, RFA MORITZ, MC MUNOZ-TORRES, TD MURPHY, DM MUZNY, IF NEWSHAM, JT REESE, HM ROBERTSON, GE ROBINSON, O RUEPPELL, V SOLOVYEV, M STANKE, E STOLLE, JM TSURUDA, MV VAERENBERGH, RM WATERHOUSE, DB WEAVER, CW WHITFIELD, Y WU, EM ZDOBNOV, L ZHANG, D ZHU, RA GIBBS, and \f \\\\$AUTHOR.LASTNAME (2014). **Finding the missing honey bee genes: lessons learned from a genome upgrade.** *BMC Genomics* 15.1, p. 86. DOI: 10.1186/1471-2164-15-86 (cit. on p. 243).
- MORIOKA, M *et al.* (2006). **Insights into social insects from the genome of the honeybee *Apis mellifera*.** *Nature* 443.7114, pp. 931–949. DOI: 10.1038/nature05260 (cit. on p. 243).
- MUDGE, JM and J HARROW (2016). **The state of play in higher eukaryote gene annotation.** *Nature Reviews Genetics* 17.12, pp. 758–772. DOI: 10.1038/nrg.2016.119 (cit. on p. 241).
- ZHANG, MQ (2002). **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 3.9, pp. 698–709. DOI: 10.1038/nrg890 (cit. on p. 241).



## **Appendix to part III**

---

### **C.1 Supplementary Notes**

#### **C.1.1 Species Set**

Terrestrial animal diversity is largest in insects. Genome and gene research has been focused strongly on humans and vertebrates. Analyses as well as problems in assembly and annotation reveal that insect genomes differ from those of vertebrates, requiring specifically tailored parameter adjustments and open-minded descriptions. The seven species chosen for the present study represent four of the larger insect orders, spanning a wide range of genome sizes and life styles. For these species, manual curations have been completed within the i5k project (I5K CONSORTIUM, 2013), which ensures basically similar methods of sequencing, gene annotation, and curation.

### C.1.2 Annotation

Finding, or locating, protein-coding genes within genomic sequence comprises the delineation of canonical coding sequences (CDSs) (gene prediction *sensu stricto*, BRENNER, 1999) as well as the more general delineation of gene structures including exons and untranslated regions (UTRs), often aided by the inclusion of transcriptomic or proteomic evidence. The result is a structural annotation of genes. Despite previous struggles of differentiated delineation (e.g., ZHANG, 2002, the terms ‘prediction’ and ‘annotation’ are often used interchangeably, reflecting the latest development in gene finding algorithm implementation, which perform both steps in one run (e.g., MAKER, HOLT and YANDELL, 2011). Attaching further information on the function or role of the gene product (protein) results in ‘functional annotation’ (more detailed definition in WILBRANDT *et al.*, 2017). The success of both structural and functional annotation critically depends on assembly quality. The outcomes depend on the aim (which kind of genes will be analyzed) and decisions on procedures, databases, and evidence used.

#### Gene definition

The operational definition of ‘gene’ has been changed and adapted in the face of incoming data and knowledge (reviewed by GERSTEIN *et al.*, 2007). The Sequence Ontology (EILBECK *et al.*, 2005) defines the term ‘gene’ (SO:0000704<sup>VII.1</sup>) as follows: “A region (or regions) that includes all of the sequence elements necessary to encode a functional transcript. A gene may include regulatory regions, transcribed regions and/or other functional sequence regions.” It is important to note that there can be protein-coding and non-coding genes (*i.e.*, that will be transcribed but not translated, e.g., tRNAs) (GERSTEIN *et al.*, 2007). In the following text, we will refer exclusively to protein-coding genes, using the term as delineated in WILBRANDT *et al.* (2017).

---

<sup>VII.1</sup> The Sequence Ontology Browser.  
<http://www.sequenceontology.org/browser/obob.cgi>. Accessed 15 November 2016.



### Automated structural annotation of protein coding genes

The history and development of automated gene predictors and annotators has been reviewed by BRENT (2005, 2008) and BRENT and GUIGÓ (2004), while principles and guidance are provided by, for example, YANDELL and ENCE (2012). Thus, only a few words on the underlying rationales will be issued here.

The basic assumption of structural gene prediction is that a gene differs from the surrounding genomic sequence by specific traits, such as codon usage, and is marked by recognizable sequence patterns or signals (BURGE and KARLIN, 1997, 1998; c.f. 'Kozak rules', HUANG *et al.*, 2016; KOZAK, 1991). The application of (derived) Markov Models to model such traits improved *ab initio* predictions (*i.e.*, based on the to be analyzed sequence alone), BURGE and KARLIN, 1998. However, these traits can be species-specific, calling for individual training (model-adjustment) of gene finders (KORF, 2004). Furthermore, pure *ab initio* approaches tend to over-predict (COGHLAN and DURBIN, 2007). Corroborating an identified potential gene sequence with RNAseq (*i.e.*, aligned transcripts) or protein evidence either by sequence or profile similarity helps in more reliable prediction (BURSET and GUIGO, 1996; FICKETT, 1996), but includes a potential bias brought forth by database completeness (including taxon coverage and biases therein) and correctness (BURSET and GUIGO, 1996). Further quality control is desirable, *e.g.*, by measuring the congruency between prediction and evidence (such as implemented by MAKER as AED, HOLT and YANDELL, 2011; EILBECK *et al.*, 2005).

### Manual curation of structural gene models

Manual curation of protein-coding gene models is conducted by experts who ideally know the structure and characteristics of the gene under review from previous experience or who are acquainted with reliable evidence supporting the modeled structure. This process is made possible by visualization tools (*e.g.*, WebApollo, KÖNIG *et al.*, 2016), to which the original gene models/predictions are given as well as evidence.

As stated previously, automated annotation depends on species-specific training (KORF, 2004) to improve model parameters and thus gene delineation.

It helps, generally speaking, to know whether the analyzed genome is large with a low gene density or rather the opposite. This becomes visible in the relative frequencies of required curator actions. For example, in the annotation of the large, 'diffuse' genome of *Oncopeltus fasciatus* were many genes wrongly split, thus "curation involved merging automatic predictions far more often (6.8x) than splitting them into separate models" (PANFILIO *et al.*, 2017). In contrast, the 'compact' *Strigamia maritima* genome showed "in a significant number of cases, the automated annotation [...] fused adjacent genes, largely on the basis of confounding RNASeq evidence" (Supplement p.3 of CHIPMAN *et al.*, 2014). Required curation actions may not only depend on assembly size, but also on assembly quality, as it directly influences gene prediction accuracy.

MISRA *et al.* (2002) provided several examples of challenging gene structures and highlighted the need for common curation rules and vocabularies as well as an in-depth documentation of curation actions. Especially when facing complicated models and contradictory evidence of barely conserved genes, it is possible that the model is deteriorated by curation.

