



Phylogenetics of the world's largest beetle family (Coleoptera: Staphylinidae) A methodological exploration

Kypke, Janina Lisa

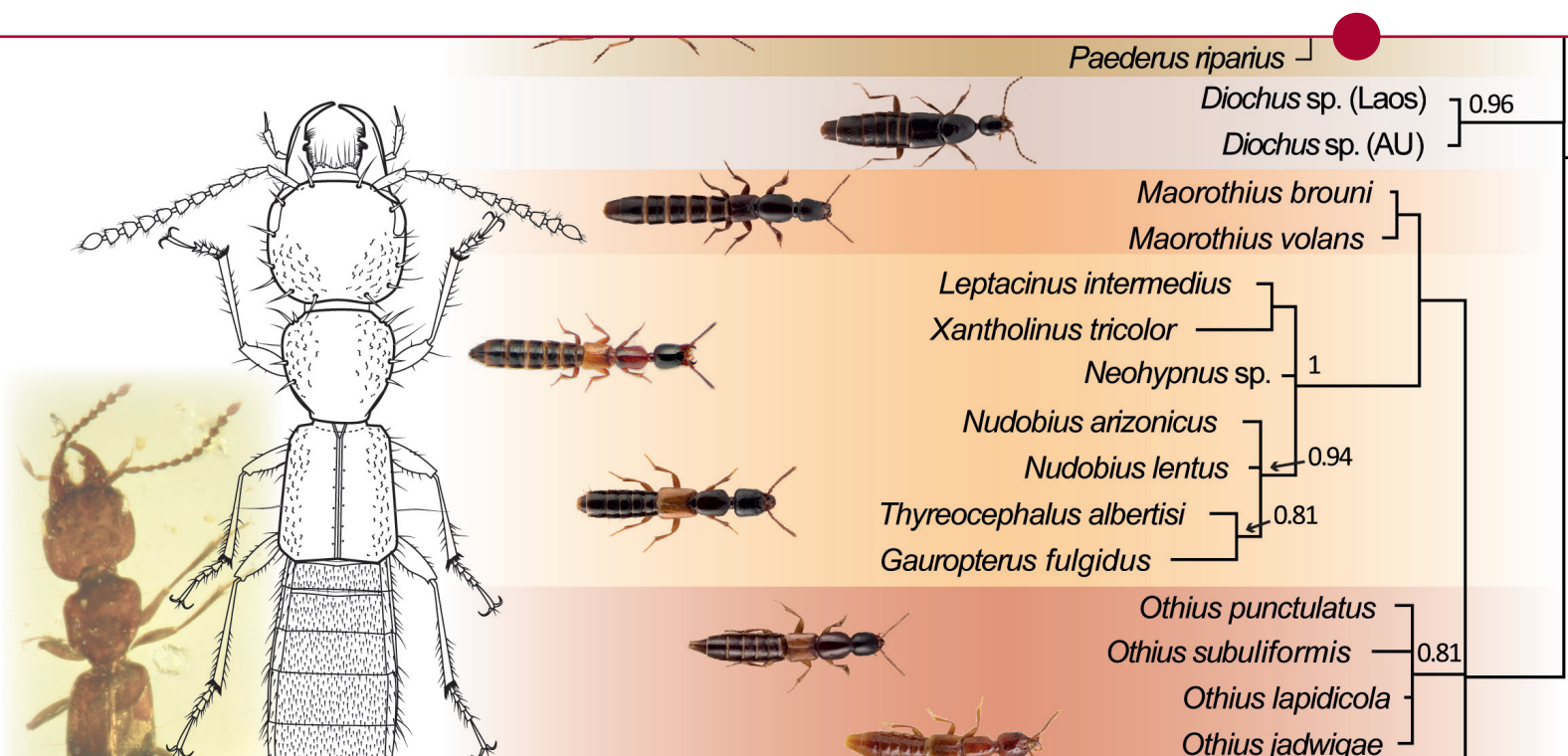
Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Kypke, J. L. (2018). *Phylogenetics of the world's largest beetle family (Coleoptera: Staphylinidae): A methodological exploration*. Natural History Museum of Denmark, Faculty of Science, University of Copenhagen.



PhD thesis | Janina L. Kypke



Phylogenetics of the world's largest beetle family (Coleoptera: Staphylinidae)

A methodological exploration

Academic advisor | Alexey Y. Solodovnikov

Submitted | 23 December 2018

PhD thesis |

Journal of

Psychology

Volume 123

Number 4

2023

ISSN 0000-0000

DOI: 10.1002/abc

Copyright © 2023

John Wiley & Sons, Ltd.

All rights reserved.

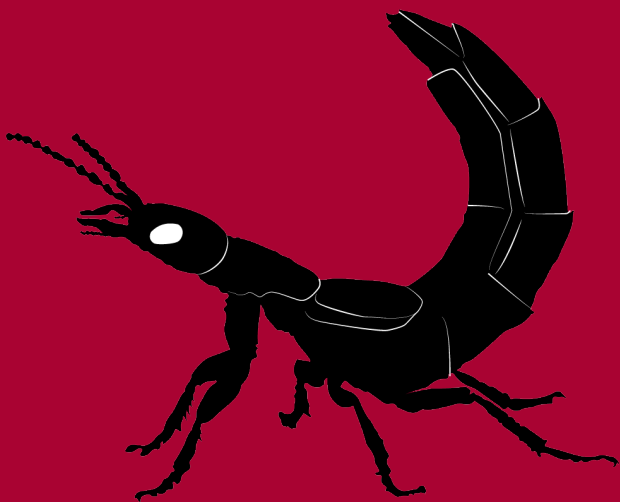
Published online

in Wiley Online

Library on [Date]

[Date]

[Date]



Institution Copenhagen University
Department Natural History Museum of Denmark
Section Biosystematics Section
Author Janina Lisa Kypke

Title **Phylogenetics of the world's largest beetle family (Coleoptera: Staphylinidae)
A methodological exploration**

Funding This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 642241. This material reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains.



Academic advisor Dr Alexey Y. Solodovnikov
Submitted 23 December 2018

*This thesis is dedicated to my family, my love,
and dear friends close and far.*

Table of Contents

Abstract	IV
Resume	V
Acknowledgements	VI
Introduction	1
Chapter 1	5
Exploration of NGS-based mixed omics data leads to first deep-level phylogeny for the world's largest animal family	
Chapter 2	75
The past and the present through phylogenetic analysis: the rove beetle tribe Othiini now and 99 Ma	
Chapter 3	94
Every cloud has a silver lining: X-ray micro-CT reveals Orsunius rove beetle in Rovno amber from a specimen inaccessible to light microscopy	

Abstract

Rove beetles (Coleoptera: Staphylinidae) are the largest family of all animals that dominate ground-based cryptic habitats of any terrestrial landscape globally, inhabiting our planet for at least 200Ma. They are the best example where sheer species-richness hinders the reconstruction a robust phylogeny, rendering this group extremely challenging for systematic studies. Despite a considerable interest in them and the growing amount of taxonomic research, this major section of the planetary Tree of Life is still largely unknown.

Systematic entomologists generally agree that rove beetles are a monophylum and the majority of their 32 subfamilies seem to be well-defined lineages. However, with regards to groupings of subfamilies or even their ranking, very little is widely accepted. All hitherto performed attempts to infer an overall backbone phylogeny of staphylinids using only morphological characters of crown groups or a small number of genes have failed.

Ultimately, a comprehensive phylogeny for this hyper-diverse family should be inferred using a combined analysis that makes use of available morphological and genetic traits, and considering both the crown and the stem group diversity.

This motivated me to target exactly these missing links during my PhD. Overall, I wanted to enforce the application of soundly established methods by testing different technological advancements. So I acquired skills that enabled me to 1) use various rapidly generated genomic markers in this group for phylogenomic inference; 2) integrate data from extant and extinct species for phylogeny reconstruction; and 3) facilitate character mining in fossil specimens that are always challenging. With respect to those three key-skills, the thesis consists of three chapters: one manuscript on phylogenomics in preparation, one published paper with the morphology-based phylogenetic analysis of both extant and extinct taxa, and one paper in press on the application of technological advancements in the study of amber fossils.

Resume

Rovbiller (Coleoptera: Staphylinidae) er den største familie af alle dyr, de dominerer jordbaserede kryptiske levesteder af ethvert landskab i hele verdenen, og har beboet vores planet i mindst 200 mio. år. De er det bedste eksempel, hvor ren artsrigdom forhindrer udviklingen af en robust fylogeni, hvilket gør gruppen ekstremt udfordrende for systematiske undersøgelser. På trods af en betydelig interesse for dem og en voksende mængde taksonomisk forskning er denne store del af livets træ stadig stort set ukendt.

Systematiske entomologer er generelt enige om, at rovbille danner et monofylum, og størstedelen af deres 32 underfamilier synes at være veldefinerede afgreninger. Men med hensyn til grupperingerne af underfamilierne og deres rangering, er der meget lidt som er bred accepteret. Alle hidtil udførte forsøg på at udlede en overordnet fylogeni af staphylinider, der kun bruger morfologiske karakterer eller et lille antal gener har mislykkedes.

I sidste ende skal en omfattende fylogeni for denne hypermangfoldige familie udledes ved hjælp af en kombineret analyse, der gør brug af tilgængelige morfologiske og genetiske træk.

Dette motiverede mig til at målrette mit arbejde mod præcis disse manglende forbindelser under min PhD. Samlet set ønskede jeg at afprøve anvendelsen af veletablerede metoder ved at teste forskellige teknologiske fremskridt. Så jeg erhvervede færdigheder, der gjorde det muligt for mig at 1) bruge forskellige hurtigt genererede genomiske markører i denne gruppe til fylogenomiske antagelser; 2) Integrere data fra eksisterende og uddøde arter til fylogenetiske rekonstruktioner; og 3) lette afsøgningen af nye karakterer i fossile prøver, der altid er udfordrende. Med hensyn til disse tre nøglekompetencer består afhandlingen af tre kapitler: et manuskript om fylogenomik der er under udarbejdelse, en udgivet artikel med den morfologibaserede fylogenetiske analyse af både eksisterende og uddøde taxa og en artikel i tryk om anvendelse af teknologiske fremskridt i undersøgelsen af ravfossiler.

Acknowledgements

They say it takes a village to raise a child and I feel the same about my PhD. The past three years I was probably influenced by so many people, they could fill a village. First and foremost, I would like to thank Dr Alexey Solodovnikov for being the best supervisor I could have hoped for during the past three years. Thank you for believing in me and for always staying positive, even if things did not always turn out the way we anticipated. I was lucky to be part of the Big4, a network that brought me to many places around Europe and connected me to fellow researchers that have become my friends. I think it's fair to say that the Zoological Museum has become something like a second home and I will surely miss the friendly atmosphere at the Biosystematics Section. Thank you Josh Jenkins Shaw and Igor Orlov for being such great office mates, for the chit-chat and the fun times, but also your help, especially with regards to identifying beetles. Many thanks also to the other members of the Solodovnikov lab and, although a former member, Dagmara Żyła. Next to these professional connections, the Danish work-life balance allowed me to build new friendships and to become a passionate Crossfitter, neither of which I want to miss again. Last, but definitely not least, I want to thank my family and my partner. Even though you can only grasp what I have been doing, you never failed to support me and even though you live abroad, you are always there for me. Thank you.

Introduction

With more than 63,000 described recent and over 400 fossil species (Alfred Newton, pers. comm.), the rove beetles (Coleoptera: Staphylinidae) are the largest family of all eukaryotic organisms (Figure 1). They are dominating various ground-based and cryptic habitats of any terrestrial landscape on the planet, except Antarctica. Based on the paleontological data, this beetle family evolved through a period of more than 200 million of years (Cai *et al.* 2012; 2014). Rove beetles are among the best examples of a super-diverse group where we know very little to nothing about their phylogeny. The sheer species-richness renders this group extremely challenging for phylogenetic and thus systematic studies on all levels. The truth is that, despite a considerable interest in rove beetles and the growing amount of taxonomic

research, this major section of the planetary Tree of Life is still largely unknown.

Systematic entomologists generally agree that rove beetles are a monophylum and the majority of their 32 subfamilies seem to be well-defined lineages. However, with regards to groupings of subfamilies or even their ranking, very little is widely accepted robust knowledge. The same applies to tribes and subtribes inside many subfamilies, especially species-rich ones like Staphylininae, Paederinae or Aleocharinae. All to date performed attempts to infer phylogenies at various levels had their limitations. While attempts to reconstruct an overall backbone phylogeny of staphylinids failed entirely. There is growing evidence that the mainstream phylogenetic work in rove beetles using morphological characters of crown groups or

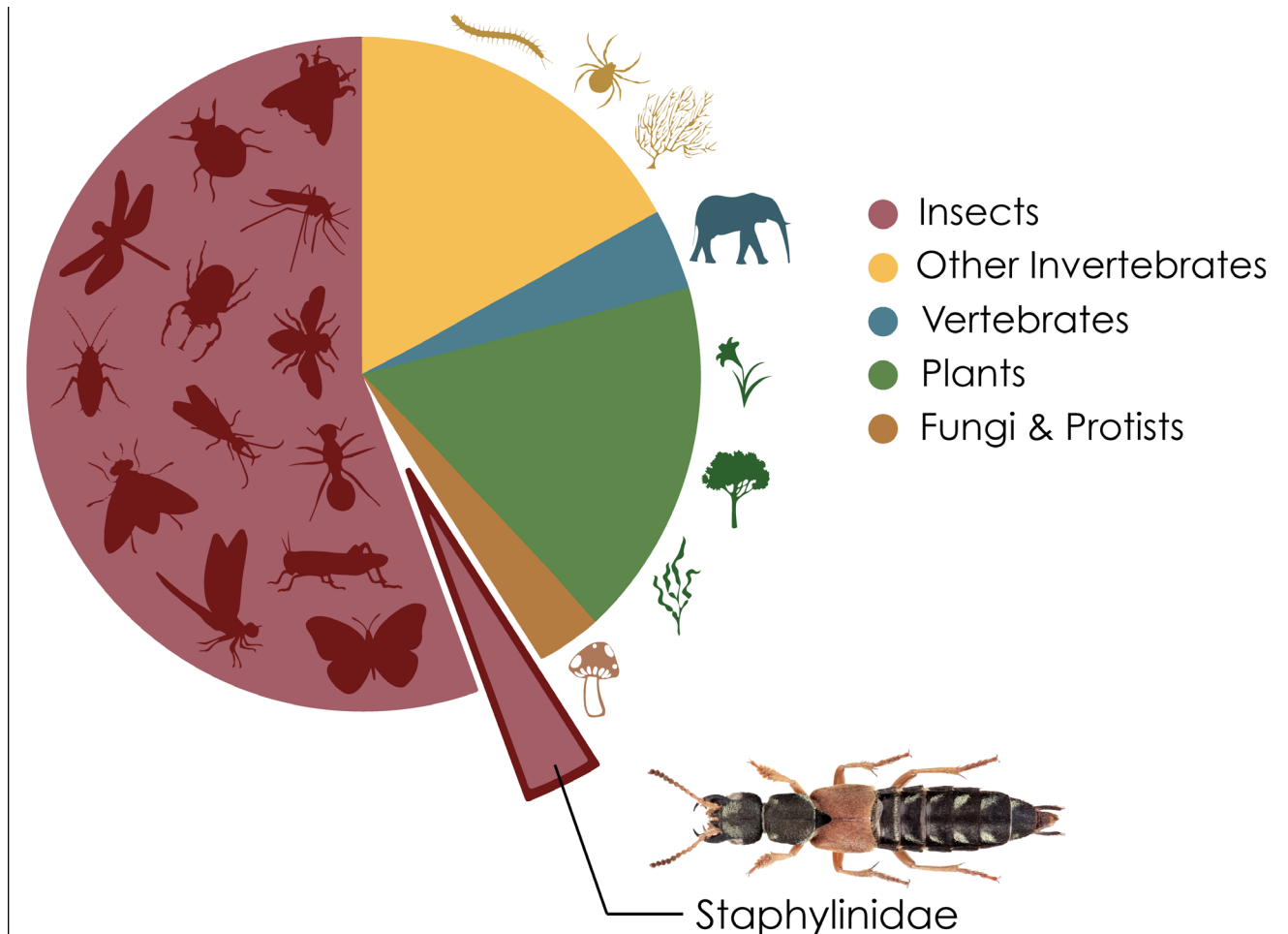


Figure 1. Total number of all catalogued species on Earth from major organismal groups of Eukaryotes. Species numbers from: IUCN, 2017; Ahn et al. 2017.

a small number of genes must be upgraded to new approaches like phylogenomics and total-evidence phylogenetics.

This is what motivated me to use the three years time of my PhD to target exactly these missing links that have hindered the progress of rove beetle systematics. Ultimately, I think that a comprehensive phylogeny for this hyper-diverse family should be inferred using a combined analysis that makes use of all available information, i.e. the morphological

as well as all possible genetic traits, not only considering the crown group diversity, but also the stem lineages as far as they are available as fossils (Figure 2). In order to become the person that can run such combined analyses, I tested some of the available technological advancements on members of this group and acquired a number of complementary skills from the fields of classical entomology and systematics, paleontology, bioinformatics and phylogenomics. These skills allow me to

- 1) use rapidly generated genomic data of various kinds for phylogenomic inference;
- 2) integrate molecular and morphological data from extant and extinct species for the phylogenetic reconstruction; and
- 3) mine characters in fossil specimens of varied preservation quality using light microscopy and modern X-ray micro-computed tomography (micro-CT).

The first and the main chapter of my thesis aims to bring rove beetle phylogenetics into the new era of phylogenomics. It is

an experimental study of 993 single-copy nuclear genes with a total alignment length of 494,743 amino-acid sites that were identified in 19 genomes, 11 transcriptomes and 40 nuclear protein-coding genes for a total of 33 and 57 species. This experiment leads to a first resolved backbone phylogeny of the Staphylinidae and shows that it is safe and promising to invest more money and time in a much larger analysis using the pipeline that I assembled in an interdisciplinary team.