### Functional annotation of protein-coding genes

The goal of protein-coding gene prediction is to find those stretches of genomic DNA that are translated into peptides (a sequence of amino acids) and folded into functional proteins. Describing their biological properties in categories (rather than assigning one specific function) is the aim of functional annotation (SASSON *et al.*, 2006). The most reliable approach to determine these properties is experimental, but since this is largely unfeasible (limitations of time and workforce in the face of a deluge of predictions also in non-model organisms), automated workarounds, *i.e.*, automated annotation, are used (GABALDÓN, 2006; MAO *et al.*, 2005; MORIYA *et al.*, 2007; reviewed by JUNCKER *et al.*, 2009). The main rationale of functional annotation (similar sequence implies similar function; 'ortholog conjecture') has been strongly debated (*e.g.*, ALTENHOFF *et al.*, 2012; NEHRT *et al.*, 2011; ROGOZIN, 2014). During automated annotation, error propagation and lacking manual curation have been identified as confining issues (BORK and KOONIN, 1998; BRENNER, 1999; SCHNOES *et al.*, 2009), although recent developments in integrating multiple methods (FORSLUND and SONNHAMMER, 2008; HUYNEN *et al.*, 2000; SJÖLANDER, 2004)

and reciprocal control (RENTZSCH and ORENKO, 2009) have led to considerable quality improvements (ŠKUNCA *et al.*, 2012). Currently, Gene Ontology terms (ASHBURNER *et al.*, 2000; THE GENE ONTOLOGY CONSORTIUM, 2001, 2009) are most widely used to describe functional roles or properties of proteins, where Evidence Codes provide a qualitative confidence measure (FORSLUND, 2011), while pathway affiliations are described using the system of the Kyoto Encyclopedia of Genes and Genomes (KEGG, KANEHISA *et al.*, 2006; OGATA *et al.*, 1999).

### C.1.3 Extended results

#### Curator experience and participation

In total, 132 annotators participated in the seven annotation projects under scrutiny and curated 6057 gene models. A third of all curators (46, 34.8 %) took part in two or more projects and provided 3368 of all curated models (56.1 %). Only ten curators (7.6 %) participated in four or more of the projects in our data set (henceforth treated as being 'experienced' irrespective of the number and quality of contributed models) and supplied together 1468 of the 6057 curated gene models, a share of 24.4 %. Since these curators handled on average ca. 147 gene models, 5x more than the overall average of ca. 28 genes/curator (III.1 a, Additional file C.2.2: Table 2b), it is likely that their experience and selection of genes for annotation has a strong influence on curation results. This is underlined by the counts of curators shared between only two projects (Additional file III.2: Table 2c); the most extreme example is the comparison of annotation projects of the hymenopterans *Athalia rosae* (33 curators) and *Orussus abietinus* (28 curators), which share 23 curators (including 15 that only participated in these two projects).

#### Comparison of BRAKER and MAKER annotations

BRAKER predicts substantially more short gene models resulting in shorter proteins (Fig. III.4, Additional file C.2.5). This is not due to a decreased length of exons or introns, but results from a much larger fraction of single-exon genes

predicted by BRAKER in comparison to MAKER. Only up to 4 % of these single-exon genes map to multi-exon genes predicted by MAKER ('single-exon gene' → 'larger exon count' in Fig. III.4 c, Additional file C.2.5: Table 5b). In contrast, one-fifth of all BRAKER gene predictions have no MAKER counterpart, and more than half of these are single-exon genes. This indicates that although iterative training in combination with extrinsic evidence (as used by MAKER) may help to bridge long or 'confusing' introns, the majority of single-exon genes in BRAKER does not consist of 'MAKER-fragments'. It is possible that MAKER ignores single-exon genes as they are either under-represented in its training sets or because they do not reach a certain cutoff or support threshold. To check whether the single-exon genes predicted by BRAKER might constitute an entire class of gene models missed by MAKER, we spot-checked the predicted genes for underlying evidence and their mapping to the assembled genome. It appears that some of these cases were not supported by evidence, while other single-exon genes were found in an area where MAKER predicted one or two models in close proximity (Additional file C.2.6).

### Correlative trends of gene composition

In the seven analyzed species, we find a negative correlation of exon/intron count and median GC content of exons/introns (Fig. III.2 b, d). This includes that gene models with many exons/introns are more restricted/less variable, in their GC content than models with simpler structure. There is also a negative correlation of exon count and median exon length along with a restriction to a certain median exon length class (ca 190 bp) in gene models with many exons (Fig. III.2 c). In contrast to ZHU *et al.* (2009), we observe mixed trends regarding the correlation of intron count and median intron length: some species exhibit a (slightly) positive correlation (*A. rosae*, *L. decemlineata*, *O. fasciatus*), some a (slightly) negative correlation (*C. lectularius*, *F. occidentalis*, *O. abietinus*) in all sets ('complete', 'analyzed', 'automatic', 'manual'). Interestingly, there is a discrepancy between the trends found for the 'analyzed' and 'complete' sets in *A. glabripennis* (Fig. III.2 e, black brace). In all species it appears that gene models with extremely short introns have been removed/modified (Fig. III.2 e), but the majority of curated models still covers the general dispersion of the complete data sets. Despite the close phylogenetic relationship of *A. glabripennis* and *L. decemlineata* (both Chrysomelidae), there appear to be differences in the

distributions of transcript and median intron length (Fig. III.3 a) as well as between correlations of intron count and median intron length per transcript (Fig. III.2 e), which might indicate high plasticity of insect or coleopteran gene structures.

#### C.1.4 Extended material and methods

##### Data sample

Download sources and used files are listed in Additional File C.2.1. Descriptions of the i5k procedures for sequencing and annotation can be found in the respective supplementary notes of the publications on *A. glabripennis* (MCKENNA *et al.*, 2016), *C. lectularius* (BENOIT *et al.*, 2016), *L. decemlineata* (SCHOVILLE *et al.*, 2017), and *O. fasciatus* (PANFILIO *et al.*, 2017). Publications for the remaining species are in preparation, the procedures were very similar/identical.

##### Data set preparation

Not all annotations included UTRs, thus these are not considered in the present study. Prior to set preparation, all non-coding genes (*i.e.*, models without mRNA) were removed. Note that due to current limitations of data formats, we count gene parts on different scaffolds as separate genes. Furthermore, all our analyses comparing the location of gene models delineated by different algorithms or by manual curation are based on overlaps of whole genes rather than subunits of these. For the comparison of manually curated gene models with their overlapping predecessors, we excluded deleted models and newly created models, as both do not have a predecessor. Up to 42 % of the manually curated genes is new, *i.e.*, has no overlapping predecessor in the automatic annotation (Additional File C.2.4). This implies a relatively high rate of false negatives, contrary to the intuitive expectation of dominating false positives. It will be interesting to fully explore, why these genes have not been found by the automatic annotation tool and which criteria lead to the manual *de novo* annotation, as discussed, for example, in the supplement of the *A. glabripennis* genome publication (MCKENNA *et al.*, 2016). The authors stated low RNAseq

support and rapid divergence (*i.e.*, a lack of protein alignments) as factors influencing the low prediction rates of receptor gene families by MAKER.