In the second chapter I demonstrate how a morphology-based phylogenetic analysis

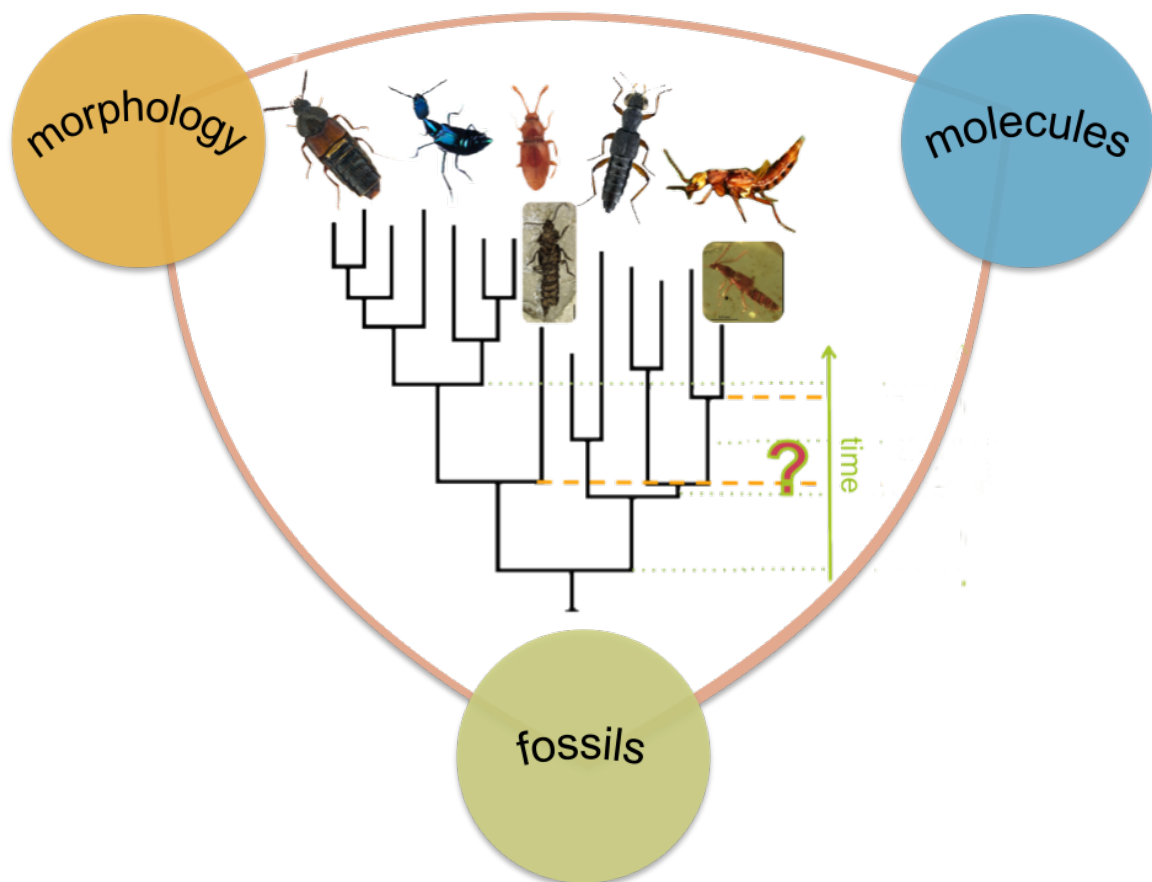


Figure 2. Desired holistic approach to the Staphylinidae Tree of Life.

involving both crown and stem groups places two newly discovered puzzling fossil species entrapped in 100 Ma Burmese amber within the extant tribe Othiini (Staphylinidae: Staphylininae). This analysis of 65 characters for 38 representatives of the subfamilies Staphylininae and Paederinae is the first time that a morphology-based phylogeny is in very high congruence with the earlier molecular analyses of 6 and 7 genes for this group.

The key importance of stem lineages for phylogenetics motivates the third and final chapter that explores the capabilities of micro-CT to study amber fossils inaccessible to light microscopy. It shows that Baltic amber fossils covered in gaseous froth that most likely formed shortly after the insect's entrapment, which made them inaccessible for light microscopy, can be studied after the 3D image reconstruction process. This technique can bring many important fossils to the light of science that are now neglected due to their 'poor' preservation.

Practical limitations of the PhD did not let me show how all these various technological advances I aimed for can work together, for example in a combined analysis of the genomic and morphological data from the extant and

extinct species (Figure 2). However, my time as a PhD fellow developed me into a researcher that can do this type of analysis in the future.

References

- Ahn, K-J, Cho, Y-B, Kim, Y-H, Yoo, I-S, Newton, AF (2017) Checklist of the Staphylinidae (Coleoptera) in Korea. *Journal of Asia-Pacific Biodiversity*, 10(3): 279–336. DOI: 10.1016/j.japb.2017.06.006.
- Cai C, Huang D, Thayer MK, Newton AF (2012) Glypholomatine Rove Beetles (Coleoptera: Staphylinidae): a Southern Hemisphere Recent Group Recorded from the Middle Jurassic of China. *Journal of the Kansas Entomological Society*, 85:239–244. DOI: 10.2317/JKES120531.1.
- Cai CY, Huang DY, Newton AF, Thayer MK (2014) *Mesapatetica aenigmatica*, a new genus and species of rove beetles (Coleoptera, Staphylinidae) from the Middle Jurassic of China. *Journal of the Kansas Entomological Society*, 87:219–224.
- IUCN (2017) Numbers of threatened species by major groups of organisms (1996–2017). Available from: http://cmsdocs.s3.amazonaws.com/summarystats/2017-1_Summary_Stats_Page_Documents/2017_1_RL_Stats_Table_1.pdf

Chapter 1

Exploration of NGS-based mixed omics data leads to first deep-level phylogeny for the world's largest animal family

Janina L. Kypke, Hermes E. Escalona, Maxim Nesterenko,
Piotr Kussakin, Alex Predeus, Alexey Solodovnikov

Exploration of NGS-based mixed omics data leads to first deep-level phylogeny for the world's largest animal family

Janina L. Kypke¹, Hermes E. Escalona², Maxim Nesterenko³, Piotr Kussakin³, Alex Predeus³, Alexey Solodovnikov¹

¹Zoological Museum, Biosystematics Section, Natural History Museum of Denmark, Copenhagen, Denmark;

²Australian National Insect Collection, CSIRO-National Research Collections, Canberra, ACT, Australia;

³Bioinformatics Institute, St. Petersburg, Russia.

Abstract

Rove beetles (Staphylinidae) are the largest family of insects and of all living animals, but their internal evolutionary relationships are unknown. We addressed this issue by formulating a phylogenomic analysis based on the design of an ortholog database from 10 reference genomes and consisting of 3,822 single-copy nuclear genes suitable for phylogenetic analyses. We demonstrated its integration with heterogeneous genomic resources into a single dataset. We *de novo* sequenced 16 staphylinid genomes with low coverage and analyzed them together with three available genomes, 10 transcriptomes and 40 additional genes. The final dataset is a comprehensive representation of Staphylinoidea (6 of 6 families) and Staphylinidae (13 of 32 subfamilies). Maximum likelihood (ML) analyses of 993 genes with a total alignment length of 494,743 amino-acid sites in datasets of 33 and 57 species led to the first resolved backbone phylogeny of Staphylinidae. Major findings are: 1) the omaliine group of subfamilies is sister to all remaining Staphylinidae; 2) the oxyteline group is monophyletic and includes the former family Silphidae; 3) the tachyporine group is polyphyletic; 4) the subfamily Tachyporinae is polyphyletic and nested within the staphylinine group of subfamilies.

Table of Contents

INTRODUCTION	8
ROVE BEETLES, STAPHYLINIDAE	9
TOWARDS A STAPHYLINID BACKBONE PHYLOGENY	10
ROLE OF MOLECULAR DATA	11
STAPHYLINOLOGY ENTERS THE PHYLOGENOMIC ERA	13
RETRIEVING PHYLOGENETICALLY INFORMATIVE PROTEIN-CODING GENES FROM FRAGMENTARY GENOMIC DATA	16
MATERIAL AND METHODS	20
NEWLY GENERATED GENOMIC DATA	20
TRANSCRIPTOMIC DATA	23
RETRIEVAL OF REMAINING DATA	24
ORTHOLOGY PREDICTION	26
DATA DEPOSITION	29
PHYLOGENETIC ANALYSIS	29
RESULTS	35
RAW DATA PRE-PROCESSING AND ASSEMBLIES	35
MINING THE ORTHOLOGOUS GENE SET	36
PHYLOGENETIC ANALYSIS	38
DISCUSSION	41
STAPHYLINID PHYLOGENOMICS VS. THE <i>STATUS QUO</i> OF SYSTEMATICS	41
ADVANCES AND DRAWBACKS OF INTEGRATE -OMIC DATASETS	46
PROSPECTS FOR THE FUTURE	49
CONCLUSIONS	50
AUTHOR CONTRIBUTIONS	51
ACKNOWLEDGEMENTS	51
REFERENCES	51
SUPPLEMENTARY MATERIAL	59

Chapter 1

Introduction

Willi Hennig's legacy, the phylogenetic method, flourished in a powerful discipline that allied with new technologies to reveal the planetary tree of life (ToL). Insects are arguably a particularly strong biotic agent that occur in all terrestrial habitats and the most speciose of the ToL (about 1 M described species, IUCN 2018). Their evolutionary success, proven by both extreme lineage diversity and ubiquity, is a strong incentive to study their evolutionary history, but also an impediment.

With the development of statistics-based inference tools that enable a researcher to infer objective degrees of confidence in their phylogenetic results, along with the emergence of large-scale methods of sequencing DNA, phylogenetics or its new derivation 'phylogenomics', matured into a powerful quantitative science. It became more rigorous at reconstructing difficult sections of the ToL, like insects. As a result, modern biology is gaining a reasonably well-resolved and widely accepted backbone phylogeny for all organismal groups and more detailed phylogenies for its major phyla. Naturally, phylogenies of various groups charismatic or relevant for humans, like for example vertebrates, are resolved to an impressive degree of confidence and precision (Teeling *et al.* 2005; Beck *et al.* 2006; Pyron & Wiens 2011; Prum *et al.* 2015; Kumar *et al.* 2017; Hara *et al.* 2018; Malinsky *et al.* 2018). But considering that ca. 95% of all animal species are invertebrates, most of which are insects (IUCN 2018), there is still much to do.

Rove beetles (Staphylinidae), the largest family of insects and of all living animals, are a fascinating group of organisms and an example of a large part of the ToL that is unresolved. There are > 63,137 described extant species of rove beetles grouped in 3,870 genera and 32 subfamilies (Grebennikov & Newton 2009; Ahn *et al.* 2017). Species estimates for rove beetles are most likely an underestimation as every year new species are described. It is widely agreed among systematic entomologists that rove beetles are a monophylum and the majority of their subfamilies seem to be well-defined lineages (Thayer 2016).

The rove beetle faunae are poorly to extremely poorly known worldwide, except the Central-European part of Eurasia. In addition, many genera and tribes that supposedly organize the rove beetle diversity into taxonomic units are non-monophyletic, more or less artificial taxonomic units, as shown by modern phylogenetic analyses and revisions. In résumé, the ToL-Staphylinidae-branch is largely unresolved. The relationships between subfamilies are still under debate, as is the status of some of them and even some staphylinoid but formally non-rove beetle families (e.g. the family of carrion beetles, Silphidae) (Figure 1) (Hansen 1997; Beutel & Molenda 1997; Hunt *et al.* 2007; McKenna *et al.* 2014, 2015; King *et al.* 2015; Zhang *et al.* 2018).

With the overall goal to resolve the backbone phylogeny of Staphylinidae, this study explores state-of-the-art techniques and presents the first phylogenomic study across Staphylinoidea, exploring

the relationships of the rove beetle's major diversifying lineages based on 993 genes and 494,743 amino-acid sites.

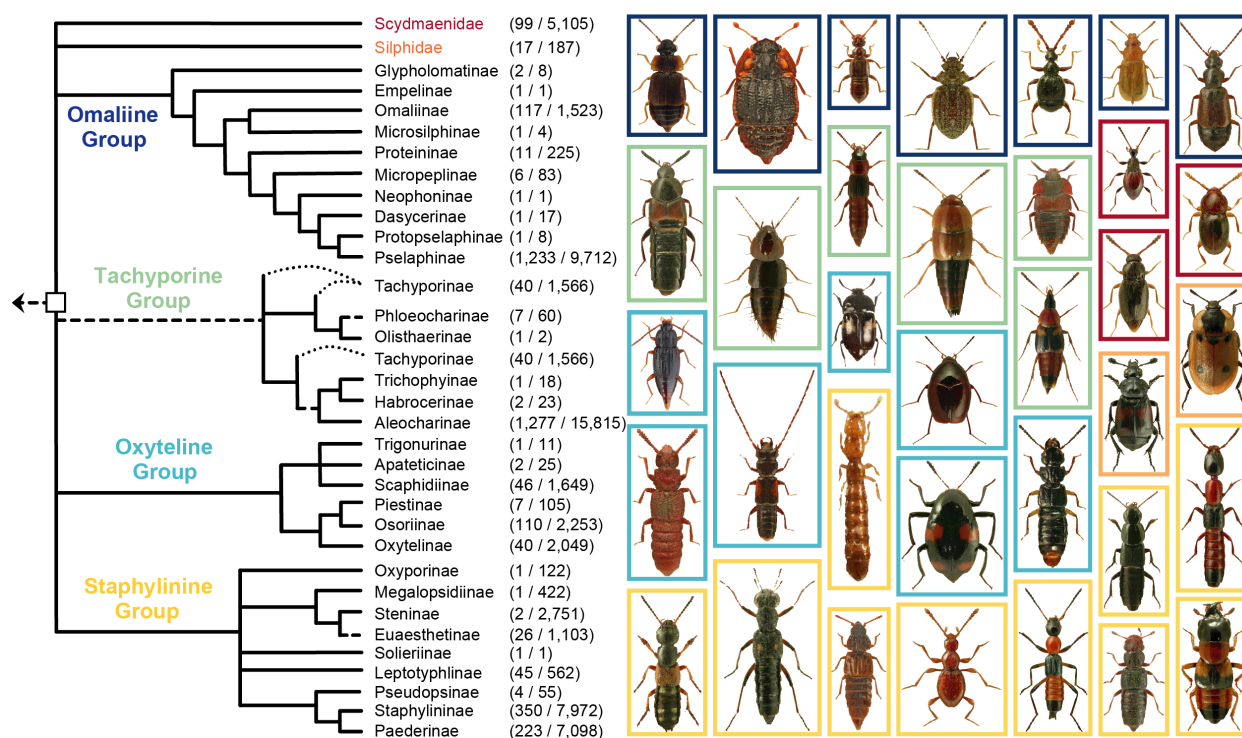


Figure 1. Current phylogeny and diversity of Staphylinidae. Left-hand side: Consensus backbone phylogeny after Thayer (2016); values in parentheses are genus / species numbers for the (sub)families. Right-hand side: rove beetles from four groups of subfamilies; frame colors refer to their respective group in the tree on the left; Red: Scydmaenidae; Orange: Silphidae; Staphylinidae: Dark blue, omaliine group; Green, tachyporine group; Light blue, oxyteline group; Yellow, staphylinine group. Images from <http://zin.ru/> (copyrights: Stanislav Krejčík, K.V. Makarov) and <http://coleoptera.ksib.pl/?l=en> (copyrights: Lech Borowiec).

Rove beetles, Staphylinidae

Most rove beetles do not look like 'typical' beetles, because of their short elytra that cover compactly folded wings. Short elytra free the flexible abdomen and enable rove beetles to penetrate small crevices and to occupy different ecological niches inaccessible to other winged insects or even beetles, without losing the ability to fly (Blum 1979; Hammond 1979; Lawrence & Newton 1982).

Rove beetles are an abundant element in all habitable terrestrial biomes of the world, mainly as ground-based predators or decomposers (Thayer 2016). They are an old lineage of beetles, documented by a diverse fossil record that starts from the Middle Jurassic (ca. 176-165 Ma) (Cai *et al.* 2012, 2014). The ca. 226 Ma old Late Triassic fossil *Leehermania prorova* Chatzimanolis *et al.*, 2012, claimed to be the earliest staphylinid (Chatzimanolis *et al.* 2012) was shown to be a myxophagan close to the family Hydroscaphidae (Fikáček *et al.* in prep.).

The rove beetle fossil record is rapidly growing - currently 430 described fossil species (Alfred Newton, pers. comm.) - and shows the former existence of a number of stem lineages (Solodovnikov *et al.* 2013; Jałoszyński *et al.* 2018), as well as a long presence of many crown groups (Yamamoto

Chapter 1

2016b; Żyła *et al.* 2017; Jałoszyński *et al.* 2018). These fossils have been used to generate time-calibrated phylogenies at different taxonomical ranks (Maruyama & Parker 2017; Brunke *et al.* 2017; Zhang & Zhou 2018; Song & Ahn 2018) and even a time-calibrated phylogeny with a few representatives of the entire family, but based on only three genes (Zhang & Zhou 2013).

The available fossil record and a number of phylogenetic analyses, even though limited, indicate complex cladogenetic patterns in time and space for this family. The habitus diversity of rove beetles is stunning (Figure 1), with many forms looking like 'typical', slender rove beetles with short elytra, as well as all sorts of derived shapes. Based on the aforementioned, reconstructing the rove beetle phylogeny would be a major contribution to the planetary ToL and form a knowledge framework to inform many biodiversity-related fields of science.

Towards a Staphylinid backbone phylogeny

There have been attempts to resolve the phylogeny of rove beetles as a whole, to delimit their major lineages and their relationships. At first, these were intuitive non-phylogenetic classification systems of early authors who were arranging taxa in some order in their collections and systematic monographs (e.g. (Ganglbauer 1895; Coiffait 1972). With the onset of statistical phylogenetic methods for the analysis of morphological characters, the first investigations towards understanding the relationships amongst Staphylinidae as a whole were conducted. The analyses were based mainly on exoskeletal characters of adults and, to a lesser extent since often unknown, also larvae (Lawrence & Newton 1982; Hansen 1997; Beutel & Molenda 1997; Beutel & Leschen 2005).