Of the four sets used in our analyses, two do not need modification as they are given (original automated annotation, OGS). To congregate the two subsets 'analyzed manual' and 'analyzed automated', the following steps were executed for each species:

- grep manually curated genes from OGS (indicated as source or tag "ManualCuration")
- remove from this set all genes without mRNA
- use the resulting file with an in-house script to
  - ⇒ find all gene models in the automated annotation (target) that overlap the manually curated models (query) and lie on the same strand
  - ⇒ produce gff3 output files, excluding manually curated models without overlapping predecessors (*de novo* models)

To extract descriptive parameters used in the following analyses, we ran COGNATE with default parameters on all four (sub)sets.

### Non-canonical start positions in MAKER predictions

A custom script was used to determine the frequencies of those amino acids, proteins encoded by gene models predicted by MAKER started with. The amino acid fasta files of the version BCM\_version\_0.5.3-Primary\_Gene\_Set were obtained for all seven species from <https://i5k.nal.usda.gov/content/data-downloads>. Results are displayed in Additional File C.2.3.

### BRAKER-vs-MAKER analyses

For the re-annotations, transcriptomic data (RNAseq, Additional File C.2.1) obtained within the respective i5k projects was mapped to the genomes of the two hymenopterans using HISAT version 2.1.0 (KIM *et al.*, 2015). The overall alignment rate was 83.96 % for *A. rosae* and 80.87 % for *O. abietinus*. The output was transformed into sorted .bam format (bamtools, BARNETT *et al.*, 2011) and used as intrinsic evidence for the annotation with BRAKER version 1.9 (HOFF *et al.*, 2016) using default parameters. When studying *O. fasciatus*,

we directly used aligned RNAseq-evidence (i5k used TopHat2, KIM *et al.*, 2013). BRAKER predicted a total of 14,319 protein-coding genes in *A. rosae*, 14,049 in *O. abietinus*, and 75,389 in *O. fasciatus*. To assess the differences between the annotation derived from BRAKER and MAKER (*i.e.*, the original i5k annotation), we analyzed the new BRAKER annotations with COGNATE version 1.0 (WILBRANDT *et al.*, 2017) with default parameters and compared the results to the COGNATE results on the i5k annotations (see above). We identified overlapping predictions lying on the same strand along with the number of exons in query and target gene model using a custom script in both directions (*i.e.*, BRAKER → MAKER and MAKER → BRAKER) (Fig. III.4). We spot-checked the annotation situation for 55 single-exon genes predicted by BRAKER but not MAKER using the Apollo web browser interface hosted by i5k@NAL<sup>VII.2</sup> with appropriate annotation and evidence tracks (MAKER, Augustus, expression evidence for assembled transcriptomes and/or mapped raw reads, depending on the availability for each species) for the following criteria:

- presence of evidence (RNAseq, Augustus / SNAP model),
- proximity to MAKER models ( $\pm 10$  kbp and/or  $< 1.5x$  of the gene locus size off),
- N content of CDS nucleotide sequence,
- blastn (CAMACHO *et al.*, 2009) (mapping to target and other scaffolds),
- proximal MAKER models, overlapping BRAKER models.

Results are given in Additional File C.2.6.

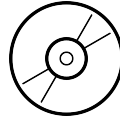
## C.2 Electronic supplements

### C.2.1 Additional file 1: Data sources and used files

List of publications, download sources, and used files for all seven species. (XLSX, 6.7 kB)

---

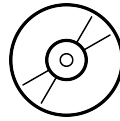
<sup>VII.2</sup> i5k Workspace @ NAL: <https://i5k.nal.usda.gov>. Last accessed 10 January 2018.



Please find this table on the attached CD at  
./electronic\_supplement/Chapter\_III/Additional\_file\_1.xlsx

### **C.2.2 Additional file 2: Curators**

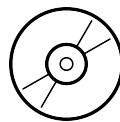
Includes a table of counts of curators and curated gene models per individual project (Supplementary Table 2a) and per 'experience group' (Supplementary Table 2b) as well as a table of counts of curators shared between pairwise compared projects (Supplementary Table 2c). In the latter table, counts of curators that participated exclusively in the compared two projects and no other are given in parentheses. (XLSX, 6.2 kB)



Please find this table on the attached CD at  
./electronic\_supplement/Chapter\_III/Additional\_file\_2.xlsx

### **C.2.3 Additional file 3: Non-canonical start codons**

List of start codons of automatically generated gene models given in percentages and absolute numbers, including median, minimum, and maximum percentages. (XLSX, 8.2 kB)

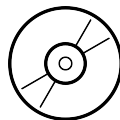


Please find this table on the attached CD at  
./electronic\_supplement/Chapter\_III/Additional\_file\_3.xlsx



### C.2.4 Additional file 4: Automated vs manually curated gene models

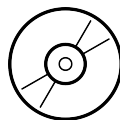
Counts of gene models that were subjected to manual curation and their overlapping automatically generated gene models, as well as counts of non-coding and de novo gene models that were not considered in this study. (XLSX, 6.2 kB)



Please find this table on the attached CD at  
`./electronic_supplement/Chapter_III/Additional_file_4.xlsx`

### C.2.5 Additional file 5: BRAKER vs MAKER

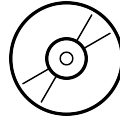
Includes as Supplementary Table 5a counts of gene models with  $x$  exons in BRAKER/MAKER overlapping gene models with  $y$  exons in MAKER/BRAKER, respectively. In Supplementary Table 5b, an excerpt of COGNATE summary statistics is given for an overview of structural parameters of gene models generated by BRAKER and MAKER. (XLSX, 8.1 kB)



Please find this table on the attached CD at  
`./electronic_supplement/Chapter_III/Additional_file_5.xlsx`

### C.2.6 Additional file 6: BRAKER-only single-exon gene models

For the three re-annotated species (*A. rosae*, *O. abietinus*, *O. fasciatus*), an excerpt of 50 gene models that were annotated only by BRAKER (no overlapping gene model in the respective MAKER annotations) and consist only of a single exon are checked for length and number of Ns. Supporting evidence and flanking/opposite MAKER predictions are listed for some of these genes. (XLSX, 18.7 kB)

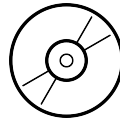


Please find this table on the attached CD at

`./electronic_supplement/Chapter_III/Additional_file_6.xlsx`

### C.2.7 Additional file 7: COGNATE results

COGNATE results of all species (except *E. occidentalis*) and all four sets used to compare automatically generated and manually curated gene models ('complete automatic' ['auto\_compl'], 'complete manual' ['man\_compl'], 'analyzed automatic' ['auto'], 'analyzed manual' ['man']) as well as COGNATE results of the three re-annotated species (*A. rosae*, *O. abietinus*, *O. fasciatus*) for the new automatically generated gene models (analogous to 'complete automatic' ['braker\_compl'] and 'analyzed automatic' ['ManCvB']). (ZIP, 170.6 MB)



Please find this file on the attached CD at

`./electronic_supplement/Chapter_III/Additional_file_7.xlsx`

---

## Supplementary bibliography C

---

- ALTENHOFF, AM, RA STUDER, M ROBINSON-RECHAVI, and C DESSIMOZ (2012). **Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs.** *PLoS Computational Biology* 8.5. Ed. by JA EISEN, e1002514. DOI: 10.1371/journal.pcbi.1002514 (cit. on p. 250).
- ASHBURNER, M, CA BALL, JA BLAKE, D BOTSTEIN, H BUTLER, JM CHERRY, AP DAVIS, K DOLINSKI, SS DWIGHT, JT EPPIG, *et al.* (2000). **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 25.1, pp. 25–29 (cit. on p. 251).
- BARNETT, DW, EK GARRISON, AR QUINLAN, MP STRÖMBERG, and GT MARTH (2011). **BamTools: a C++ API and toolkit for analyzing and managing BAM files.** *Bioinformatics* 27.12, pp. 1691–1692. DOI: 10.1093/bioinformatics/btr174 (cit. on p. 254).
- BENOIT, JB, ZN ADELMAN, K REINHARDT, A DOLAN, M POELCHAU, EC JENNINGS, EM SZUTER, RW HAGAN, H GUJAR, JN SHUKLA, F ZHU, M MOHAN, DR NELSON, AJ ROSENDALE, C DERST, V RESNIK, S WERNIG, P MENEGAZZI, C WEGENER, N PESCHEL, JM HENDERSHOT, W BLENAU, R PREDEL, PR JOHNSTON, P IOANNIDIS, RM WATERHOUSE, R NAUEN, C SCHORN, MC OTT, F MAIWALD, JS JOHNSTON, AD GONDHALEKAR, ME SCHARF, BF PETERSON, KR RAJE, BA HOTTEL, D ARMISÉN, AJJ CRUMIÈRE, PN REFKI, ME SANTOS, E SGHAIER, S VIALA, A KHILA, SJ AHN, C CHILDERS, CY LEE, H LIN, DST HUGHES, EJ DUNCAN, SC MURALI, J QU, S DUGAN, SL LEE, H CHAO, H DINH, Y HAN, H DODDAPANENI, KC WORLEY, DM MUZNY, D WHEELER, KA PANFILIO, IM VARGAS JENTZSCH, EL VARGO, W BOOTH, M FRIEDRICH, MT WEIRAUCH, MAE ANDERSON, JW JONES, O MITTAPALLI, C ZHAO, JJ ZHOU, JD EVANS, GM ATTARDO, HM ROBERTSON, EM ZDOBNOV, JMC RIBEIRO, RA GIBBS, JH WERREN, SR PALLI, C SCHAL, and S RICHARDS (2016). **Unique features of a global human ectoparasite identified through sequencing of the bed bug genome.** *Nature Communications* 7, p. 10165. DOI: 10.1038/ncomms10165 (cit. on p. 253).

- BORK, P and EV KOONIN (1998). **Predicting functions from protein sequences—where are the bottlenecks?** *Nature Genetics* 18.4, pp. 313–318. DOI: 10.1038/ng0498-313 (cit. on p. 250).
- BRENNER, SE (1999). **Errors in genome annotation.** *Trends in Genetics* 15.4, pp. 132–133. DOI: 10.1016/S0168-9525(99)01706-0 (cit. on pp. 248, 250).
- BRENT, MR (2005). **Genome annotation past, present, and future: How to define an ORF at each locus.** *Genome Research* 15.12, pp. 1777–1786. DOI: 10.1101/gr.3866105 (cit. on p. 249).
- (2008). **Steady progress and recent breakthroughs in the accuracy of automated genome annotation.** *Nature Reviews Genetics* 9.1, pp. 62–73. DOI: 10.1038/nrg2220 (cit. on p. 249).
- BRENT, MR and R GUIGÓ (2004). **Recent advances in gene structure prediction.** *Current Opinion in Structural Biology* 14.3, pp. 264–272. DOI: 10.1016/j.sbi.2004.05.007 (cit. on p. 249).
- BURGE, C and S KARLIN (1997). **Prediction of complete gene structures in human genomic DNA.** *Journal of Molecular Biology* 268.1, pp. 78–94. DOI: 10.1006/jmbi.1997.0951 (cit. on p. 249).
- (1998). **Finding the genes in genomic DNA.** *Current Opinion in Structural Biology* 8.3, pp. 346–354. DOI: 10.1016/S0959-440X(98)80069-9 (cit. on p. 249).
- BURSET, M and R GUIGO (1996). **Evaluation of gene structure prediction programs.** *Genomics* 34.3, pp. 353–367 (cit. on p. 249).
- CAMACHO, C, G COULOURIS, V AVAGYAN, N MA, J PAPADOPOULOS, K BEALER, and TL MADDEN (2009). **BLAST+: architecture and applications.** *BMC Bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421 (cit. on p. 255).
- CHIPMAN, AD *et al.* (2014). **The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*.** *PLOS Biology* 12.11, e1002005. DOI: 10.1371/journal.pbio.1002005 (cit. on p. 250).
- COGHLAN, A and R DURBIN (2007). **Genomix.** *Bioinformatics (Oxford, England)* 23.12, pp. 1468–1475. DOI: 10.1093/bioinformatics/btm133 (cit. on p. 249).
- EILBECK, K, SE LEWIS, CJ MUNGALL, M YANDELL, L STEIN, R DURBIN, and M ASHBURNER (2005). **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biology* 6.5, R44. DOI: 10.1186/gb-2005-6-5-r44 (cit. on pp. 248, 249).
- FICKETT, JW (1996). **Finding genes by computer: the state of the art.** *Trends in Genetics* 12.8, pp. 316–320. DOI: 10.1016/0168-9525(96)10038-X (cit. on p. 249).
- FORSLUND, K and ELL SONNHAMMER (2008). **Predicting protein function from domain content.** *Bioinformatics* 24.15, pp. 1681–1687 (cit. on p. 250).
- FORSLUND, K (2011). *The relationship between orthology, protein domain architecture and protein function.* Doctoral thesis: Stockholm University. 112 pp. (cit. on p. 251).
- GABALDÓN, T (2006). **Computational approaches for the prediction of protein function in the mitochondrion.** *American Journal of Physiology - Cell Physiology* 291.6, pp. C1121–C1128. DOI: 10.1152/ajpcell.00225.2006 (cit. on p. 250).