However, results of these analyses were largely conflicting with each other. One of the high impact and long lasting arrangements of rove beetle subfamilies was the four groups of subfamilies by (Lawrence & Newton 1982, 1995) (Figure 1). These groups were predominantly established based on morphological characters of larvae and adults, but also integrating other organismal characteristics, e.g. mode of feeding. Out of those four, the tachyporine group was the most uncertain and potentially polyphyletic. Upon erection of those groups, there were no hypotheses regarding their relationships at the base of the family-ToL. Scydmaenidae and Silphidae were also left unresolved and outside Staphylinidae, because at that time existing hypotheses about their respective sister subfamily were conflicting (Hansen 1997; Beutel & Molenda 1997).

Scydmaenidae have since been integrated in Staphylinidae based on the comprehensive phylogenetic analysis of 206 morphological characters of larvae and adults (Grebennikov & Newton 2009), which was later confirmed with molecular (McKenna *et al.* 2014) and fossil data (Żyła *et al.* 2017). Silphidae continue to branch out within the tachyporine group in more recent molecular phylogenetic analyses, but the taxon sampling in those was too incomplete to be certain about the family's placement (McKenna *et al.* 2014, 2015; Yamamoto *et al.*, in prep), so that it remains treated as a family

close to Staphylinidae. The rove beetle subfamily groups by Lawrence and Newton (1982, 1995) are still in use, even though more and more evidence, most often based on molecular data, appears against their monophyly. Due to that, Chatzimanolis (2018) recently deliberately rejected them, however, without proposing any alternative.

The integration of Scydmaenidae within Staphylinidae is one example where analyses based on morphology successfully resolved the phylogenetic relationships on the subfamily level. An even earlier example is the peculiar-looking family Pselaphinae (former Pselaphidae) that was shown to be a lineage inside Staphylinidae first based on cladistic analysis of morphological characters (Newton & Thayer 1995). The consideration of stem lineages preserved as fossils, in addition to the morphology of recent taxa, can help to solve phylogenetic puzzles that occurred deeper in time, even when formal cladistic analysis is not done. More examples are the close relationships between the morphologically puzzling subfamilies Dasycerinae and Neophoninae inferred from a transitional fossil (Yamamoto 2016a), also confirmed in the molecular study by McKenna *et al.* (2014). These examples show that the key to unequivocal phylogenetic results is a carefully selected, large-enough taxon sampling of species that represent the lineage in question, preferably including fossils, with a well-composed morphological character matrix. But most of the attempts to resolve the deeper divergences of the rove beetle phylogeny as a whole did not fulfill these criteria and lacked some subfamilies or used homoplasious characters within the taxa sampled as terminals for analyses.

Role of molecular data

As recently summarized (Gusarov 2018), molecular phylogenetics of rove beetles came into play near the end of the last century and by now brought us towards a better understanding of some lineages. However, it did not improve our knowledge of the rove beetle backbone phylogeny either. This is mainly because rove beetle molecular phylogenetics did not go beyond a taxon- and marker-limited Sanger sequencing approach. The biggest phylogeny in terms of taxon coverage (130 taxa of 31 subfamilies) (McKenna *et al.* 2014) used only two markers, while the biggest phylogeny in terms of marker diversity (95 genes, Zhang *et al.* 2018) used only 14 rove beetle representatives from 10 subfamilies. Even though Zhang *et al.* (2018) used the more cost-effective next-generation sequencing (NGS) technology to rapidly sequence those 95 genes for each taxon, their approach is still primer-based and requires running PCR amplification for the target genes before sequencing is possible. On the positive side, this approach only sequences the targeted genes, which reduces the overall costs, compared to sequencing all extracted genetic information and afterwards selecting the target regions that are suitable for phylogenetic inference. On the downside, experience has shown that especially the PCR amplification step can be tricky. Often, the protocols need to be adjusted from one gene to

Chapter 1

another and also amongst species. To apply this approach across the entire family and many taxa, long times in the laboratory are required before the samples can be sequenced quickly.

However, there are other ways to generate large amounts of data across a sizeable taxon sampling in conjunction with NGS techniques that have been applied in other beetle superfamilies or insect orders. Examples are phylogenetic inferences based on

- 1) sequencing genomes (Neafsey *et al.* 2015);
- 2) mitochondrial genomes (mitogenomes) (Cameron *et al.* 2007; Fenn *et al.* 2008; Hua *et al.* 2009; Talavera & Vila 2011; Simon & Hadrys 2013; Li *et al.* 2015; Timmermans *et al.* 2015; Song *et al.* 2016; López-López & Vogler 2017; Linard *et al.* 2018);
- 3) transcriptomes (Letsch & Simon 2013; Misof *et al.* 2014; Dell'Ampio *et al.* 2014; Boussau *et al.* 2014);
- 4) target capturing specific DNA fragments prior to sequencing, also referred to as anchored hybrid enrichment, a reduced representation technique that uses oligonucleotide 'baits' and is independent of the genome size; 4 a) based on specific regions in transcriptomes and genomes (Young *et al.* 2016; Peters *et al.* 2017; Bank *et al.* 2017; Haddad *et al.* 2018; Espeland *et al.* 2018; Shin *et al.* 2018); or 4 b) noncoding ultraconserved elements (UCEs) (Blaimer *et al.* 2015; Branstetter *et al.* 2017; Van Dam *et al.* 2017); and
- 5) mixed genomic datasets (Kawahara & Breinholt 2014; Kusy *et al.* 2018a,b).

Genomic data has been generated unevenly across insects and was initially often driven by the model organisms in the group. The primary reason for that were very high sequencing costs that did not encourage researchers to experiment with little-studied organisms on a large scale. However, since the 2000s sequencing costs have continuously been decreasing and it has become a viable option. Many of the aforementioned studies already sequenced libraries of non-model organisms and demonstrated that these can be integrated as well. Additionally, -omic data of all sorts (transcriptomes, genomes, target-captured genes) can relatively easily be analyzed jointly, even if the data was initially sequenced for something other than phylogenetic analysis. Finally, genomic datasets for phylogenomic inference can be developed further by adding morphological characters of extant lineages (e.g. in Peters *et al.* 2014), fossils, and other relevant traits. These are encouraging incentives to aim for a robust phylogenetic hypothesis based on an NGS approach, rather than Sanger sequencing.

In summary, rove beetles are a lineage for which we currently have only a vague idea of their phylogeny (Figure 1), mostly based on morphology and some molecular analyses of very few genes (except Zhang *et al.* 2018). Previous attempts to resolve the backbone of this family's phylogeny failed either because few species were sampled or few characters (morphological or genetic) were analyzed or both. This is a megadiverse group where NGS technologies have never been used to reconstruct their phylogeny but where they could prove to be very promising. Among major phylogenetic ambigu-

ities relevant for Staphylinidae are the early phylogenetic splits, the rank and sister group of Silphidae within Staphylinoidea, as well as the monophyly of both the tachyporine group and the subfamily Tachyporinae within Staphylinidae. Considering the technological advancements and lower NGS sequencing costs, here, we aim to perform the first experimental phylogenomic analyses of the rove beetles using all available whole genome markers and investing in shotgun sequencing of several new genomes. Given a potential risk of failure, we limited new genomes to sixteen and thus, our overall taxon sample was small. However, it was comprehensive enough to assess our results for clades where, as believed, we know the sister-group relationships, and, on the contrary, we know little to nothing. In particular, this study tests if different genomic data sources can be combined to make use of the available heterogeneous data in addition to new data generated in a more standardized way.

Staphylinology Enters the Phylogenomic Era

In 2015, when this project began, the only publicly available relevant genomic data were the sequenced and annotated genome of the carrion beetle (Silphidae) *Nicrophorus vespilloides* Herbst, 1783 (Cunningham *et al.* 2015). Furthermore, there was a transcriptome of *Aleochara curtula* (Goeze, 1777) (Misof *et al.* 2014). The challenge was to find the most suitable approach to use NGS technology to find the necessary number of markers for the phylogenetic reconstruction of the family, considering the available time and existing data in conjunction with the diversity of the group. The different types of genomic data usually require specific input material as well as processing of the sequenced reads before the data can be used in an analysis. Like two sides of a coin, there are benefits and disadvantages, which will briefly be explained in this section.

The largest data source that can be generated are nuclear genomes. However, thorough sequencing and annotation of entire genomes for a representative number of rove beetles for their phylogenetic analysis is beyond the current capacity of any systematic entomology lab. This is because the *de novo* assembly and annotation of just one well-sequenced genome consumes a lot of time and resources. It usually requires to estimate the genome size, then the use of different sequencing platforms to obtain both short and long reads to cover the entirety of the genome several times to obtain a reliable consensus (Mahmoud *et al.* 2017). Additionally, for the *de novo* genome assembly the reads are assembled and a genome map that fluorescently labels specific sequence recognition sites is first generated and then used to correct errors in the assembly (O'Connor 2008). For the annotation transcripts of the same species and proteomes of well-known model organisms are mapped onto the genome (Adams 2008). If not already available, for a more complete annotation several transcriptomes of the same species should be generated as the gene expression changes at differing times and differs

Chapter 1

amongst tissues. The functional annotation of predicted gene models then requires further computational processing.

A second option, that has become available recently for more researchers as sequencing costs decreased, is called whole genome re-sequencing, which leads to low-coverage genomes. In comparison to whole genomes, the idea is not to sequence as much of the entire genome as possible, but rather to sequence enough to assemble reads to scaffolds and to identify a large number (few thousands) of genes that are suitable for phylogenetic inference. This approach does not enable one to annotate the obtained genome, as assemblies are too crude. Prior to sequencing, the genome size should be estimated to make sure that a reasonable and even sequence coverage can be reached across all species. In comparison to other organisms, insect genomes are of moderate size (ca. 500 Megabases (Mb) on average) (He *et al.* 2016). However, due to common gene or chromosome and even rare entire genome duplications (Tsutsui *et al.* 2008), the size of a genome can vary from species to species - even amongst congeneric species (Geraci *et al.* 2007), and in some species even amongst opposite sexes (Montiel *et al.* 2012), and it is not correlated with organismal complexity (He *et al.* 2016). It can be estimated prior to sequencing using either densitometry or flow cytometry (Hare & Johnston 2011), or afterwards by conducting a k-mer analysis (He *et al.* 2016). Prior estimation is recommended to be done with fresh or flash-frozen samples, but has been demonstrated to work on ethanol-preserved samples of Crustaceans (Jeffery & Gregory 2014). The actual difficulty of using low-coverage genomes for phylogenetic reconstructions is to find the suitable genes without being able to annotate the genome and without a well-sequenced and annotated reference genome of a close relative (at best congeneric) that the new genome could directly be mapped to. The only way to find suitable genes is bioinformatically intensive and time-consuming: In a nutshell, one needs to create a 'database' of suitable genes identified in well-sequenced and annotated reference genomes, which can be used to pull out complementary gene sequences in the low-coverage genomes using gene models and reciprocal searches that apply the BLAST algorithm (Altschul *et al.* 1990).

The same principle is applied for the 'target capture' (also called 'target enrichment') approach, in which 'baits' (also called 'probes') are designed based on the sequences of the genes in the database. Those baits are then used prior to sequencing. They literally capture those genes that are most similar to themselves. Furthermore, they are magnetic and can be retained using magnetic beads, to separate the sequences of interest from the remaining extracted DNA. Like this, only captured DNA will be sequenced (Mayer *et al.* 2016). This approach is very smart and reduces sequencing costs since only genes of interest are sequenced. However, the development of a properly working bait kit is a time-consuming process, that is bioinformatically intensive and involves at least one test-run, because theoretically designed baits do not always manage to capture the genes they should, for example if divergences are too large. Also, the cheaper sequencing costs are outweighed by expensive custom-

designed bait kits and additional time needed in the laboratory to perform the target capture and slightly more difficult library preparation.

Sequencing whole mitochondrial genomes in comparison to nuclear genomes is very easy, because in multicellular eukaryotes mitogenomes usually occur in high copy number per cell and are much smaller. Additionally, it has been shown that they can be sequenced from pooled extracts of specimens collected in bulk using metagenomic sequencing techniques and without the need for prior identification (Malé *et al.* 2014; Linard *et al.* 2016). The metagenomic approach for mitogenomes makes this a very cheap option. One of the upsides of sequencing mitogenomes is at the same time a downside for phylogenetics: with their small size the number of genes is limited to ca. 40 (Simon & Hadrys 2013). Furthermore, they evolve at a much faster rate than nuclear genes, which can lead to unresolved deep phylogenetic splits (Talavera & Vila 2011). They often show signs of heterotachy (within site rate variations throughout time (Lopez *et al.* 2002). Substitution rate models that try to detect heterotachy and to account for it exist for both bayesian inference and maximum likelihood (Pons *et al.* 2010), but especially using mitogenomes, this problem needs special attention and models that assume constant substitution rates (gamma distribution) should be avoided. Failure to detect heterotachous sites can lead to long branch attraction (LBA) and biased phylogenetic inferences that converge on the wrong tree (Philippe *et al.* 2005). Another difficulty working with mitochondrial data is high base compositional heterogeneity (Song *et al.* 2010), which describes uneven nucleotide (or for protein data amino-acid) frequencies. Since most models in phylogenetic inference assume stationarity, that is they assume equal transition probabilities and base frequencies at equilibrium, non-stationarity can lead to clusters of species with similar base frequencies, even though unrelated (Song *et al.* 2010; López-López & Vogler 2017). These are the reasons why phylogenies based on mitochondrial genomes are often wrong, or at least in conflict with more congruent analyses based on nuclear or morphological data.

A transcriptome constitutes the entirety of RNA molecules expressed from genes in a cell or tissue (Thompson *et al.* 2016). Gene expression *per se* and the degree to which a gene is expressed varies amongst tissues; it is regulated internally, but this regulation is influenced externally by the environment. This is why most transcriptomes constitute a snapshot in time, rather than a complete set of RNAs. Transcriptomes usually can yield hundreds to a few thousand protein-coding genes, which offers a greater variety of genes to choose from compared to mitogenomes. Methods to assemble transcriptomes *de novo* exist, as do tools to find suitable genes for phylogenetic inference (same bioinformatic effort as for low-coverage genomes). Drawbacks are slightly more difficult RNA extraction protocols compared to those for the more stable DNA. More importantly though, and in contrast to all previously mentioned methods, is the need for special chemicals and cooling devices when specimens are collected in the field, that are especially difficult to bring to isolated and/or tropical regions. This

Chapter 1

excludes all specimens collected and stored in ethanol and demands freshly collected material. However, phylogenetics asks for a representative taxon sampling (Nabhan & Sarkar 2012) and for a group as diverse as Staphylinidae, where some subfamilies are endemic to certain regions of the world, it would be impossible to collect enough species for a balanced taxon sampling.

Whichever strategy one chooses, there are advantages and drawbacks and there might be no prevailing reasons why one method should be preferred over another. In our case, determining variables were time, money, and a collaboration with members of the 1KITE-team that had both successfully mined transcriptomes and done target capture to reconstruct phylogenies (Misof *et al.* 2014; Mayer *et al.* 2016; Young *et al.* 2016). This provided us access to a total of ten transcriptomes for species within the superfamily Staphylinoidea, one published and nine unpublished. These data, together with the *N. vespilloides* genome, were the foundations on which we chose the sequencing type. Transcriptomes give access to nuclear protein-coding genes, which excluded mitochondrial genomes as an option. Sequencing more transcriptomes was impossible, because of the special collection requirements. In order to find suitable genes in the transcriptomes for the phylogenetic reconstruction of Staphylinidae and related families, a database as described earlier had to be generated. With such database new data could either be generated by doing target capture or whole genome re-sequencing of material collected in ethanol. Without an existing bait kit for Coleoptera at the time and the testing phase for the specificity of the baits after bioinformatically intensive bait design process, we decided to add taxa with the whole genome re-sequencing approach.

Retrieving Phylogenetically Informative Protein-Coding Genes from Fragmentary Genomic Data

Phylogenetic reconstruction from molecular markers based on DNA or RNA sequence data requires the use of homologous, in particular orthologous sequences per gene (ideally single-copy, but see Emms & Kelly 2018), meaning that they stem from the ancestral gene of the last common ancestor of the compared species and did not develop within the same lineage, i.e. via gene duplication (Fitch 1970). Only then are genes comparable across species and suitable to infer their phylogenetic relationships. While in theory this is easy to comprehend, in practice it is rather difficult to determine the 'original' copy in a family of genes that have a common origin (*homologs*). In order to be able to distinguish the gene passed on by an ancestral lineage (*ortholog*), from those acquired via gene duplication after the evolutionary split into different species (*paralog*) or via horizontal gene transfer (*xenolog*), well-sequenced data of high quality are required (Koonin 2005; Petersen *et al.* 2017). Existing software tries to disentangle these evolutionary histories of genes either via phylogeny-based or graph-based approaches (Kristensen *et al.* 2011). Phylogeny-based approaches usually use a form of

tree reconciliation method where species-specific gene family trees are mapped onto a species tree (Goodman *et al.* 1979). The rationale to disentangle the relationships between genes is that paralog sequences will group in clades of the same species, in contrast to ortholog sequences that will group in clades with different species. Such trees then enable to detect gene loss or gene duplication events amongst the compared taxa, for example while applying the principle of maximum parsimony, but also maximum likelihood (ML) or Bayesian inference approaches are available (Gabaldón 2008; Kristensen *et al.* 2011). Since orthology relationships are the result of an evolutionary process, a phylogenetic analysis appears to be the most appropriate tool to understand them. However, in reality, depending on the data at hand, tree-based approaches are problematic. In order to reconstruct a correct gene tree, all members of a gene family need to be identified, which means that only proteomes of well-sequenced and annotated genomes should be used. Species prone to hybridization are troublesome as the bifurcating structure of a phylogenetic tree does not allow for horizontal gene transfer (Gabaldón 2008). Additionally, available software have shown to have a high rate of false positives, i.e. a gene is identified as ortholog although it is not (Chen *et al.* 2007). Most concerning is that the majority of algorithms usually assume a known species tree, even though newer algorithms are relaxing this assumption (Yang & Smith 2014). However, if a reliable inference of phylogenetic relationships is the goal, using an approach that requires a known species-tree to find the appropriate data becomes meaningless.