- GERSTEIN, MB, C BRUCE, JS ROZOWSKY, D ZHENG, J DU, JO KORBEL, O EMANUELSSON, ZD ZHANG, S WEISSMAN, and M SNYDER (2007). **What is a gene, post-ENCODE? History and updated definition.** *Genome Research* 17.6, pp. 669–681. DOI: 10.1101/gr.6339607 (cit. on p. 248).
- HOFF, KJ, S LANGE, A LOMSADZE, M BORODOVSKY, and M STANKE (2016). **BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.** *Bioinformatics* 32.5, pp. 767–769. DOI: 10.1093/bioinformatics/btv661 (cit. on p. 254).
- HOLT, C and M YANDELL (2011). **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 12.1, p. 491. DOI: 10.1186/1471-2105-12-491 (cit. on pp. 248, 249).
- HUANG, Y, SY CHEN, and F DENG (2016). **Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction.** *Computational and Structural Biotechnology Journal* 14, pp. 298–303. DOI: 10.1016/j.csbj.2016.07.002 (cit. on p. 249).
- HUYNEN, M, B SNEL, W LATHE, and P BORK (2000). **Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences.** *Genome Research* 10.8, pp. 1204–1210. DOI: 10.1101/gr.10.8.1204 (cit. on p. 250).
- I5K CONSORTIUM (2013). **The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment.** *Journal of Heredity* 104.5, pp. 595–600. DOI: 10.1093/jhered/est050 (cit. on p. 247).
- JUNCKER, AS, LJ JENSEN, A PIERLEONI, A BERNSEL, ML TRESS, P BORK, G VON HEIJNE, A VALENCIA, CA OUZOUNIS, R CASADIO, *et al.* (2009). **Sequence-based feature prediction and annotation of proteins.** *Genome Biology* 10.2, p. 206 (cit. on p. 250).
- KANEHISA, M, S GOTO, M HATTORI, KF AOKI-KINOSHITA, M ITOH, S KAWASHIMA, T KATAYAMA, M ARAKI, and M HIRAKAWA (2006). **From genomics to chemical genomics: new developments in KEGG.** *Nucleic acids research* 34 (suppl 1), pp. D354–D357 (cit. on p. 251).
- KIM, D, B LANGMEAD, and SL SALZBERG (2015). **HISAT: a fast spliced aligner with low memory requirements.** *Nature Methods* 12.4, pp. 357–360. DOI: 10.1038/nmeth.3317 (cit. on p. 254).
- KIM, D, G PERTEA, C TRAPNELL, H PIMENTEL, R KELLEY, and SL SALZBERG (2013). **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biology* 14, R36. DOI: 10.1186/gb-2013-14-4-r36 (cit. on p. 255).
- KÖNIG, S, LW ROMOTH, L GERISCHER, and M STANKE (2016). **Simultaneous gene finding in multiple genomes.** *Bioinformatics* 32.22, pp. 3388–3395. DOI: 10.1093/bioinformatics/btw494 (cit. on p. 249).
- KORF, I (2004). **Gene finding in novel genomes.** *BMC Bioinformatics* 5.1, p. 59. DOI: 10.1186/1471-2105-5-59 (cit. on p. 249).

- KOZAK, M (1991). **An analysis of vertebrate mRNA sequences: intimations of translational control.** *The Journal of Cell Biology* 115.4, pp. 887–903. DOI: 10.1083/jcb.115.4.887 (cit. on p. 249).
- MAO, X, T CAL, JG OLYARCHUK, and L WEI (2005). **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 21.19, pp. 3787–3793. DOI: 10.1093/bioinformatics/bti430 (cit. on p. 250).
- MCKENNA, DD, ED SCULLY, Y PAUCHET, K HOOVER, R KIRSCH, SM GEIB, RF MITCHELL, RM WATERHOUSE, SJ AHN, D ARSALA, JB BENOIT, H BLACKMON, T BLEDSOE, JH BOWSER, A BUSCH, B CALLA, H CHAO, AK CHILDERS, C CHILDERS, DJ CLARKE, L COHEN, JP DEMUTH, H DINH, H DODDAPANENI, A DOLAN, JJ DUAN, S DUGAN, M FRIEDRICH, KM GLASTAD, MAD GOODISMAN, S HADDAD, Y HAN, DST HUGHES, P IOANNIDIS, JS JOHNSTON, JW JONES, LA KUHN, DR LANCE, CY LEE, SL LEE, H LIN, JA LYNCH, AP MOCZEK, SC MURALI, DM MUZNY, DR NELSON, SR PALLI, KA PANFILIO, D PERS, MF POELCHAU, H QUAN, J QU, AM RAY, JP RINEHART, HM ROBERTSON, R ROEHRDANZ, AJ ROSENDALE, S SHIN, C SILVA, AS TORSON, IMV JENTZSCH, JH WERREN, KC WORLEY, G YOCUM, EM ZDOBNOV, RA GIBBS, and S RICHARDS (2016). **Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle–plant interface.** *Genome Biology* 17.1, p. 227. DOI: 10.1186/s13059-016-1088-8 (cit. on p. 253).
- MISRA, S, MA CROSBY, CJ MUNGALL, BB MATTHEWS, KS CAMPBELL, P HRADECKY, Y HUANG, JS KAMINKER, GH MILLBURN, SE PROCHNIK, CD SMITH, JL TUPY, EJ WHITFIELD, L BAYRAKTAROGLU, BP BERMAN, BR BETTENCOURT, SE CELNIKER, AD de GREY, RA DRYSDALE, NL HARRIS, J RICHTER, S RUSSO, AJ SCHROEDER, S SHU, M STAPLETON, C YAMADA, M ASHBURNER, WM GELBART, GM RUBIN, and SE LEWIS (2002). **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biology* 3.12, research0083.1–83.22. DOI: 10.1186/gb-2002-3-12-research0083 (cit. on p. 250).
- MORIYA, Y, M ITOH, S OKUDA, AC YOSHIZAWA, and M KANEHISA (2007). **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Research* 35 (suppl 2), W182–W185 (cit. on p. 250).
- NEHRT, NL, WT CLARK, P RADIVOJAC, and MW HAHN (2011). **Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals.** *PLoS Computational Biology* 7.6. Ed. by A RZHETSKY, e1002073. DOI: 10.1371/journal.pcbi.1002073 (cit. on p. 250).
- OGATA, H, S GOTO, K SATO, W FUJIBUCHI, H BONO, and M KANEHISA (1999). **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Research* 27.1, pp. 29–34 (cit. on p. 251).
- PANFILIO, KA, IMV JENTZSCH, JB BENOIT, D EREZYILMAZ, Y SUZUKI, S COLELLA, HM ROBERTSON, MF POELCHAU, RM WATERHOUSE, P IOANNIDIS, MT WEIRAUCH, DST HUGHES, SC MURALI, JH WERREN, CGC JACOBS, EJ DUNCAN, D ARMISÉN, BMI VREEDE, P BAA-PUYOULET, CS BERGER, Cc CHANG, H CHAO, MJM CHEN, YT

- CHEN, CP CHILDERS, AD CHIPMAN, AG CRIDGE, AJJ CRUMIÈRE, PK DEARDEN, EM DIDION, H DINH, H DODDAPANENI, A DOLAN, S DUGAN-PEREZ, CG EXTAVOUR, G FEBVVAY, M FRIEDRICH, N GINZBURG, Y HAN, P HEGER, T HORN, Ym HSIAO, EC JENNINGS, JS JOHNSTON, TE JONES, JW JONES, A KHILA, S KOELZER, V KOVACOVA, M LEASK, SL LEE, CY LEE, MR LOVEGROVE, HL LU, Y LU, PJ MOORE, MC MUNOZ-TORRES, DM MUZNY, SR PALLI, N PARISOT, L PICK, M PORTER, J QU, PN REFKI, R RICHTER, R RIVERA-POMAR, AJ ROSENDALE, S ROTH, L SACHS, ME SANTOS, J SEIBERT, E SGHAIER, JN SHUKLA, RJ STANCLIFFE, O TIDSWELL, L TRAVERSO, Mvd ZEE, S VIALA, KC WORLEY, EM ZDOBNOV, RA GIBBS, and S RICHARDS (2017). **Molecular evolutionary trends and feeding ecology diversification in the Hemiptera, anchored by the milkweed bug genome.** *bioRxiv*, p. 201731. DOI: 10.1101/201731 (cit. on pp. 250, 253).
- RENTZSCH, R and CA ORENGO (2009). **Protein function prediction – the power of multiplicity.** *Trends in Biotechnology* 27.4, pp. 210–219. DOI: 10.1016/j.tibtech.2009.01.002 (cit. on p. 251).
- ROGOZIN, IB (2014). **Complexity of Gene Expression Evolution after Duplication: Protein Dosage Rebalancing.** *Genetics Research International* 2014, e516508. DOI: 10.1155/2014/516508 (cit. on p. 250).
- SASSON, O, N KAPLAN, and M LINIAL (2006). **Functional annotation prediction: All for one and one for all.** *Protein Science* 15.6, pp. 1557–1562. DOI: 10.1110/ps.062185706 (cit. on p. 250).
- SCHNOES, AM, SD BROWN, I DODEVSKI, and PC BABBITT (2009). **Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies.** *PLoS Computational Biology* 5.12, e1000605. DOI: 10.1371/journal.pcbi.1000605 (cit. on p. 250).
- SCHOVILLE, SD, YH CHEN, MN ANDERSSON, JB BENOIT, A BHANDARI, JH BOWSER, K BREVIK, K CAPPELLE, MJM CHEN, AK CHILDERS, C CHILDERS, O CHRISTIAENS, J CLEMENTS, EM DIDION, EN ELPIDINA, P ENGSONTIA, M FRIEDRICH, I GARCIA-ROBLES, RA GIBBS, C GOSWAMI, A GRAPPUTO, K GRUDEN, M GRYNBERG, B HENRISSAT, EC JENNINGS, JW JONES, M KALSI, SA KHAN, A KUMAR, F LI, V LOMBARD, X MA, A MARTYNOV, NJ MILLER, RF MITCHELL, M MUNOZ-TORRES, A MUSZEWSKA, B OPPERT, SR PALLI, KA PANFILIO, Y PAUCHET, LC PERKIN, M PETEK, MF POELCHAU, E RECORD, JP RINEHART, HM ROBERTSON, AJ ROSENDALE, VM RUIZ-ARROYO, G SMAGGHE, Z SZENDREI, GWC THOMAS, AS TORSON, IMV JENTZSCH, MT WEIRAUCH, AD YATES, GD YOCUM, JS YOON, and S RICHARDS (2017). **A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae).** *bioRxiv*, p. 192641. DOI: 10.1101/192641 (cit. on p. 253).
- SJÖLANDER, K (2004). **Phylogenomic inference of protein molecular function: advances and challenges.** *Bioinformatics* 20.2, pp. 170–179 (cit. on p. 250).
- ŠKUNCA, N, A ALTENHOFF, and C DESSIMOZ (2012). **Quality of Computationally Inferred Gene Ontology Annotations.** *PLOS Computational Biology* 8.5, e1002533. DOI: 10.1371/journal.pcbi.1002533 (cit. on p. 251).

- THE GENE ONTOLOGY CONSORTIUM (2001). **Creating the Gene Ontology Resource: Design and Implementation.** *Genome Research* 11.8, pp. 1425–1433. DOI: 10.1101/gr.180801 (cit. on p. 251).
- (2009). **The Gene Ontology in 2010: extensions and refinements.** *Nucleic Acids Research* 38 (Database), pp. D331–D335. DOI: 10.1093/nar/gkp1018 (cit. on p. 251).
- WILBRANDT, J, B MISOF, and O NIEHUIS (2017). **COGNATE: comparative gene annotation characterizer.** *BMC Genomics* 18.1, p. 535. DOI: 10.1186/s12864-017-3870-8 (cit. on pp. 248, 255).
- YANDELL, M and D ENCE (2012). **A beginner’s guide to eukaryotic genome annotation.** *Nature Reviews Genetics* 13.5, pp. 329–342. DOI: 10.1038/nrg3174 (cit. on p. 249).
- ZHANG, MQ (2002). **Computational prediction of eukaryotic protein-coding genes.** *Nature Reviews Genetics* 3.9, pp. 698–709. DOI: 10.1038/nrg890 (cit. on p. 248).
- ZHU, L, Y ZHANG, W ZHANG, S YANG, JQ CHEN, and D TIAN (2009). **Patterns of exon-intron architecture variation of genes in eukaryotic genomes.** *BMC Genomics* 10, p. 47. DOI: 10.1186/1471-2164-10-47 (cit. on p. 252).



# D

---

## Appendix to part IV

---

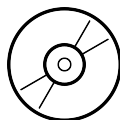
### D.1 Electronic supplements

#### D.1.1 Additional file 1: Species and COGNATE median data

Supplementary table with data version information and an ordered representation of COGNATE batch results; comprising several sheets:

- a) Sources
- b) COGNATE - Assembly parameters
- c) COGNATE - Component sizes
- d) COGNATE - Median transcript features
- e) COGNATE - Summary statistics

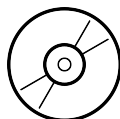
(XLSX, 128 kB)



Please find this table on the attached CD at  
`./electronic_supplement/Chapter_IV/Additional_file_1.xlsx`

### **D.1.2 Additional file 2: COGNATE results**

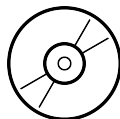
Complete COGNATE result sets for the 13 analyzed species. (ZIP, 113.7 MB)



Please find this file on the attached CD at  
`./electronic_supplement/Chapter_IV/Additional_file_2.zip`

### **D.1.3 Additional file 3: Density plots**

Density plots of selected parameters measured by COGNATE. (PDF, 631 kB)



Please find this document on the attached CD at  
`./electronic_supplement/Chapter_IV/Additional_file_3.pdf`

## Appendix to part V

---

### E.1 Supplementary Notes

#### E.1.1 Assembly calls and library combinations

For the internal assemblathon, 22 assemblies were generated with the tools Platanus and AllPaths-LG as described below. Both tools used quality trimmed reads.

For each species, 21 different assemblies were generated with Platanus version 1.2.4 (KAJITANI *et al.*, 2014). These differed in the library combinations used in the three consecutive steps of assembly, as outlined in Table VII.4.