Graph-based approaches on the other hand, use the genome-wide best reciprocal hit (BRH) criterion. Here, the underlying theory says that when sequences of a homologous gene are compared to each other and across pairs of genomes, then two orthologous sequences of the same gene will be more similar across genomes than to sequences of any other gene since they descend from a single ancestral gene (Altenhoff *et al.* 2012). In contrast to phylogeny-based approaches, using the BRH criterion to find orthologs has proven to have a low false positive rate (Chen *et al.* 2007). However, even when applying the BRH criterion, most algorithms assume that gene loss is rare and that the compared gene sets are complete. That problem in fact exists in both phylogeny-based or graph-based approaches. In reality, these assumptions are not valid, because empirical gene loss is generally common (Wyder *et al.* 2007) and for most species there are no full well-sequenced and annotated genomes available. There are various reasons why gene sets may be incomplete: 1) the genes have been lost in the studied organism in an evolutionary process (such statement requires a full genome at hand); 2) the genomes are shallowly sequenced and are incomplete; 3) there are no available genomes for the organism of interest, but only transcriptomes or other NGS data (e.g. from target capture approaches).

In order to circumvent the problem of not being able to infer the genealogical relationships from incomplete data, we first created a gene orthology set based on published genomes. These ge-

Chapter 1

omes are either considered as draft genomes or full genomes, but are all sequenced well enough to be annotated. There is a variety of software available for these purposes and we chose OrthoDB (Zdobnov *et al.* 2017). The software identifies clusters of orthologous sequence groups (i.e. genes) by using the BRH criterion to find the shortest path through the speciation node genes on a distanced-based gene tree (Kriventseva *et al.* 2015). Details about this part of the analysis and the taxon sampling can be found in the Materials and Methods section: 'Orthology prediction'. Orthologs were likewise extracted from transcriptome and low-coverage genome data using this graph-based approach.

Because the already available data were transcriptomic, we designed the database for single-copy protein-coding genes and considered only the proteome of each genome. They will be referred to as 'reference proteomes'. Since orthologous sequence groups (in short: ortholog groups, orthologs or OGs) by definition stem from the last common ancestor of compared species, the taxonomic choice of species for assessing ortholog sets is important. The more closely related the taxa whose genomes are being compared, the more specific the orthologs for this clade will be and vice versa. A wider taxon sampling will hence lead to a selection of more conserved genes.

The final ortholog gene set consists of OGs where each OG contains the protein-coding sequences of all the reference taxa for which an ortholog sequence was identified. In each OG, a reference taxon is only represented by one sequence. Only OGs that occur in single copy in each reference taxon were considered. Once generated, the ortholog gene set was then used to assign sequences from incomplete genetic data, e.g. from transcriptomes or low-coverage sequenced genomes, to ortholog groups. A variety of software has been developed for these purposes; we chose Orthograph (Petersen *et al.* 2017). It is designed especially for incomplete data, e.g. targeting expressed sequence tags (ESTs) similar to the software HaMStR (Ebersberger *et al.* 2009). They both perform forward and reverse searches to assign whether or not sequences are ortholog to a set of provided OGs. In a first step, for each OG (including the sequences of the reference species) a multiple sequence alignment is created and then converted into profile Hidden Markov Models (pHMMs). Contrary to methods that search using the BLAST algorithm (Altschul *et al.* 1990), that only allow pairwise sequence comparisons, a pHMM considers all sequences in an alignment at once and goes through an alignment column by column (Eddy 1998). Based on the frequency of each of the amino-acid in a given column, a score is calculated and saved. This procedure also allows for insertions or deletions. If a new sequence is then compared to the pHMM, based on the previously calculated scores, a probability is calculated, which decides again whether or not the new sequence is homologous to the given alignment or not (Eddy 1998; Compeau & Pevzner 2015). To then determine whether or not a sequence is orthologous, a reciprocal search is performed applying the BRH criterion: the candidate sequence is searched with BLASTP (Ebersberger *et al.* 2009) against the reference proteomes.

The sequence in question will only be accepted as being orthologous to the set of sequences of the reference species (OGs) if at least one protein-coding sequence of a reference taxon that contributed to the pHMM provides the best BLASTP hit. One difference between HaMStR and Orthograph is that HaMStR does not check whether or not a candidate transcript might have already been assigned to another OG (therefore creating more than one assignment of the same transcript). This is problematic since it allows genes to be co-ortholog and ideally, such co-orthologs should be removed from the dataset. Instead, Orthograph assigns the sequence in question with a positive reciprocal match to the overall best matching OG. In addition, it takes into account all reference species at once by creating one database, contrary to HaMStR that considers each reference species separately. Once all sequences went through the forward and reverse searches, they are ranked according to their scores in both the forward and reverse searches in descending order. Beginning from the highest scores of the forward search, a sequence will be considered as ortholog and assigned to the OG, if the best reciprocal BLASTP hit of the reverse search stems from a sequence in the same OG the pHMM search was based on. If this is fulfilled, no other sequence transcript with a lower score will be considered for that OG anymore (Petersen *et al.* 2017), unless they are not overlapping. In case two transcripts do not overlap, they may be assigned to one OG and are concatenated; an option that can also be turned off. This ensures that only the best matching sequence with the highest score is selected as ortholog. Details about the use of Orthograph and the genomes used in this study can be found in the Material and Methods section 'Orthology prediction'.

Amongst the many available orthology inference tools, HaMStR is the most widely used (Yang & Smith 2014). Orthograph is a newer software that has not been used extensively compared to HaMStR to make such statements, but has already been used to assign orthologs in transcriptomes for a comprehensive phylogenetic reconstruction of hymenoptera (Peters *et al.* 2017) and to generate probe sets for target DNA enrichment subsequently used in conjunction with transcriptomic data in hymenoptera phylogenetics (Mayer *et al.* 2016; Bank *et al.* 2017). When the work on this project began three years ago, there was no comparable study published on Coleoptera. Recently, two studies have been published with mixed datasets, but both focusing on members of click beetles (Elateridae). The authors used nuclear protein-coding genes extracted from transcriptomes in addition to one and three genomes respectively (Kusy *et al.* 2018a,b). With a much wider taxon sampling, the i5K initiative also just published a phylogenetic analysis on arthropod evolution using protein-coding genes of 76 sequenced and annotated genomes, including six beetles (Thomas *et al.* 2018). These genomes were used here to generate a new ortholog set covering Coleoptera. We are the first to mine transcriptomes and low-coverage genomes for suitable markers to infer the phylogenetic relationships of the mega-diverserove beetles, even though it might be a risky endeavor.

Chapter 1

Material and Methods

Newly generated genomic data

Collection of specimens

Specimens intended for whole-genome re-sequencing (WGS) belonged to 16 species. They were collected in various parts of the world between the years 2012 and 2018 and dropped alive in the absolute ethanol. They were kept there for days or weeks under normal field conditions preferably at least in the shade. After they were brought to the Natural History Museum of Denmark, they were stored in the freezer at -20°C until further processing, each species in their own cryovial, possibly with other conspecifics if several specimens had been collected at the same site. Information about collection time, localities, collectors and number of specimens per species can be found in Table 1.

DNA extraction

DNA extractions were carried out at the Centre for GeoGenetics, Natural History Museum of Denmark, where pre- and post-PCR work was physically separated from each other to minimize cross-contamination. Total DNA was extracted from specimens, of which head, prothorax, and abdomen were carefully disarticulated from each other to enable the digestion buffer to reach more soft tissues and yield higher DNA amounts. All tools used for the separation of specimens were sterilized with absolute ethanol between processing different species. For most specimens, DNA was extracted using the DNeasy® Blood and Tissue kit (Quiagen, Valencia, CA), according to the manufacturer's protocol. For two species, one of the genus *Gymnusa* Gravenhorst, 1806 and another of the genus *Diochus* Erichson, 1839, an isopropanol precipitation protocol was used as described in Gilbert *et al.* (2007) and Thomsen *et al.* (2009), but omitting the phenol:phenol:chloroform purification step before the isopropanol precipitation. For all samples, DNA quantity was first measured using a Qubit® 2.0 Fluorometer with the Qubit dsDNA BR Assay kit (Thermo Fisher Scientific, Waltham, MA) and the quality was assessed using the 4200 TapeStation System (Agilent Technologies, Inc., Santa Clara, CA). Extracts of the latter two samples needed to be purified to remove tiny tissue particles before they could be processed further. Purification was carried out at the laboratories of Novogene Corporation Limited. In order to have sufficient amounts of total DNA (a minimum 0.6 μg of total DNA), sometimes conspecific specimens needed to be pooled prior to DNA extraction. The number of specimens per species used for extractions is listed in Table 1. The remaining beetle parts (what has not been lysed by the buffer) are stored in the freezer collection of the Natural History Museum, Denmark (ZMUK) at -20°C .

Table 1. Detailed list of species chosen for whole genome re-sequencing (WGS). Each species is listed including their taxonomic rank, collection year, country it was collected in, and the number of specimens used for DNA extractions.

Superfamily	Family	Subfamily	Tribe	Species	Author, year	Year collected	Locality	# specimens
Staphylinoidea	Staphylinidae	Aleocharinae	Athetini	gen. sp.	Casey, 1910	2018	Denmark	12
Staphylinoidea	Staphylinidae	Aleocharinae	Gymnusi	<i>Gymnusa</i> sp.	Gravenhorst, 1806	2015	Russia	12
Staphylinoidea	Staphylinidae	Omalinae	Anthrophagini	<i>Lesteva longoelytrata</i>	Goeze, 1777	2018	Denmark	13
Staphylinoidea	Staphylinidae	Omalinae	Omalini	<i>Acidota crenata</i>	(Fabricius, 1793)	2018	Denmark	6
Staphylinoidea	Staphylinidae	Oxytelinae	Deleasterini	<i>Deleaster dichrous</i>	(Gravenhorst, 1802)	2016	Italy	7
Staphylinoidea	Staphylinidae	Oxytelinae	Oxytelini	<i>Anotylus rugosus</i>	(Fabricius, 1775)	2018	Denmark	6
Staphylinoidea	Staphylinidae	Paederinae	Lathrobiini	<i>Lathrobium brunnipes</i>	(Fabricius, 1793)	2018	Denmark	2
Staphylinoidea	Staphylinidae	Paederinae	Paederini	<i>Paederus littoralis</i>	Gravenhorst, 1802	2018	Denmark	5
Staphylinoidea	Staphylinidae	Staphylininae	Othiini	<i>Othius punctulatus</i>	Goeze, 1777	2017	Sweden	4
Staphylinoidea	Staphylinidae	Staphylininae	Diachini	<i>Diachus</i> sp.	Erichson, 1839	2017	Queensland (AU)	9
Staphylinoidea	Staphylinidae	Staphylininae	Staphylinini	<i>Quedius fuliginosus</i>	Gravenhorst, 1802	2018	Denmark	1
Staphylinoidea	Staphylinidae	Staphylininae	Staphylinini	<i>Philonthus decorus</i>	Gravenhorst, 1802	2017	Denmark	3
Staphylinoidea	Staphylinidae	Steninae	-	<i>Stenus bimaculatus</i>	Gyllenhal, 1810	2018	Denmark	5
Staphylinoidea	Staphylinidae	Tachyporinae	Mycetoporini	<i>Lordithon lunulatus</i>	Linnaeus, 1761	2014	Denmark	4
Staphylinoidea	Staphylinidae	Tachyporinae	Tachyporini	<i>Tachinus rufipes</i>	(Linnaeus, 1758)	2017	Denmark	1
Staphylinoidea	Silphidae	Silphinae	Silphini	<i>Silpha</i> sp.	Linnaeus, 1758	2015	Russia	1

Library preparation & genome sequencing

Library preparation, generation of low-coverage WGS data and quality control were carried out by Novogene Corporation Limited (Beijing, China). A total amount of 0.6 µg DNA per sample was used as input material for the library preparation. Sequencing libraries were generated using NEBNext Ultra DNA Library Prep Kit for Illumina (New Eng-

land Biolabs, Beijing, China) following manufacturer's recommendations and indices were added to each sample. The genomic DNA was randomly fragmented to a size of 500 bp (350 bp for *Diachus* sp.) by shearing (AFA Process with Covaris Focused-ultrasonicator). Then, DNA fragments were end polished, A-tailed, and ligated with the NEBNext Universal PCR primers for Illumina (New England Biolabs, Beijing, China) sequencing, and further PCR enriched with P5 and indexed P7 oligos. The PCR products were purified using AMPure XP (Beckman Coulter, Shanghai, China). Resulted libraries were analyzed for size distribution with the Agilent 2100 Bioanalyzer and quantified using real-time PCR on an ABI Veriti Thermal Cycler (Applied Biosystems). Sequencing was performed on an Illumina HiSeq X Ten Platform. In a first trial run only two species were sequenced: *Diachus* sp. was sequenced with a shorter insert size of 350 bp and *Gymnusa* sp. with an insert size of 500 bp. Although two different species cannot directly be compared, FastQC v. 0.11.7 (Andrews 2010) checks and summary statis-

Chapter 1

tics of an initial assembly with untrimmed reads using Quast (Gurevich *et al.* 2013) led to superior metrics for the specimen sequenced with larger insert size (e.g. more longer contigs, higher N50). A summary of FastQC and Quast results for the quick assemblies of the two genomes can be found in Supplementary Material, Table S1). Therefore, we chose to sequence all remaining samples with an insert size of 500 bp. All specimens were sequenced to 6 Gigabases (Gb) raw data. Post sequencing, all reads containing adapters, reads containing N > 10% ('N' represents the base that cannot be determined) and reads containing a low quality base ($Q_{\text{score}} \leq 5$) which is over 50% of the total bases were removed by Novogene Corporation Limited.

Genome assembly & quality assessment

Raw sequence reads were checked prior to and after trimming the reads with FastQC v. 0.11.7 (Andrews 2010). K-mers were analyzed with Jellyfish v. 2.2.7 (Marçais & Kingsford 2011) and perl scripts by Joseph Ryan (available at http://josephryan.github.io/estimate_genome_size.pl/) with a k-mer size set to $k = 31$, counting both strands and using a hash size of 10 billion and an upper count limit of 500. While Jellyfish counts the k-mers and generates a histogram, the perl scripts help to find the first peak in the k-mer plot, which is used to estimate the genome size and coverage. The genome size of each species was estimated whenever possible. Additionally, k-mer plots were also used to estimate the genome size using the online platform GenomeScope (Vurture *et al.* 2017) with a k-mer length of $k = 31$, read length = 150 bp and max k-mer coverage = 1000. Trimmomatic v. 0.36 (Bolger *et al.* 2014) was used to trim the raw reads at their beginnings and ends and to clean raw reads from potential leftover adapter contamination. The first 5 bases at the beginning of each read were always removed as well as all subsequent bases below a quality threshold of 10. Bases at the end of each read were removed if < 5 . Furthermore, every read was scanned with a sliding window of four bases and cut whenever the average quality per base dropped below 20. The minimum read length was set to 40 bases, thus shorter reads were removed.

Next, the trimmed raw reads were assembled using SparseAssembler (Ye *et al.* 2012) under default settings except that the k-mer size was set to 31, a value smaller than the minimum fragment length chosen in Trimmomatic and allowing 10 intermediate k-mers to be skipped. There is evidence that this assembler apparently performs better with low-coverage genome sequencing data than other assemblers, leading to longer continuous contigs (A. Predeus, unpublished). Since the genomes were sequenced only to low coverage, and in many cases stem from several pooled specimens, a high level of heterozygosity was expected. Heterozygous regions are usually assembled in separate contigs when assembling short reads, which leads to an increased and erroneous number of gene copies and

additionally overestimates the genome size. The software Redundans (Pryszcz & Gabaldón 2016) was therefore used to remove alternative heterozygous contigs from the assemblies, to further scaffold the reduced scaffolds and to close gaps. Final assemblies were assessed with Quast (Gurevich *et al.* 2013).

The number of genes and their properties in newly assembled genomes were assessed using BUSCO software v. 3.0.0 (Simão *et al.* 2015) in conjunction with HMMER v 3.1 (Eddy 2011), Blast+ (Camacho *et al.* 2009) and Augustus (Keller *et al.* 2011) using the ‘endopterygota’ database (odb9) with 2,442 OGs.

Transcriptomic data

Retrieval of transcriptome sequence data

The raw reads of the following five species have kindly been provided by members of the 1KITE project: one undetermined species of the genus *Acrotrichis* Motschulsky, 1848, *Aleochara curtula* (Goeze, 1777), *Hydraena subimpressa* Rey, 1885, *Ocypus brunnipes* (Fabricius, 1781), and *Oiceoptoma thoracicum* (Linnaeus, 1758) (Misof *et al.* 2014). Furthermore, the raw reads of five more unpublished species were kindly provided by and in agreement with the Australian National Insect Collection (CSIRO), TransANIC project: *Anisotoma castanea* (Herbst, 1791), *Paederus cruenticollis* Germar, 1848, *Paragyrtonodes modestus* Szymczakowski, 1966, one undetermined species of the genus *Philagarica* Deane, 1930, and one undetermined species of the genus *Silphotelus* Broun, 1895. A detailed list of the taxon sampling can be found in Table 2.

Table 2 Detailed list of species, for which their raw transcript reads were obtained by institutions and people listed below.