The three steps for the Platanus assembly (contig assembly, scaffolding, and gap closing) were called as follows (this is a schematic call example and would require adaption to run).

```

# contig assembly
platanus assemble -o SPECIES_ASSEMBLY_ID \
  -n 0 -t $NSLOTS -m 60 \
  -f LIB_A_1.fq LIB_A_2.fq LIB_B_1.fq LIB_B_2.fq \
  LIB_C_1.fq LIB_C_2.fq LIB_D_1.fq LIB_D_2.fq

# scaffolding
platanus scaffold -o SPECIES_ASSEMBLY_ID \
  -t $NSLOTS -c SPECIES_ASSEMBLY_ID_contig.fa \
  -b SPECIES_ASSEMBLY_ID_contigBubble.fa \
  -IP1 LIB_A_1.fq LIB_A_2.fq \
  -IP2 LIB_B_1.fq LIB_B_2.fq \
  -OP3 LIB_C_1.fq LIB_C_2.fq \
  -OP4 LIB_D_1.fq LIB_D_2.fq

# gap closing
platanus gap_close -o SPECIES_ASSEMBLY_ID \
  -t $NSLOTS -c SPECIES_ASSEMBLY_ID_scaffold.fa \
  -IP1 LIB_A_1.fq LIB_A_2.fq \
  -IP2 LIB_B_1.fq LIB_B_2.fq \
  -OP3 LIB_C_1.fq LIB_C_2.fq \
  -OP4 LIB_D_1.fq LIB_D_2.fq

```

In each of the three steps, between two and four libraries could be employed (indicated in the above call by libraries A, B, C, and D). In the table indicating the different combinations for each step and each assembly run (Table VII.4), the four libraries are encoded as follows:

1. 250 bp, paired-end reads
2. 800 bp, paired-end reads
3. 3 kbp, mate-pair reads
4. 8 kbp, mate pair reads

The platanus-assemblies were compared to a single Allpaths-LG (version 52488, BUTLER *et al.*, 2008) assembly. This tool proceeds in a two-step process of configuration and running. For the configuration step, two files are required to indicate locations and specifications of used libraries. These files follow this scheme:

Assembly ID	Step 1	Step 2	Step 3
0	1,2	1,2,3	1,2,3,4
1	1,2	1,2,3,4	1,2
2	1,2	1,2,3,4	1,2,3
3	1,2	1,2,3,4	1,2,3,4
4	1,2	1,2,3,4	2,3,4
5	1,2	2,3,4	1,2
6	1,2	2,3,4	1,2,3
7	1,2	2,3,4	1,2,3,4
8	1,2,3	1,2,3,4	1,2
9	1,2,3	1,2,3,4	1,2,3
10	1,2,3	1,2,3,4	1,2,3,4
11	1,2,3	2,3,4	1,2
12	1,2,3	2,3,4	1,2,3
13	1,2,3,4	2,3,4	1,2,3,4
14	1,2,3,4	2,3,4	1,2
15	1,2	2,3	1,2
16	1,2	2,3	1,2,3
17	1,2	1,2,3	2,3
18	1,2	1,2,3	1,2,3
19	1,2,3	2,3	1,2,3
20	1,2,3	1,2,3	1,2,3

**Table VII.4 – Assembly with Platanus: library combinations.** For each assembly ID and each step of the Platanus assembly process (contig assembly, scaffolding, and gap closing), a combination of DNA read libraries is given. See text for a disclosure of the libraries.

```

# library locations: in_groups.csv
group_name, library_name, file_name
PE1, illumina1, /DIR/SPECIES_250_trimmed_paired_*.fq
Jump1, illumina2, /DIR/SPECIES_800_trimmed_paired_*.fq
Jump2, illumina3, /DIR/SPECIES_3kb_trimmed_paired_*.fq
Jump3, illumina4, /DIR/SPECIES_8kb_trimmed_paired_*.fq

# library specifications: in_libs.csv
library_name, project_name, organism_name, type, paired, \
    frag_size, frag_stddev, insert_size, insert_stddev, \
    read_orientation, genomic_start, genomic_end
illumina1, SPECIES, SPECIES_FULLL, fragment, 1, 250, 20, \
    , , inward, ,
illumina2, SPECIES, SPECIES_FULLL, jumping, 1, 800, 80, \
    , , inward, ,
illumina3, SPECIES, SPECIES_FULLL, jumping, 1, 3000, 300, \
    , , outward, ,
illumina4, SPECIES, SPECIES_FULLL, jumping, 1, 8000, 800, \
    , , outward, ,

```

The steps of configuration and assembly with Allpaths-LG were run with the following call scheme. Note that the `SIZE_BP` parameter was adjusted for each species based on the genome size estimation using Jellyfish (MARÇAIS and KINGSFORD, 2011).

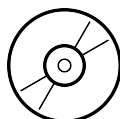
```
# configuration
PrepareAllpathsInputs.pl \
    DATA_DIR=/DIR/SPECIES_trimmed/allpaths/mydata \
    PICARD_TOOLS_DIR=/DIR/picard-tools-2.0.1/ \
    PHRED_64=1 \
    PLOIDY=2 \
    GENOME_SIZE=SIZE_BP

# run assembly
RunPathsLG PRE=./SPECIES_trimmed \
REFERENCE_NAME=allpaths \
DATA_SUBDIR=mydata \
RUN = myrun \
TARGETS=standard \
THREADS=$NSLOTS
OVERWRITE=TRUE
```

## E.2 Electronic supplements

### E.2.1 Additional file 1: MySQL database

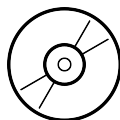
MySQL database scheme and database dump. (ZIP, 168.3 MB)



Please find this file on the attached CD at `./electronic_supplement/Chapter_V/Additional_file_1.zip`

### E.2.2 Additional file 2: MySQL commands

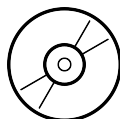
MySQL commands. (TXT, 7 kB)



Please find this file on the attached CD at  
`./electronic_supplement/Chapter_V/Additional_file_2.txt`

### **E.2.3 Additional file 3: Perl and R scripts**

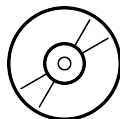
Custom scripts in Perl (data handling, etc.) and R (plotting). (ZIP, 36 kB)



Please find this file on the attached CD at  
`./electronic_supplement/Chapter_V/Additional_file_3.zip`

### **E.2.4 Additional file 4: COGNATE results**

Complete COGNATE data for the 26 analyzed species. (ZIP, 346.8 MB)



Please find this file on the attached CD at  
`./electronic_supplement/Chapter_V/Additional_file_4.zip`



---

## Supplementary bibliography E

---

- BUTLER, J, I MACCALLUM, M KLEBER, IA SHLYAKHTER, MK BELMONTE, ES LANDER, C NUSBAUM, and DB JAFFE (2008). **ALLPATHS: De novo assembly of whole-genome shotgun microreads**. *Genome Research* 18.5, pp. 810–820. DOI: 10.1101/gr.7337908 (cit. on p. 268).
- KAJITANI, R, K TOSHIMOTO, H NOGUCHI, A TOYODA, Y OGURA, M OKUNO, M YABANA, M HARADA, E NAGAYASU, H MARUYAMA, Y KOHARA, A FUJIYAMA, T HAYASHI, and T ITOH (2014). **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads**. *Genome Research* 24.8, pp. 1384–1395. DOI: 10.1101/gr.170720.113 (cit. on p. 267).
- MARÇAIS, G and C KINGSFORD (2011). **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers**. *Bioinformatics* 27.6, pp. 764–770. DOI: 10.1093/bioinformatics/btr011 (cit. on p. 270).