Superfamily	Family	Subfamily	Tribe	Species	Author, year	ID	Data provided by
Staphylinoidea	Hydraenidae	Hydraeninae	Hydraenini	<i>Hydraena subimpressa</i>	Rey, 1885	INShkeTAFRAAPEI-90	1KITE
Staphylinoidea	Leiodidae	Leiodinae	Agathidiini	<i>Anisotoma castanea</i>	Herbst, 1791	CWHPE16090185	CSIRO
Staphylinoidea	Leiodidae	Camiarinae	Camiarini	<i>Paragyrtonodes modestus</i>	Szymczakowski, 1966	CWHPE16110001	CSIRO
Staphylinoidea	Ptiliidae	Acrotrichinae	Acrotrichini	<i>Acrotrichis</i> sp.	Motschulsky, 1848	INShkeTAIRAAPEI-95	1KITE
Staphylinoidea	Ptiliidae	Ptiliinae	Nanosellini	<i>Philagarica</i> sp.	Deane, 1930	CWHPE16110015	CSIRO
Staphylinoidea	Silphidae	Nicrophorinae	Nicrophorini	<i>Nicrophorus vespilloides</i>	Herbst, 1783	GSE72225	CB Cunningham & AJ Moore
Staphylinoidea	Silphidae	Silphinae	Silphini	<i>Oiceoptoma thoracica</i>	Linnaeus, 1758	RINSinITCDRAAPEI-61	1KITE
Staphylinoidea	Staphylinidae	Aleocharinae	Aleocharini	<i>Aleochara curtula</i>	Goeze, 1777	INShauTBERAAPEI-33	1KITE
Staphylinoidea	Staphylinidae	Staphylininae	Staphylinini	<i>Ocypus brunnipes</i>	Fabricius, 1781	INShkeTCMRAAPEI-45	1KITE
Staphylinoidea	Staphylinidae	Paederinae	Paederini	<i>Paederus cruenticollis</i>	Germar, 1848	CWHPE16090138	CSIRO
Staphylinoidea	Staphylinidae	Proteininae	Silphotelini	<i>Silphotelus</i> sp.	Broun, 1895	CWHPE16090163	CSIRO

Chapter 1

De novo transcriptome assembly & quality assessment

Sequencing adaptors and low-quality regions were removed using Trimmomatic v. 0.3 (Bolger *et al.* 2014) (using settings ILLUMINA-CLIP: TruSeq-3-PE.fa :2:30:10:2:True LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). The quality of prepared paired-end read data was manually assessed using FastQC v. 0.11.7. De novo assembly for comparative purposes was performed using three different assemblers with different settings: TransABYSS v. 1.5.5 (Robertson *et al.* 2010) (k-mer size of 32, 48, and 64), SOAPdenovo-Trans v. 1.04 (Xie *et al.* 2014) (k-mer size of 33, 49, and 65) and Trinity v. 2.3.2 (Grabherr *et al.* 2011). As for the newly assembled genomes, BUSCO software v. 3.0.0 (e-value: 1e-3) with the 'endopterygota' dataset (odb9 with 2,442 OGs) were used to verify the presence and completeness of orthologs in assembled transcriptomes. The assembly quality evaluation of individual contigs was carried out using TransRate v. 1.0.1 (Smith-Unna *et al.* 2016).

Retrieval of remaining data

Short read archives (SRAs) of three low-coverage aleocharine genomes *Dalotia coriaria* (Kraatz, 1856), *Deinopsis erosa* (Stephens, 1832) and *Mimaenictus wilsoni* Kistner & Jacobson, 1975 were downloaded from NCBI using sratoolkit v. 2.8.1 (SRA Toolkit Development Team, <http://ncbi.github.io/sra-tools/>) with the 'prefetch' option. These genomes were sequenced and submitted to NCBI by Joe Parker (California Institute of Technology, California, USA). An overview of the three species used is provided in Table 3. SRA files were converted to fastq files using 'fastq-dump --split3'. The downloaded raw reads were inspected with FastQC v. 0.11.7. Genome size and coverage were estimated using Jellyfish v. 2.2.7 and perl scripts by Joseph Ryan (Whitney Laboratory for Marine Bioscience, Florida, USA), as well as GenomeScope with the same settings as for the newly-sequenced genomes (see above). K-mer plots were generated with Jellyfish. Next, the reads were trimmed with Trimmomatic using the same settings as for the newly sequenced genomes, and results were again checked in FastQC. The Platanus Genome Assembler v. 1.2.4 (Kajitani *et al.* 2014)

Table 3. Detailed list of three aleocharine species (Staphylinoidea: Staphylinidae), for which their raw sequence reads were obtained via whole genome re-sequencing; downloaded from GenBank.

Voucher #	Subfamily	Tribe	Species	Author, year	Locality	Year collected	Collector Identifier	Bioproject accession
CI110	Aleocharinae	Aenictoteratini	<i>Mimaenictus wilsoni</i>	Kistner & Jacobson, 1975	USA	2016	Joe Parker	PRJNA397438
BHU26	Aleocharinae	Athetini	<i>Dalotia coriaria</i>	(Kraatz, 1856)	USA	2016	Joe Parker	PRJNA342849
BWP27	Aleocharinae	Deinopsini	<i>Deinopsis erosa</i>	(Stephens, 1832)	unknown	2016	Joe Parker	PRJNA361286

was used to assemble the trimmed reads. Platanus was chosen because it is specifically made for the *de novo* assembly of short DNA reads from heterozygous diploids obtained via shotgun sequencing. The software first assembles reads to contigs, testing differing k-mer sizes. Heterozygous regions in the assemblies were identified and then removed using Redundans as described for the newly sequenced genomes (see above). Finally, Quast was used to obtain comparative metrics (e.g. N50, N75, GC-content, number of Ns per 100 bp) of the assembled scaffolds.

In order to increase the taxon sampling, nucleotide sequences of 95 nuclear protein-coding genes published by (Zhang *et al.* 2018) were downloaded from NCBI and added to the genomic data set. This incorporated representatives of five additional subfamilies of Staphylinidae to the existing dataset and also increased the taxon sampling of already sampled in- and outgroup members from 33 to 57 taxa. In the following, the dataset of Zhang *et al.* (2018) will be called ‘primer-based’ to distinguish it from the genomic and transcriptomic data. The latter will be referred to as ‘omic-only’. We included the 95 genes in our orthology assignment using our ortholog set (see Material and Methods section ‘Orthology prediction’). An overview of the taxa included in the primer-based dataset and used in this study can be found in Table 4.

Table 4. Detailed list of species, for which 95 nuclear-protein coding genes were downloaded and 40 genes were added to the primer-based dataset after orthology assignment.

Voucher No	Superfamily	Family	Subfamily	Tribe	Species	Author, year	Locality
INB201	Staphylinoidea	Agyrtidae	Pterolomatinae	-	<i>Pteroloma forstromii</i>	Gyllenhal, 1810	China, Jilin prov., Changbaishan
INB108	Staphylinoidea	Hydraenidae	Hydraeninae	Hydraenini	<i>Hydraena</i> sp.	Kugelann, 1794	China, Hongkong
CSR026	Derodontoidea	Jacobsoniidae	-	-	<i>Derolathrus</i> sp.	Sharp, 1908	Australia, Queensland, Garradunga
CSR109	Derodontoidea	Jacobsoniidae	-	-	<i>Derolathrus</i> sp.	Sharp, 1908	Australia, New South Wales
CSR030	Staphylinoidea	Leiodidae	Camariinae	Agyrttonini	<i>Agyrtodes</i> sp.	Portevin, 1907	Australia, Canberra, Black mountain
CSR032	Staphylinoidea	Leiodidae	Leiodinae	Agathidiini	<i>Agathidium</i> sp.	Panzer, 1797	Australia, Queensland, Garradunga
CSR038	Derodontoidea	Nosodendridae	-	-	<i>Nosodendron</i> sp.	Latreille, 1804	Australia, Queensland, Garradunga
CSR044	Staphylinoidea	Ptiliidae	-	-	gen. sp.	Heer, 1843	Australia, Queensland, Garradunga
INB022	Staphylinoidea	Silphidae	Nicrophorinae	Nicrophorini	<i>Nicrophorus nepalensis</i>	Hope, 1831	China, Guizhou prov., Leigongshan
INB078	Staphylinoidea	Silphidae	Silphinae	Necrodini	<i>Necrodes littoralis</i>	(Linnaeus, 1758)	China, Sichuan prov., Hailuogou
INB085	Staphylinoidea	Staphylinidae	Apateticinae	-	<i>Apatetica</i> sp.	Westwood, 1848	China, Yunnan prov., Baihualing
INB176	Staphylinoidea	Staphylinidae	Omaliinae	-	gen. sp.	MacLeay, 1825	China, Gansu prov., Guan'ergou
INB086	Staphylinoidea	Staphylinidae	Osoriinae	Leptochirini	<i>Priochirus</i> sp.	Sharp, 1887	China, Sichuan prov., Emeishan
INB087	Staphylinoidea	Staphylinidae	Osoriinae	Osoriini	<i>Osorius</i> sp.	Guérin-Méneville, 1829	China, Sichuan prov., Emeishan
INB030	Staphylinoidea	Staphylinidae	Paederinae	Paederini	<i>Megalopaederus</i> sp.	Scheerpeltz, 1957	China, Sichuan prov., Wolong
INB091	Staphylinoidea	Staphylinidae	Pselaphinae	Tyrini	<i>Centrophthalmus</i> sp.	Schmidt-Göbel, 1838	China, Guangdong prov., Heishiding
CSR054	Staphylinoidea	Staphylinidae	Scaphidiinae	Scaphidiini	<i>Scaphidium</i> sp.	Olivier, 1790	Australia, Queensland
INB089	Staphylinoidea	Staphylinidae	Scaphidiinae	Scaphidiini	<i>Scaphidium</i> sp.	Olivier, 1790	China, Yunnan prov., Yulongshan
CSR146	Staphylinoidea	Staphylinidae	Scydmaeninae	-	gen. sp.	Leach, 1815	Australia, Queensland
INB008	Staphylinoidea	Staphylinidae	Staphylininae	Staphylinini	<i>Staphylinus</i> sp.	(Linnaeus, 1758)	China, Guangdong prov., Dadongshan
INB174	Staphylinoidea	Staphylinidae	Staphylininae	Staphylinini	gen. sp.	Latreille, 1802	China, Guizhou prov., Leigongshan
INB175	Staphylinoidea	Staphylinidae	Steninae	-	<i>Dianous</i> sp.	Leach, 1819	China, Hubei prov., Dashennongjia
INB090	Staphylinoidea	Staphylinidae	Tachyporinae	Tachyporini	<i>Tachinus</i> sp.	Gravenhorst, 1802	China, Sichuan prov., Wawushan

Publication DOI: 10.1038/s41467-017-02644-4

Chapter 1

Orthology prediction

Obtaining the ortholog gene set

The first step towards being able to find single-copy orthologs in the primer-based, as well as omic-only data of the target species, was creating an ortholog gene set of nuclear protein-coding genes from a range of reference genomes. Those reference genomes are characterized by being formally assembled, annotated and published. The ortholog gene set was created using the online platform OrthoDB9 (Zdobnov *et al.* 2017), a comprehensive hierarchical catalog of single copy orthologs, using the following ten reference species: *Anoplophora glabripennis* (Motschulsky, 1854) (McKenna *et al.* 2016); *Agrilus planipennis* Fairmaire, 1888 (i5K Consortium 2013); *Leptinotarsa decemlineata* (Say, 1824) (i5K Consortium 2013); *Onthophagus taurus* (Schreber, 1759) (i5K Consortium 2013); *Orussus abietinus* (Scopoli, 1763) (i5K Consortium 2013) (official gene sets of these five species were downloaded from BCM v. 0.5.3 (Poelchau *et al.* 2015); *Bombyx mori* (Linnaeus, 1758) (official gene set downloaded from <http://silkworm.big.ac.cn/jsp/download.jsp>; Wang 2004); *Dendroctonus ponderosae* Hopkins, 1902 (genome downloaded from http://metazoa.ensembl.org/Dendroctonus_ponderosae/Info/Index; Keeling *et al.* 2013); *Drosophila melanogaster* Meigen, 1830 (v. R6.7, downloaded from Flybase.org; Hoskins *et al.* 2015); *Tribolium castaneum* (Herbst, 1797) (v. 5.2 downloaded from <http://ibeetle-base.uni-goettingen.de/help/resources>; Richards *et al.* 2008) and *Nicrophorus vespilloides* (v. 1.0 downloaded from <https://www.ncbi.nlm.nih.gov>; Cunningham *et al.* 2015). More information about the reference species can be found in Table 5. Since *N. vespilloides* was not available in OrthoDB, we used a 3-step approach to add it as a reference species in the ortholog set: First, if more than one isoform for a given gene was present, then only the longest isoform was selected for downstream analysis as it is done for genomes integrated in OrthoDB. Then, the predicted protein sequences were mapped onto the other reference genomes (see above) present in the OrthoDB database for genes that have been found to occur in single-copy on the Endopterygota hierarchical-level in our reference species. They were saved in a table with EOG identifiers and respective sequence headers of the genomic data. This resulted in a file containing the regions successfully mapped, with both the identifier used in the *N. vespilloides* genome annotation, as well as the EOG (Eukaryotic Ortholog Group) identifier used in OrthoDB. OGs that occurred in single copy in all reference species, but with multiple occurrences in *N. vespilloides* were discarded due to putative paralogy. Finally, every OG had to be presented in at least three reference species, else, the OG was excluded. The final ortholog set consists of 3,822 OGs used for downstream analyses. An overview of OGs for each reference species is provided in Table 6. A detailed table listing the species and gene IDs (individual species IDs and EOG IDs) can be viewed in

the online deposited Supplementary Material
(<https://drive.google.com/open?id=1PDA7OnBJieoEMZWYXaA-n5-5-TuHbCHf>: Table S2).

Even though the ortholog set was specifically generated to include genes from *N. vespilloides*, which is the closest reference genome to all Staphylinidae and the only draft genome within Staphylinidae, it needs to be mentioned that we missed that the identified genes were actually not present within the OG fasta files. Hence, these sequences had an influence on the 3,822 OGs that were chosen, but they were not used in the following steps for the identification of single-copy orthologs in the target species or at any other point in the analysis. We are currently working on a solution before publishing this study, starting from Orthograph, once the *N. vespilloides* sequences have been added.

Table 5. Detailed list of species, for which their reference genomes were downloaded to generate the OG sets. Information provided: the species' classification, their common name, gene version, number (#) of available genes, the download source including link and DOI to the corresponding publication.

#	Superfamily	Family	Subfamily	Tribe	Species	Common name	Author, year
1	Elateriformia	Buprestidae	Agrilinae	-	<i>Agrilus planipennis</i>	Emerald ash borer	Fairmaire, 1888
2	Bombycoidea	Bombycidae	Bombycinae	-	<i>Bombyx mori</i>	Silk Moth SilkDB	(Linnaeus, 1758)
3	Curculionioidea	Curculionidae	Scolytinae	Hylesinini	<i>Dendroctonus ponderosae</i>	Mountain pine beetle	Hopkins, 1902
4	Ephydroidea	Drosophilidae	Drosophilinae	-	<i>Drosophila melanogaster</i>	Fruit Fly FlyBase	Meigen, 1830
5	Chrysomeloidea	Cerambycidae	Lamiinae	Lamiini	<i>Anoplophora glabripennis</i>	Asian long-horned b	Motschulsky, 1854
6	Chrysomeloidea	Chrysomelidae	Chrysomelinae	Chrysomelini	<i>Leptinotarsa decemlineata</i>	Colorado potato bee	Say, 1824
7	Staphylinioidea	Silphidae	Nicrophorinae	Nicrophorini	<i>Nicrophorus vespilloides</i>	Burying beetle	Herbst, 1783
8	Scarabaeoidea	Scarabaeidae	Scarabaeinae	-	<i>Onthophagus taurus</i>	Bull-headed dung be	(Schreber, 1759)
9	Orussoidea	Orussidae	-	-	<i>Orussus abietinus</i>	Parasitic wood wasp	(Scapoli, 1763)
10	Cucujiformia	Tenebrionidae	Tenebrioninae	Triboliini	<i>Tribolium castaneum</i>	Red Flour Beetle	(Herbst, 1797)

#	Gene set version	# genes	Download source	Publication DOI	Link to data
1	BCM_version_0.5.3	15,497	i5K ftp server	10.1093/nar/gku983	https://i5k.nal.usda.gov/content/data-downloads
2	BGI version	14,623	SilkDB	10.1093/nar/gki116	http://silkworm.bgi.ac.cn/jsp/download.jsp
3	DendPond_male_1.0	13,088	Ensembl Metazoa	10.1186/gb-2013-14-3-r27	http://metazoa.ensembl.org/Dendroctonus_ponderosae/info/Index
4	R6.07	13,919	Fly Base	10.1093/nar/gku983	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.07_FB2015_04/
5	BCM_version_0.5.3; primary gene set	22,035	i5K ftp server	10.1186/s13059-016-1088-8	https://i5k.nal.usda.gov/content/data-downloads
6	BCM_version_0.5.3; primary gene set	24,671	i5K ftp server	10.1093/nar/gku983	https://i5k.nal.usda.gov/content/data-downloads
7	Nicve_v1.0_ncbi.nih.gov	12,585	NCBI	10.1093/gbe/evv194	https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=110193&lvl=3&lin=f&keep=1&srchmode=1&unlock
8	BCM_version_0.5.3	17,483	i5K ftp server	10.1093/nar/gku983	https://i5k.nal.usda.gov/content/data-downloads
9	BCM_version_0.5.3	10,959	i5K ftp server	10.1093/nar/gku983	https://i5k.nal.usda.gov/content/data-downloads
10	Tcas5.2	16,631	iBeetle Base	10.1038/nature06784	http://ibeetle-base.uni-goettingen.de/help/resources

Table 6. Number of OGs identified for each reference species. The final ortholog set consists of 3,822 OGs in total.

Species	# OGs
<i>Agrilus planipennis</i>	3,327
<i>Anoplophora glabripennis</i>	3,565
<i>Bombyx mori</i>	2,811
<i>Dendroctonus ponderosae</i>	3,169
<i>Drosophila melanogaster</i>	2,067
<i>Leptinotarsa decemlineata</i>	3,448
<i>Nicrophorus vespilloides</i>	3,790
<i>Onthophagus taurus</i>	3,394
<i>Orussus abietinus</i>	2,907
<i>Tribolium castaneum</i>	3,819

Identification of single-copy protein-coding orthologs

We used the Orthograph package v. 0.6.2 (Petersen *et al.* 2017) to build an SQLite (an embedded SQL database engine, see: <https://www.sqlite.org/index.html>) with the reference genomes and the EOG table from the previous step to assign single-copy orthologs from all 19 low-coverage genomes (16 newly sequenced, 3 from the Parker Lab), 11 transcriptome assemblies and 24 addition-

Chapter 1

al species with 95 genes ('primer-based' dataset). For simplicity, those genomes, transcriptomes and single genes will collectively be referred to as 'target sequences' and the organisms they belong to are hence the 'target species'. The ortholog set with all 3,822 OGs formed the basis of the Orthograph pipeline, which depends on a variety of other software. In a nutshell, Orthograph uses a graph-based approach by applying the BRH criterion and a subsequent reciprocal BLAST search against the complete official gene set of all reference species to identify single-copy orthologs of the target species. A more detailed description of Orthograph's functionality can be found in the earlier section 'Retrieving Phylogenetically Informative Protein-Coding Genes from Fragmentary Genomic Data'. Unless otherwise stated, Orthograph was run using default settings.

To create pHMMs for each OG, sequences within an ortholog group (OG) were aligned using the L-INS-i option in MAFFT v. 7.402 (Kato & Standley 2013). Subsequently, the software HMMER v. 3.1b2 (Eddy 2011) was used to generate the required pHMMs to search for candidate orthologs in the target sequences on the translational level (step 1). All candidate orthologs, the pHMMs they best mapped to, and the corresponding bit score are stored in descending order in a separate file. The bit score is an indicator how well the target sequence mapped to a pHMM; the higher the bit score, the better the fit. Step one was done using default settings in Orthograph.

For the reciprocal search, the sequences of candidate orthologs become the query in a protein BLAST (BLASTP) search against the full official gene sets of all reference species on a translational level (step 2). This was done with the help of NCBI BLAST v. 2.6.0+ (Camacho *et al.* 2009). The OG to which the reference sequence belongs to, against which the candidate ortholog revealed the best hit, was then saved in the list created in step one. In the final step, the list with the results from step one and two is checked starting from the top (the candidate ortholog with the highest bit score from step one). A candidate ortholog is kept if the sequence with the best BLAST hit (step 2) is part of the OG that was used to generate the pHMM with the highest bit score (step 1), i.e. the BRH criterion is fulfilled and the new orthologous sequence is added to the respective OG it mapped best to. When the BRH criterion was rejected, the second best hit score for the same candidate ortholog was considered (in case it matched to more than one pHMM), then the third and so on. In this phase, we changed 'max-blast-searches' to 50 instead of 100 (default), meaning that maximally the first 50 pHMM search hits were considered for each candidate ortholog. The maximum number of BLAST hits, 'blast-max-hits', was also changed to 50 (default = 100). Once a candidate sequence was added to an OG (BRH fulfilled), no further sequences of the same organism were added, even if the BRH criterion would be fulfilled again. This ensures that within each OG a species is only represented once. Overall, new sequences were added to 3,812 OGs (initially 3,822 OGs). Of those sequences, Exonerate v. 2.4.0 (Slater & Birney 2005) inferred the open reading frames (ORF) by calculating a pairwise alignment with the reference sequence that scored the best hit in the reciprocal search. We set the option 'ex-

tend-orf' = 1: if possible, it will extend the ORFs beyond the pHMM alignment region, but keeping the orthologous region intact. If not specified otherwise, the extended region must at least have an overlap of 50% of the original pHMM alignment region of the reference species. The software also takes the original nucleotides and then corrects frameshift errors and handles stop codons. While terminal stop codons were removed, internal stop codons were exchanged to 'X' on amino-acid -and 'NNN' on nucleotide level.

When summarizing Orthograph results (the software creates final OG files in FASTA format with all found orthologous sequences, amino-acid and nucleotide level with corresponding sequences), the user can specify whether or not any sequences of the reference species included in the ortholog set should be kept (default: sequences of all taxa, reference, as well as target species remain). We removed sequences of all reference species except for the beetles *T. castaneum*, *A. planipennis* and *O. taurus*. Moreover, the selenocysteine symbol 'U' was replaced with 'X' on amino-acid level and 'NNN' on nucleotide level respectively, since 'U' causes many problems in other downstream software.

Data deposition

Data from the 1KITE beetle consortium and CSIRO will be publicly available in NCBI, GenBank, depending on the timeline of their specific projects. The genomic data of the 16 newly generated low-coverage genomes will be deposited in NCBI, GenBank prior to publication. The corresponding DNA voucher specimens have been deposited at the frozen tissues collection of the Natural History Museum of Denmark in Copenhagen (NHMD).

Phylogenetic analysis

Data arrangement

For this study we designed a variety of different datasets. Taxa across all Staphylinoidea (plus outgroups) are represented by published and unpublished genomes of varying quality, transcriptomes, and the 95 genes shotgun sequenced using designed primers and PCR amplification. Therefore, we analyzed the data not only combined, but also separately. We generated two supermatrices: one comprising species derived from genome and transcriptome data, and a second one with an additional 40 genes from the originally 95 genes of Zhang *et al.* (2018) that had passed the orthology prediction step. The two supermatrices were analyzed on the translational (amino-acid) level. Unless otherwise stated, all genes were processed in the same pipeline and the data were concatenated into the supermatrices in the final step before the phylogenetic analysis.

Chapter 1

Supermatrix 1: gene data blocks 'omic-only'

The data blocks are constituted by the final set of 993 protein-coding orthologs identified in three reference genomes, nineteen low-coverage genomes and eleven transcriptomes. It will be referred to as 'omic-only'.

Supermatrix 2: gene data blocks 'omic & primer-based'

This is an extension of supermatrix 1, to which data from Zhang *et al.* (2018) was added (40 assigned orthologs out of 95 genes). It extends the taxon sampling, however, the number of genes added for each new species is about a thirtyfold less (maximum 40 genes, compared to maximum 993). It will be referred to as 'omic & primer-based'.

Inference and masking of multiple sequence alignments

First, multiple sequence alignments (MSAs) for each OG were generated with the L-INS-i algorithm implemented in MAFFT v. 7.402 using the translational (amino-acid) level. To identify ambiguous or randomly similar aligned regions within each MSA, we used the software Aliscore v. 2.2 (Misof & Misof 2009) under default settings, except setting -r to one octillion (so high that all or close to all possible non-overlapping pairs will be compared) and using the -e option, so that the software takes the data type (amino-acids) into account. This step was carried out because ambiguous and randomly similar sites can hinder phylogenetic analyses, since they may lead to erroneous estimations of substitution model parameters (Kück *et al.* 2010). Ambiguously aligned sections were subsequently removed (i.e. masked) from the alignments with the help of Alicut v. 2.31 (https://github.com/mptrsen/scripts/blob/master/ALICUT_V2.3.pl). The average length of the masked protein MSAs was 370.81 bp (median: 302 bp; minimum: 36 bp; maximum: 3,636 bp).

Removal of non-informative partitions (OGs)

To estimate the information content within each OG, we used MARE software v. 0.1.2-rc (Misof *et al.* 2013; downloaded from: <http://mare.zfmk.de>). MARE requires all OGs to be concatenated to a supermatrix, in order to be able to assess the information content of each OG in relation to the overall information content. Before the MSAs were concatenated, trailing gaps were 'closed' using a customized perl script (provided by 1KITE consortium) by replacing any '-' with an 'X'. The supermatrix and a partition (charset file), that contains information about start and end positions of each MSA, both serving as input for MARE, were then generated with FasConCat v. 1.11 (Kück & Meusemann 2010; downloaded from: <https://github.com/PatrickKueck/FASconCAT>). The information content was assessed for the 3,812 OGs, one time with and a second time without the primer-based taxa under default settings. A customized perl script (provided by the 1KITE consortium) was then used to remove OGs with an information content equal to zero. Using the MARE output files, this was

done on both amino-acid and nucleotide level. This removed a total of 28 MSAs, none of which contained any primer-based taxa. MARE was rerun on the remaining 3,784 OGs, with and without the sequences from primer-based taxa.

OG selection by taxonomic representation

In this step OGs were excluded if they did not represent specific taxa. Two different groupings were tested, one with a stricter and one with a more relaxed grouping. The strict grouping lead to the removal of any OG that did not contain all species belonging to Staphylinidae and Silphidae. The relaxed grouping required one representative per subfamily within Staphylinidae and one member of Silphidae. In both scenarios, the remaining taxa could but did not have to be present. The 40 OGs that also contained primer-based taxa had previously been excluded from both test runs, so that no OG would be lost. The reduced datasets of both groupings were obtained using a customized perl script (provided by the 1KITE consortium) and they were compared via MARE. The dataset applying the strict grouping (1,043 OGs) had a higher information content (70.4%) than the dataset applying the relaxed grouping (2,867 OGs, information content 62.6%). We chose the dataset with the higher information content for downstream analyses, hence continued with the dataset after the strict grouping was applied and subsequently added the 40 OGs that contained the primer-based taxa

Removal of long-branching species

Duplication events that occurred a very long time ago (deep duplications) can lead to falsely identified orthologs in OGs (Yang & Smith 2014) and might be reflected by very long branches in a phylogenetic tree. The same can happen if genome or transcriptome assemblies are poor so that erroneous sequences appear to evolve faster. To see whether there are any sequences within a MSA that generate disproportionally long branches, we performed maximum likelihood analyses in IQ-Tree v. 1.6.5 (Nguyen *et al.* 2015) to generate single-gene (OG = gene) trees.

We performed the ModelFinder option for each ortholog MSA (Kalyaanamoorthy *et al.* 2017), considering only general amino-acid models and using *T. castaneum* as the outgroup taxon and subsequently one maximum likelihood tree search using default settings. The following steps have been done using customized bash scripts (by J.L. Kypke, adapted from H. Escalona): In order to be able to compare branch lengths across the different topologies for each phylogenetic tree, we calculated the length of every leaf's parent branch in percent. Branch lengths were extracted from each tree file with the help of Newick Utilities v. 1.6.0 (Junier & Zdobnov 2010) (using 'nw_distance' with the options '-n' to print the species name next to the distance and '-m p' to print the length of a selected node's parent branch). For each species the obtained branch length was divided by the total tree length (listed in the IQ-tree log file of each run) and multiplied by 100. The obtained value reflects the length of a branch in relation to the other branches within a tree and can therefore be compared across all single

Chapter 1

gene trees. Although it has been shown that a removal of erroneous sequences in an alignment of one OG can improve multi-gene phylogenies (Yang & Smith 2014), it has not been tested sufficiently to define clear cut-offs. To be able to define such cut-off, i.e. a value beyond which branches would be defined as too long, samples of phylogenetic trees were visually inspected. All species with values ≥ 10 showed very long branches (Figure 2). To lower possible long branch attraction (LBA), we subsequently removed all sequences from each OG of species with values ≥ 10 .

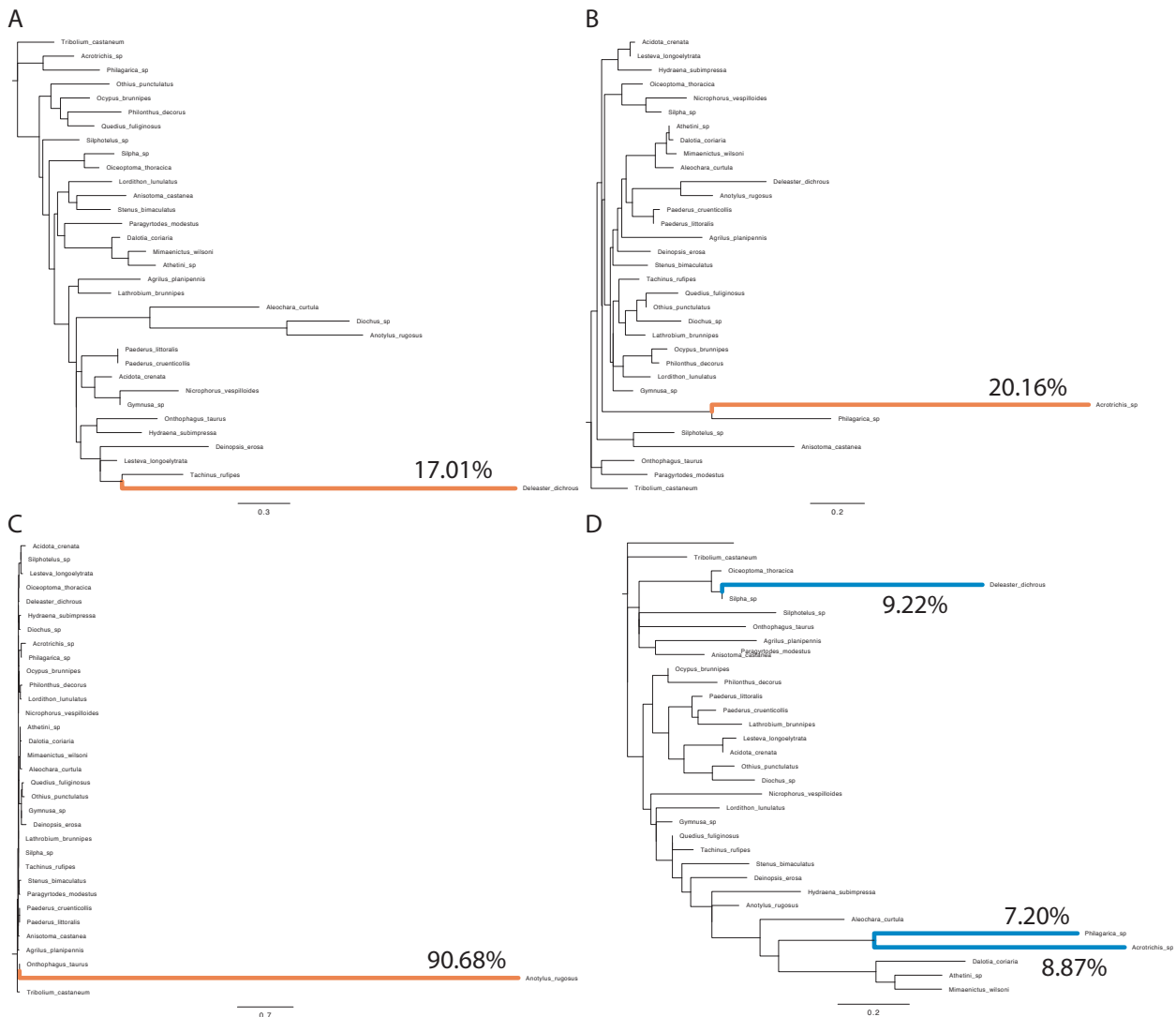


Figure 2. Examples of single-gene trees for every multiple sequence alignment (MSA) to identify sequences leading to long-branches. Sequences of species with branch lengths $> 10\%$ were subsequently removed from the analysis. A-C each show species on long branches (marked in orange); A: EOG090R0A72; B: EOG090R0BLW; C: EOG090R0CBF; D: EOG090R0HXM, an example with sequences leading to long branches but below the cut-off were kept (in blue).

Removal of short sequence alignments (OGs)

The sequence lengths before the alignment procedure differed drastically both within and amongst species (Supplementary Materials, Table S5). We critically doubt that sequences shorter than

30 bp are reliable in downstream analyses (e.g. to estimate parameters, also orthology might be inferred erroneously). Since there is unfortunately no concrete cut-off recommended in the literature that would suggest which size would be appropriate and in fact it might depend on the data type and goal of the analysis, we generated frequency plots in order to get the distribution of sequence lengths in the OGs in R v. 3.5.1 (R Core Team 2018) using the packages 'ggplot' (Wickham 2016), 'plyr' (Wickham 2011), 'grindExtra' (Auguie 2017) and 'tidyverse' (Wickham 2017), Figure 3. We decided to choose a conservative cut-off and removed all OGs with sequences < 200 bp, which is equivalent to 8.31% of the dataset at that point. At the same time this cut-off did not lead to the removal of any OGs with sequences of the primer-based taxa. The remaining 993 OGs formed the final datasets.

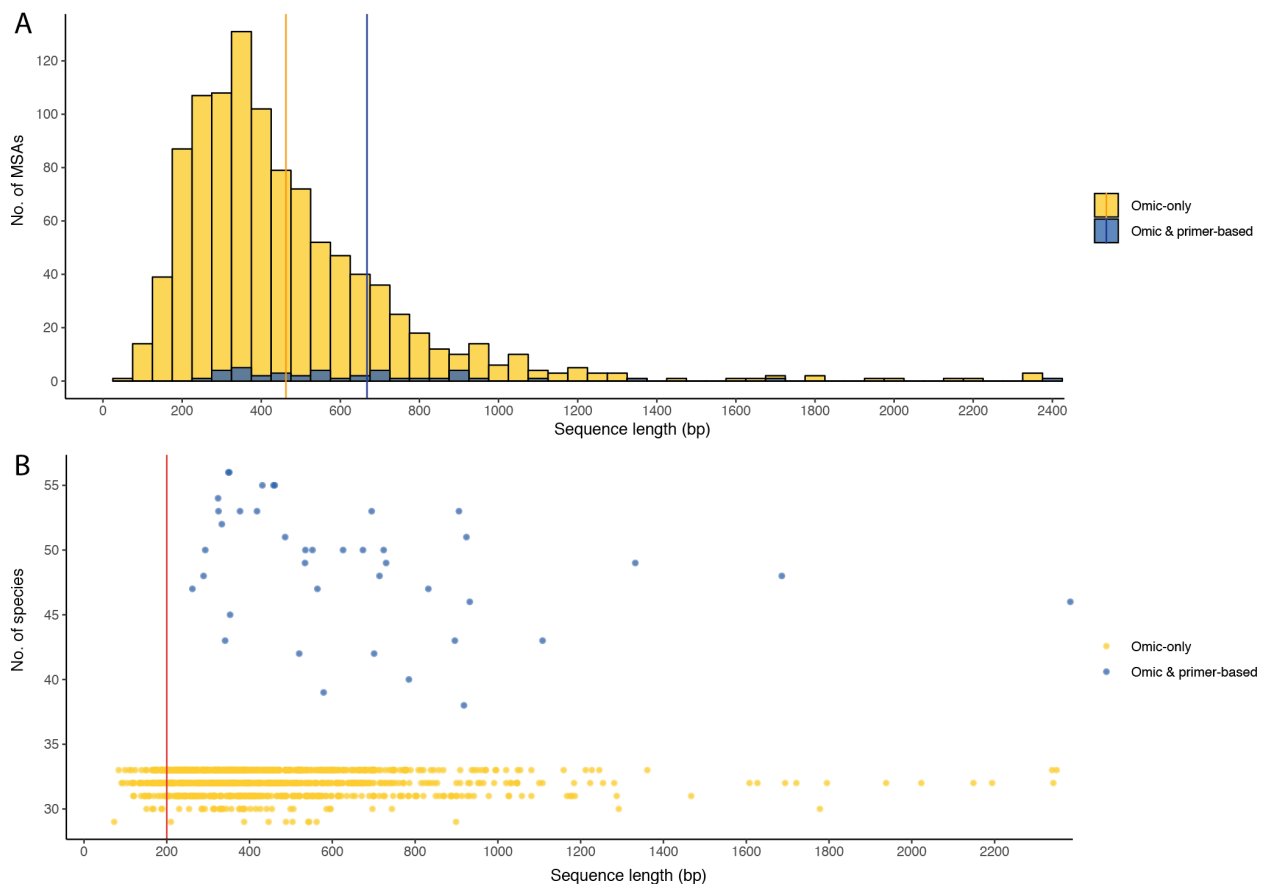


Figure 3. Frequency plots by the omic-only (yellow) and omic & primer-based (blue) datasets. **A:** distribution of the number (No.) of multiple sequence alignments (MSAs) by their length in base pairs (bp); averages are marked in orange (omic-only) and blue (omic & primer-based); **B:** number of species for each MSA by its sequence length; the red line marks the cut-off at 200 bp.