---

# List of Figures

---

## I GENERAL INTRODUCTION

I.1	Poster: The cassette metaphor . . . . .	6
I.2	Publications in genomics . . . . .	12
I.3	Research map . . . . .	30

## II THE TOOL COGNATE

II.1	Information flow in COGNATE . . . . .	58
II.2	Comparing means and medians . . . . .	66
II.3	Missing data in publications . . . . .	68

## III AUTOMATICALLY GENERATED VS MANUALLY CURATED MODELS

III.1	Curator experience and dilemmas . . . . .	90
III.2	Automatically generated and manually curated annotation comparison I . . . . .	95
III.3	Automatically generated and manually curated annotation comparison II . . . . .	97
III.4	Comparison of different annotation pipelines . . . . .	100

## IV HYMENOPTERAN REPERTOIRE FEATURES

IV.1	Genome and component sizes . . . . .	135
------	--------------------------------------	-----

IV.2	Trends of assembly size and genomic and gene structure parameter correlations . . . . .	136
IV.3	Comparison of median lengths and counts of gene elements	138
IV.4	Comparison of lengths and counts of gene elements for all genes . . . . .	140
IV.5	Comparison of GC content across species . . . . .	142
IV.6	Comparison of median GC content to count of gene element	144

## V CONSERVATION CLASSES OF THE REPERTOIRES

V.1	Phylogenetic relationships between species of the sample .	164
V.2	Gene repertoire and BUSCO values . . . . .	177
V.3	Gene structural parameters for all transcripts . . . . .	179
V.4	Comparison of two plot types illustrating the universality and duplicability of ortholog groups . . . . .	180
V.5	Ortholog groups show similar patterns of universality and single-copyness . . . . .	181
V.6	Medians of structural gene parameters compared across conservation classes and copy states . . . . .	183
V.7	Transcript lengths of conservation classes and copy states .	185
V.8	Protein lengths of conservation classes and copy states . . .	187
V.9	Exon lengths of conservation classes and copy states . . . .	188
V.10	Exon counts of conservation classes and copy states . . . . .	189
V.11	Intron lengths of conservation classes and copy states . . .	190
V.12	Protein domain count per transcript . . . . .	191
V.13	Core-specific protein domain count per transcript . . . . .	193
V.14	Protein domain and arrangement diversity . . . . .	195

## VII APPENDICES

VII.1	Poster: Towards a better understanding . . . . .	234
VII.2	Poster: Data basis, tool choice, and human review . . . . .	236

---

# List of Tables

---

## **I GENERAL INTRODUCTION**

I.1	Early milestones of genome sequencing . . . . .	10
-----	---	----

## **III AUTOMATICALLY GENERATED VS MANUALLY CURATED MODELS**

III.1	Overview of basic parameters . . . . .	93
-------	--	----

## **IV HYMENOPTERAN REPERTOIRE FEATURES**

IV.1	Species sample . . . . .	130
------	--------------------------	-----

## **V CONSERVATION CLASSES OF THE REPERTOIRES**

V.1	Species sample . . . . .	166
-----	--------------------------	-----

## **VII APPENDICES**

VII.1	Definitions I . . . . .	240
VII.2	Definitions II . . . . .	241
VII.3	Definitions III . . . . .	242
VII.4	Assembly with Platanus: library combinations . . . . .	269



---

## List of Abbreviations

---

**bp** basepairs.

**CDS** coding sequence.

**CpG** cytosine-phosphor-guanine, a dinucleotide.

**DNA** deoxyribonucleic acid.

**Gbp** gigabasepairs.

**GC** guanine and cytosine (two DNA bases).

**HGT** horizontal gene transfer.

**kB** kilobyte.

**kbp** kilobasepairs.

**LCA** last common ancestor.

**MB** megabyte.

**Mbp** megabasepairs (sometimes also referred to as Mb).

**NGS** next generation sequencing.

**SCS** scaffold or contig sequence.

**TE** transposable element.

**UTR** untranslated region.





---

## Danksagung

---

**E**S GAB ZEITEN, in denen ich nicht daran geglaubt habe, dass ich an diesen Punkt kommen und eine fertige Doktorarbeit vorlegen können würde. Dazu, dass ich diesen Weg gehen konnte, haben viele Menschen Wichtiges beigetragen, von fruchtbaren Diskussionen bis hin zu süßer Nervenahrung. Bei ihnen allen, auch den hier nicht genannten, aber nicht vergessenen, möchte ich mich bedanken.

**M**EINE BEIDEN Doktorväter OLIVER NIEHUIS und BERNHARD MISOF haben mich beständig unterstützt mit Ansporn, Verständnis, Rat und Kritik. Ich danke ihnen dafür und nehme sie mir zum Vorbild.

**B**ESONDERER DANK gebührt auch JAN PHILIP OEYEN, ohne den die letzten Wochen unendlich viel schwerer (und weniger bunt) gewesen wären. Außerdem möchte ich mich bei den Doktoranden und Studierenden des Museums bedanken, nicht zuletzt für ihr Vertrauen und den bereitwilligen Austausch. Ich habe viel von ihnen gelernt.

**I**CH DANKE ebenso meinen Koautoren und Kollegen — insbesondere RACHEL WERNECK, TANJA ZIESMANN, JONAS EBERLE, ALEXANDER DONATH, SIMON KÄFER, MALTE PETERSEN und PANOS PROVATARIS — für ihre hilfreichen Beiträge und alles andere, sowie den weiteren Mitgliedern meiner Kommission, GABRIELE KÖNIG und DIETMAR QUANDT, für ihre Bereitschaft, diese Arbeit

zu begutachten. Für die Möglichkeit zu Forschen danke ich dem Zoologischen Forschungsmuseum A. Koenig und der Deutschen Forschungsgemeinschaft.

**Z**U GUTER LETZT möchte ich meiner Familie für den bedingungslosen und geduldigen Rückhalt danken. Am wichtigsten aber waren Unterstützung, Bekräftigung und der unerschütterliche Glaube an mich von HANNES.

**D**ANKE.

---

# Erklärung

---

Hiermit versichere ich an Eides statt, dass ich diese Arbeit selbständig verfasst, keine anderen Quellen und Hilfsmittel als die angegebenen benutzt und die Stellen der Arbeit, die anderen Werken dem Wortlaut oder Sinn nach entnommen sind, kenntlich gemacht habe. Beiträge von Koautoren und Kollegen zur Originalpublikation oder Datensammlung sind am jeweiligen Kapitelanfang gelistet.

Für die Erstellung der vorliegenden Arbeit wurde keine fremde Hilfe, insbesondere keine entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen.

Diese Arbeit hat in dieser oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegen und ich habe keine früheren Promotionsversuche unternommen.

Bonn, 11. Juni 2018

Jeanne Wilbrandt