Generating the final supermatrices

The final supermatrices, cleaned from sequences, which might lead to long branches and removal of sequences < 200 amino-acids, were generated with FasConCat v. 1.11 (Kück & Meusemann 2010) as outlined above (final matrices 1 and 2). All two final data matrices consist of the same 993 OGs, but in supermatrices 1 the sequences of the primer-based taxa have been removed (re-

Chapter 1

ferred to as omic-only). The overall information content of the supermatrices 1 and 2 was assessed again using MARE as described earlier. Since no OGs had an information content = 0, we kept all OGs for subsequent analyses.

As mentioned before, sequences within each OG are of different length due to the data type that led to more or less complete assemblies. An uneven distribution of missing data in MSAs can negatively affect phylogenetic analyses and lead to erroneous trees (Cho *et al.* 2011; Dell’Ampio *et al.* 2014). To be able to spot potentially problematic taxa, not only on the OG level (present/absent from an OG), but also on the site-level within the supermatrix, we analyzed the distribution of missing data in our superalignments using Alistat v. 1.7 (Wong & Jermin; downloaded from <https://github.com/thomaskf/AliStat>). The software generates heat maps that show the distribution of missing data in pairwise comparisons of aligned sequences. We used the -r option to reorder both rows and columns in those maps. Once generated by Alistat, Adobe Illustrator CS6 v. 16.0.4 (Adobe Systems Inc. 1987-2012) was used to merge the two heat maps in places where they were identical (the ‘omic-only’ data), species names were made readable and data type labels were added. Additionally, we used the generated R scripts under slight modification to plot the distribution of completeness scores for individual sequences (rows) as well as sites (columns) in each supermatrix.

Data partitioning and substitution model selection

ModelFinder (Kalyaanamoorthy *et al.* 2017) as implemented in IQ-Tree was used to search for suitable models for the supermatrices considering the corrected Akaike information criterion (AICc) values to measure the goodness of fit for any tested model. The supermatrices were partitioned by gene boundaries. We chose the edge-proportional partition models where all partitions share the same set of branch lengths but are allowed to have different evolutionary rates, which rescales a partition’s branch lengths (-spp option). We also considered free rate models, the protein mixture models LG4X and LG4M, and otherwise defaults (23 models with empirical amino-acid exchange rate matrices, plus two protein mixture models). To reduce a possible over-parameterization and to increase the model fit, we chose the option to merge partitions, once they were identified per gene. To speed this very time consuming process up a little bit, we used the fast relaxed clustering algorithm of Partition-Finder2 (Lanfear *et al.* 2017) and set the absolute maximum number of partition pairs in the partition merging phase to 15,000 (ca. 15 times the number of partitions).

Phylogenetic and bootstrap analysis

Under the partition schemes and with the models identified as described above, we inferred phylogenetic trees with IQ-Tree using the maximum likelihood approach. Always starting from a random start tree, we calculated 20 ML trees per supermatrix using *T. castaneum* as the outgroup. The best tree was chosen based on the smallest absolute logLikelihood value. Branch supports were as-

sessed with 100 standard non-parametric bootstrap replicates in IQ-Tree using again the option -spp. We then mapped all bootstrap trees onto the best ML tree using the IQ-Tree -sup option, thus obtaining the best ML tree including bootstrap support. PDF files of the best tree for either of the datasets were generated in FigTree v. 1.4.3 (Andrew Rambaut; downloaded from: <https://github.com/rambaut/figtree>) that were graphically edited in Adobe Illustrator CS6 v. 16.0.4 (Adobe Systems Inc. 1987-2012). Bootstrap convergence was assessed to find the minimal number of bootstrap replicates needed in order to get stable support according to (Pattengale *et al.* 2010). This was done with all bootstrap trees in RAxML v. 8.2.11 (Stamatakis 2014) choosing the Weighted Robinson-Foulds (WRF) distances (default option, -B 0.03, -I autoMRE) and running 10,000 permutations. For the same dataset, bootstrap convergence was assessed five independent times with a different random seed.

To check whether there were any ‘rogue’ species in our dataset, meaning that their unstable position might weaken the overall bootstrap support and their position could be variable in a set of trees, we used the software RogueNaRok v. 1.0 (Aberer *et al.* 2013); downloaded from: <https://github.com/aberer/RogueNaRok>). If a taxon is identified as rogue then their position in a phylogenetic tree needs to be considered with caution. Additionally, it has been shown that rogue taxa can lead to low bootstrap supports on branches of otherwise stable evolutionary relationships (Misof *et al.* 2014). RogueNaRok identifies rogue taxa given the final ML tree and all bootstrap trees.

The unpublished software Uniquetree v. 1.9 (Thomas Wong, Australian National University, Australia; kindly provided by the 1KITE consortium) was used to compare the topology among the obtained 20 ML trees generated for each supermatrix with each other. The goal is to find out how many and how many times different topologies were found, specifically the one of the best tree.

Results

Raw data pre-processing and assemblies

The statistics of the assemblies obtained using whole genome re-sequencing (nineteen taxa, sixteen newly sequenced), generated by the different software in the process of assembling the genomes, were collected in Supplementary Materials, Table S3. For each species the table reports: the quality of the raw sequence reads before and after being trimmed (FastQC output), genome size estimates based on k-mer counts of size 31 on the raw sequence reads (generated with Jellyfish and GenomeScope), assembly statistics for the sixteen newly generated genomes (reported by SparseAssembler), and summary statistics for the final genome assemblies, i.e. after removing heterozygous regions with Redundans (generated by Quast).

Chapter 1

K-mer frequency plots generated after counting the k-mers of size 31 can be found in Supplementary Material, Figure S1. Genome size estimates differed between software Jellyfish (J) and GenomeScope (GS) but were estimated for seven species: *D. coriaria*: 376 Mb (J) and 228 Mb (GS); *D. erosa*: 251 Mb (J) and 515 Mb (GS average); *M. wilsoni*: 268 Mb (J) and 71 Mb (GS average); *L. longoelytrata*: 24 Mb (J) and 44 Mb (GS average); *Gymnusa* sp.: 190 Mb (J); *L. brunnipes*: 47 Mb (J). Graphs were generated during the reduction process of the genome assemblies by Redundans (Supplementary Materials, Figure S2). They depict the identity between contigs as well as scaffolds, i.e. the number of locations a given contig or scaffold potentially aligns to the assembly.

The BUSCO assessments for the final assemblies of the three publicly available Aleocharinae genomes obtained 87.47% of the Endopterygota universal single-copy orthologs, the newly sequenced genomes obtained 26.16% and the transcriptomes obtained 42.51%. The complete results of the BUSCO search per species can be found in Supplementary Materials, Table S4.

Mining the orthologous gene set

The identification of orthologous sequences in single-copy protein-coding genes of the 54 target species resulted in a total of 3,812 OGs out of 3,822 possible OGs, with a total alignment length of 2,269,610 amino-acid sites. In the omic-only dataset we identified on average 97.09% out of all possible OGs in each species (on average 97% in the reference genomes, 98.65% in the low-coverage genomes and 94.15% in the transcriptomes). The additional 24 primer-based species were only present in a total of 40 OGs, which is equivalent to 3.03%. Adding them lowers the total average down to only 57.48% in the omic & primer-based dataset. A detailed summary table of the number of genes found for each target species and their general composition (number of amino-acids, 'X's, stop codons, lengths) can be found in Supplementary Material, Table S5.

Once multiple sequence alignments (MSAs) were generated for the orthologous sequences of every OG, randomly similar sites were removed (Aliscore). Overall, 30.74% ambiguous or randomly similar sites were identified on average (median: 26.92%; minimum: 0%; maximum: 91.53%), shortening the total alignment length of all OGs to 1,418,910 sites.

The information content was assessed for those 3,812 MSAs (=partitions), with and without the primer-based taxa under default settings using MARE software: the information content and taxa were weighed in a ratio of 3:1, and 20,000 quartets were evaluated. The supermatrix that included all species (omic & primer-based) had an overall information content of 32% (57 species; matrix saturation in terms of partitions being present or absent: 50%) and the omic-only supermatrix of 55% (33 species; matrix saturation: 86%). Twenty-eight MSAs with an information content of zero were subsequently removed, none of which contained any primer-based taxa. MARE was rerun on the remaining 3,784 MSAs, with and without the sequences from primer-based taxa. The information content of the new

supermatrix with sequences of all species improved by only 0.2% (new information content 55%; 57 species; matrix saturation: 50%), for the omic-only supermatrix by 0.4% (new information content 32%; 33 species; matrix saturation: 86%). The new alignment length spanned a total of 1,413,546 sites. The removed MSAs contained amino-acid sequences of comparatively few species (average: 9.64; median: 4; min: 2; max: 28). Additionally, 25% on average of the sites in the sequences of each removed MSA were uninformative (average gap content: 5.15%; average 'X' count: 20.37%) and all but one MSA were short (average length: 295.79 bp; median: 200.50 bp; minimum: 36 bp; maximum: 2,446 bp).

The inference of single-gene (individual MSAs) phylogenies and the comparison of branch lengths and inspection of phylogenies for a subsample of all 1,083 single-gene trees, lead to the decision to remove all sequences of species with values > 10 from individual MSAs. Overall, 54.29% of all OGs contained sequences with branches ≥ 10 and altogether 728 sequences belonging to 35 species (average: 20.8 sequences; median: 8 sequences from *Lordithon lunulatus*; minimum: one sequence in *D. coriaria* and *Silphotelus* sp. & the primer-based taxa *Megalopaederus* sp., *Osorius* sp. and *Tachinus* sp. (INB088); maximum: 208 sequences in *Acrotrichis* sp.) we removed. The number of sequences removed per species and the data type is shown in Supplementary Material, Table S6.

Once the final supermatrices of 993 concatenated MSAs with a total alignment length of 494,743 sites had been generated, and before running the phylogenetic analysis, the information content was again assessed using MARE. Supermatrix 1 (omic-only) had an information content of 65.50% (matrix saturation: 97.10%), and supermatrix 2 (omic & primer-based) of 39.20% (matrix saturation: 57.50%). The aligned sequences of both supermatrices were subsequently compared in pairs to assess their completeness (amino-acid site coverage). The overall completeness score (C_a) for the superalignments differed between datasets and was higher in the omic-only dataset (59%) than when the sequences of the primer-based species were added (35%). Minimum values of sequence comparisons of individual sites, either by rows (C_r) or by columns (C_c) differed between datasets, while maximum values were identical or very similar (Supplementary Materials, Table S7). When supermatrices are compared by columns, then the overall distribution of ambiguous sites in supermatrix 2 (omic & primer-based) is skewed to the right (Supplementary Material, Figure S3, Aa & Ba). This is indicative for a large amount of missing data in supermatrix 2 and can hence be attributed to added species that are only represented by a maximum of 40 OGs in the superalignment. On the other hand, when supermatrices are compared by rows (species-wise), then the distribution of ambiguous sites in supermatrix 1 is bimodal (Supplementary Materials, Figure S3, Ab) and in supermatrix 2 trimodal (Supplementary Materials, Figure S3, Bb). The two modes indicate that the omic-only dataset is heterogeneous, where the sequences of a group of species have more missing data than another. The additional peak in supermatrix 2 can then be attributed to the added species that are highly incom-

Chapter 1

plete when entire rows in the superalignments are considered. The generated heat map of the pairwise comparisons of sequences (rows) in the superalignments shows the amount of missing data per species and data type (Supplementary Materials, Figure S4). Low shared-site coverage is displayed in dark blue and high shared-site coverage in white.

Phylogenetic analysis

ModelFinder selected 14 protein substitution models as best suitable for the partitions in supermatrix 1 (omic-only) and 15 for supermatrix 2 (omic & primer-based) respectively. A summary of the chosen models and the number of times each model was used for merged partitions in each supermatrix can be found in Supplementary Materials, Table S8. Once the substitution models had been identified for individual partitions, they were merged to increase model fit and to reduce the number of parameters in the analysis. The 993 partitions were merged into 828 (supermatrix 1) and 400 (supermatrix 2) partitions. The best tree topology out of the twenty ML tree searches for each supermatrix had a log-likelihood of -7,748,901.04 (supermatrix 1, Figure 4) and -8,262,145.86 (supermatrix 2, Figure 5) (Supplementary Materials, Table S8). In supermatrix 1 the 20 ML trees all had the exact same topology, while in supermatrix 2 twenty different topologies were found. Inspecting the topologies showed that three maximum agreement subtrees of 45 taxa were found and the topologies were changed by the four taxa that were subsequently identified as rogue, i.e. species with unstable position in the 100 bootstrap trees (see below). For each dataset (supermatrix 1 and 2) bootstrap convergence was assessed five independent times, to test whether sufficient bootstrap replicates had been drawn. For each dataset the five replicates were identical: Supermatrix 1 (omic-only) always reached bootstrap convergence after 50 replicates (WRF: 0.18%; 10,000 permutations $\geq 3\%$); supermatrix 2 (omic & primer-based) did not yet reach bootstrap convergence after the 100 replicates (WRF: 3.8%; 1,586 permutations $\geq 3\%$). While there were no rogue taxa in the topology of the omic-only dataset, the following 4 species were identified as rogue in the omic & primer-based dataset. All of them are part of the primer-based dataset and are listed in decreasing order of their 'improvement value', i.e. from most to least rogue: *Scydmaenus* sp.; Ptiliidae gen. sp.; *Centrophthalmus* sp. and *Osorius* sp.

The phylogenetic tree topologies of both datasets were very similar, although the overall support was lower in the tree from the expanded dataset (omic & primer-based), while the tree based on the omic-only dataset was fully supported in all but one node. The rove beetles were inferred paraphyletic with respect to Silphidae. Two of the long established morphology-based groups of subfamilies were recovered monophyletic: the omaliine group and the oxyteline group. Additionally, the backbone of the Staphylinidae phylogeny was resolved and shows that the monophyletic omaliine group (Omaliinae + (Proteininae + Pselaphinae)) is sister to all other rove beetles including Silphidae.

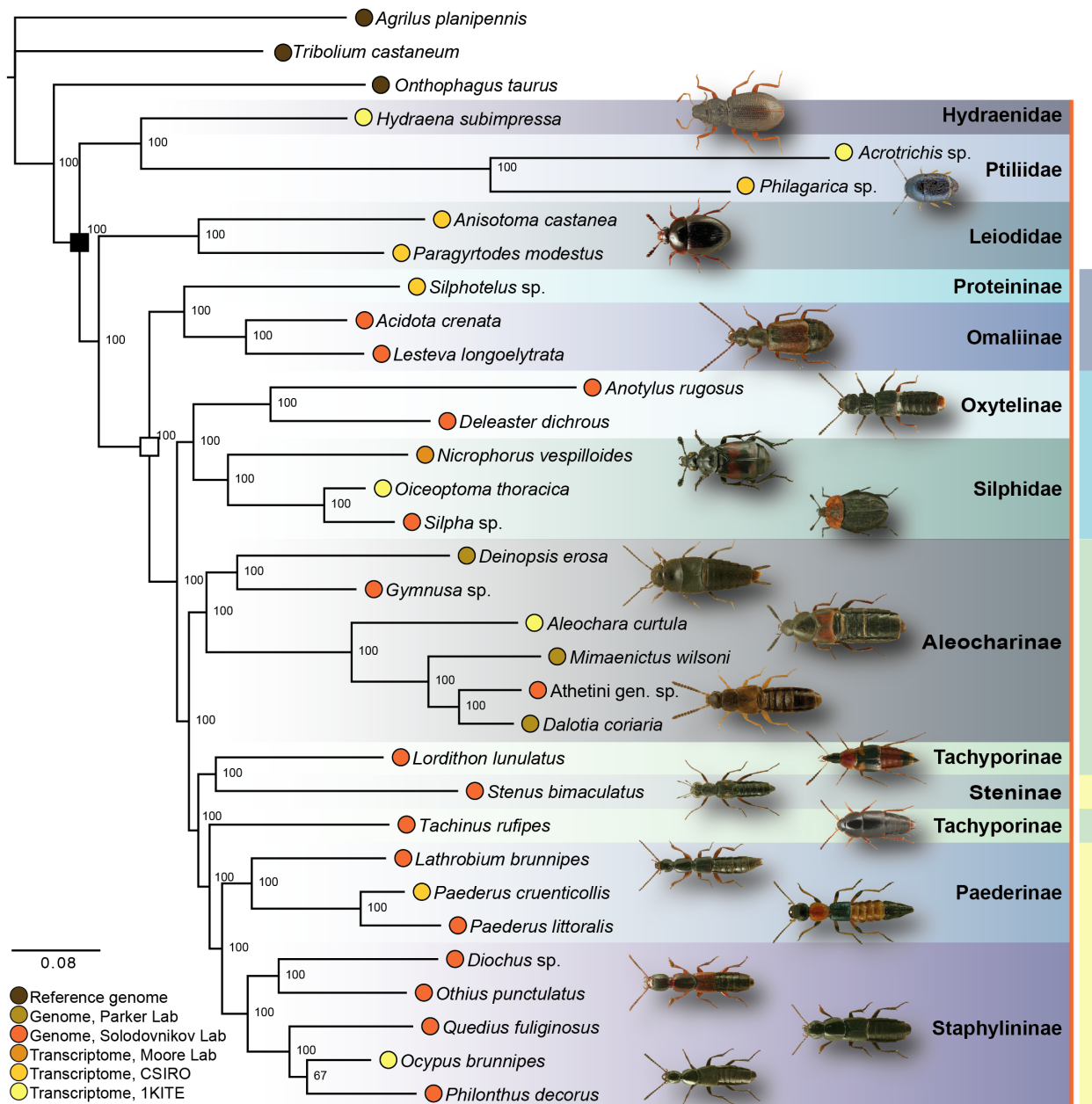


Figure 4. Best maximum likelihood (ML) tree inferred from supermatrix 1 (omics-only dataset), with bootstrap support. Node marked by filled square indicates Staphyloidea; node marked by empty square indicates Staphylinidae. Color blocks right of the orange line correspond to the four subfamily groups: omaline- (dark blue); tachyporine- (light green); oxyteline- (light blue); and staphylinine group (yellow).

The tachyporine group that had earlier been suspected as polyphyletic (Thayer 2016) did not form a monophylum. Interestingly, both members of the subfamily Tachyporinae rendered the staphylinine group paraphyletic in two places. The same pattern can be observed in the expanded dataset: Members of the genus *Tachinus* form the sister clade to (Paederinae + Staphylininae), while *Lordithon lunulatus* branches out sister to Steninae (omic-only dataset), or (Steninae + Scydmaeninae) (omic & primer-based dataset) in the clade sister to ((Paederinae + Staphylininae) + *Tachinus* sp.). Based on our analyses, the monophyletic Silphidae are the sister to all other members of the oxyteline group. In the

Chapter 1

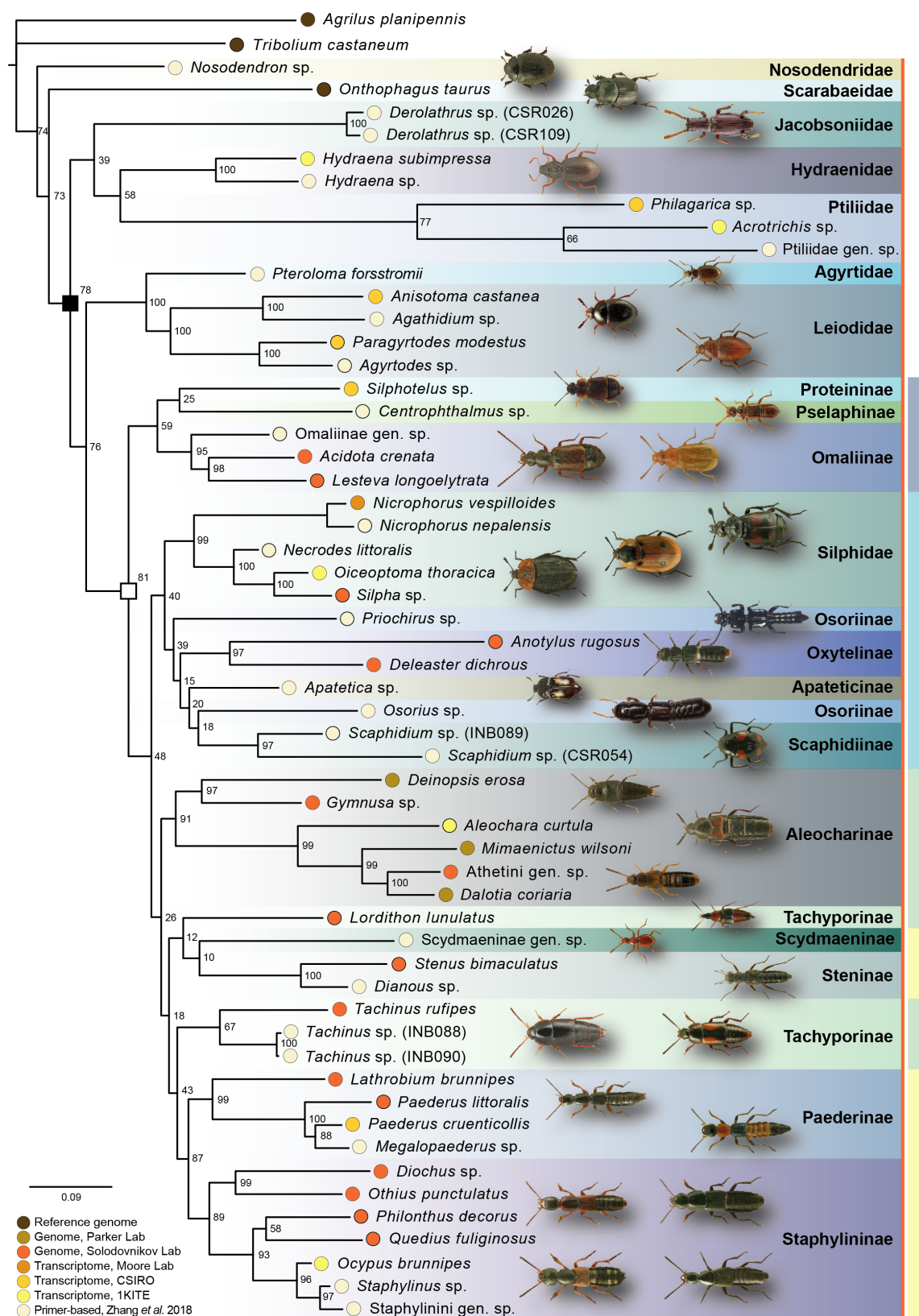


Figure 5. Best maximum likelihood (ML) tree inferred from supermatrix 2 (expanded dataset), with bootstrap support. Node marked by filled square indicates Staphylinoidea; node marked by empty square indicates Staphylinidae. Color blocks right of the orange line correspond to the four subfamily groups: omaliine- (dark blue); tachyporine- (light green); oxyteline- (light blue); and staphylinine group (yellow).

omic-only dataset, the only sampled members of the oxyteline group were two species from the subfamily Oxytelinae. However, in the expanded dataset Oxytelinae is associated with (Apateticinae + (Scaphidiinae + *Osorius* sp.)) and the second osoriine, *Priochirus* sp., is sister to all of them.

In our analyses, the sister to all Staphylinidae (+Silphidae) is the family Leiodidae in the omics-only dataset, and (Leiodidae + Agyrtidae) in the expanded dataset. In the phylogeny of the expanded dataset members of the superfamily Derodontoidea, the families Nosodendridae and Jacobsoniidae do not form a monophylum. The family Jacobsoniidae was recovered as the sister to (Hydraenidae + Ptiliidae) within Staphylinoidea, while Nosodendridae were found sister to (Scarabaeoidea + Staphylinoidea).

Discussion

The main purpose of this experiment was to explore if an assemblage of mixed genomic markers of varying size and quality can be used to infer the placement and robust backbone phylogeny of the rove beetles with the following specific objectives:

- 1) design a database (ortholog set) that contains markers that can assign OGs to species across all Staphylinoidea;
- 2) test if low-coverage genomes are a viable option, especially without prior genome size estimation; and
- 3) design a specific bioinformatic pipeline for rove beetles to integrate the different data sources.

We have performed a number of carefully chosen and rather conservative (in terms of OG acceptance) analytical steps to select only the most suitable OGs for the phylogenetic inference. We discuss our results by comparing congruence between our best trees with the *status quo* of phylogenetic knowledge of the rove beetles to identify the predictability of our methods.

Staphylinid phylogenomics vs. the *status quo* of systematics

The sister-group relationships of Staphylinidae within Staphylinoidea as well as some sister-group relationships among tribes and subtribes within the Staphylinidae are the phylogenetic levels where we have a plethora of widely agreed knowledge (summary in Thayer 2016). On the contrary, sister-group relationships among the subfamilies of Staphylinidae, i.e. those at the intermediate level, are where most of the open questions and controversies are found. All hitherto performed phylogenetic work failed to reveal an unambiguous signal at that intermediate level, where rove beetle phylogenetics needs a breakthrough, and it is where we expect our approach will be helpful. Therefore, first we consider better-known deeper (inter-familial) and more shallow (intra-subfamilial)

Chapter 1

relationships in our results in comparison with published data. After that we discuss the intermediate nodes, i.e. the basal phylogeny of Staphylinidae that lacks a consensus. This is the level our approach targeted and where our data can be considered as novel results worth further exploration.

Sister-group relationships among families of Coleoptera correspond in both phylogenies (small and large). Nosodendridae is sister to (Scarabaeidae + Staphylinidae), which is equivalent to the well established (Scarabaeoidea + Staphylinoidea) given our reduced taxon sample outside Staphylinidae. Using various morphological and molecular data, the enigmatic family Nosodendridae has previously been placed in remote superfamilies or even series of Polyphaga without any agreement among studies (Newton & Thayer 1995; Ge *et al.* 2007; Hunt *et al.* 2007; Lawrence *et al.* 2011; Bocak *et al.* 2014; McKenna *et al.* 2015). In the latest phylogeny of Coleoptera Zhang *et al.* (2018), contrary to all previous results, Nosodendridae was robustly recovered in a novel position as a sister clade to Staphyliniformia, Bostrichiformia, and Cucujiformia. We did not sample Bostrichiformia and Cucujiformia, but the position of Nosodendridae in our analysis is consistent with Zhang *et al.* (2018) (Figures 5 & 6C).

In agreement with many previous studies (Kukalová-Peck & Lawrence 1993; Hansen 1997b; Caterino *et al.* 2005; McKenna *et al.* 2015; Timmermans *et al.* 2015; Zhang *et al.* 2018), we recovered Staphylinoidea (including Jacobsoniidae) as sister to Scarabaeoidea, thus supporting a traditional monophyletic Haplogastra. This is congruent in both our analyses, i.e. with the full dataset and the one restricted to the genomic data only. Consistently with Zhang *et al.* (2018), *Derolathrus* (Jacobsoniidae), which is currently assigned to the superfamily Derodontoidea, was recovered as sister to (Hydraenidae + Ptiliidae) with high support. Altogether they were assigned sister to the rest of Staphylinoidea (Figure 5). The earlier well established topology ((Hydraenidae + Ptiliidae) + remaining Staphylinoidea) was recovered in our analyses with and without Jacobsoniidae, which were only present in the omic & primer-based dataset (Figures 5 & 6). The same or similar placements were earlier recovered for *Derolathrus* sp. on the basis of both morphological and molecular data (Lawrence *et al.* 2011; McKenna *et al.* 2015), thus strongly suggesting that Jacobsoniidae should be transferred to Staphylinoidea.

As in earlier morphology-based (Hansen 1997; Beutel & Molenda 1997; Lawrence *et al.* 2011), restricted molecular- and morphology-based (Caterino *et al.* 2005), restricted molecular-based (McKenna *et al.* 2014, 2015) and the latest more robust molecular phylogeny by Zhang *et al.* (2018), both our datasets recovered Staphylinidae paraphyletic with respect to Silphidae (Figures 4 & 5). Among all controversial positions of Silphidae inside Staphylinidae, our results are consistent with Zhang *et al.* (2018) in placing Silphidae as sister to the oxyteline group of subfamilies. Regarding Silphidae, our result conflict with the adult and larval morphology-based analysis by Grebennikov & Newton (2012) who revealed Silphidae as sister to Staphylinidae. However, even in that study

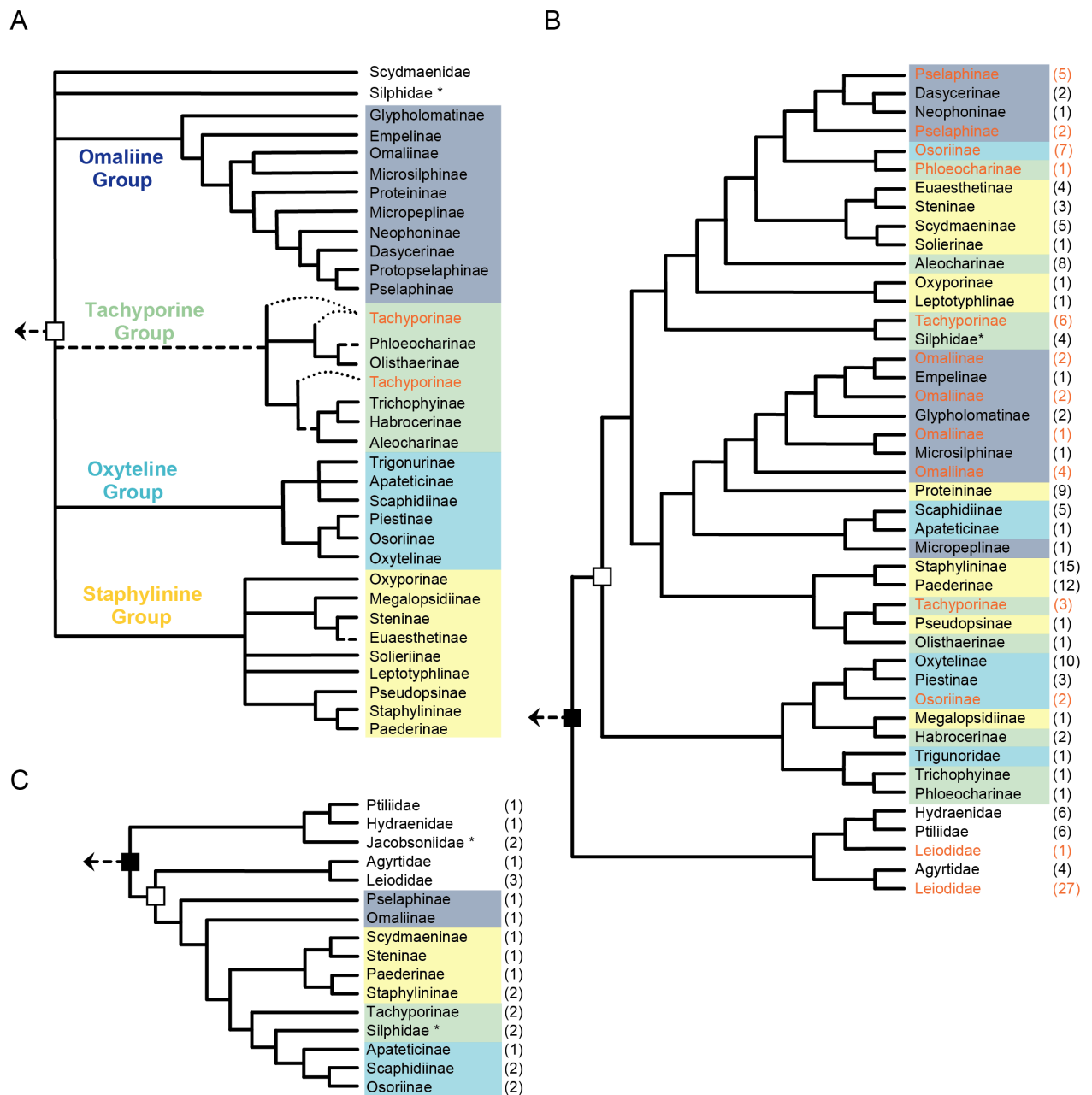


Figure 6. Phylogenies and diversity of Staphylinidae. A: Consensus backbone phylogeny after Thayer (2016); B: Simplification and extract from McKenna *et al.* (2014); C: Simplification and extract of Zhang *et al.* (2018); Colors refer to the subfamily groups in A; Dark blue: omaliine group; Green: tachyporine group; Light blue: oxyteline group; Yellow, staphylinine group. Node marked by filled square indicates Staphylinoidea; node marked by empty square indicates Staphylinidae.

Silphidae were assigned phylogenetically close to the subfamilies of the oxyteline group, and in the molecular part of their paper with the limited 18S rDNA-based analysis, Silphidae were always nested within Staphylinidae, even though with low resolution. Our results also show the former Scydmaenidae nested within Staphylinidae. Such placement was first shown based on adult and larval morphology, which is why they were downgraded to a subfamily of Staphylinidae in (Grebennikov &

Chapter 1

Newton 2009). The same topology that placed both Silphidae and Scydmaeninae inside Staphylinidae was recovered by both analyses, with and without Zhang *et al.* (2018) data (Figures 4 & 5). Staphylinidae including Scydmaeninae and Silphidae were assigned sister to the (Leiodidae + Agyrtidae) clade, consistently with Zhang *et al.* (2018) (Figure 6C). There is a broad concept of Staphylinidae including Scydmaenidae and Silphidae that coincides with the staphylinid group proposed as early as in Lawrence & Newton (1982), however, Silphidae have not been integrated yet, because their definitive placement could not be assigned in the aforementioned studies. This is also true for the (Leiodidae + Agyrtidae) clade as a sister group to Staphylinidae, often recovered as such in the above cited references.

At the more terminal level of relationships among taxa within the families and subfamilies, and among some subfamilies of Staphylinidae whose sister group relationships are less problematic, our topology from both datasets is largely consistent with the results widely agreed as well supported. All non-Staphylinidae families sampled with more than one terminal are revealed monophyletic. Internal topology of the relatively well sampled family Silphidae corresponds with the phylogenetic knowledge about this group (Newton 1997; Dobler & Müller 2000). All sampled (omic-only dataset: 8; omic & primer-based dataset: 13) subfamilies of Staphylinidae, except Tachyporinae and Osoriinae (only in the omic & primer-based dataset), are recovered monophyletic.

The monophyly of the tachyporine group of subfamilies (here represented by Tachyporinae and Aleocharinae) and the subfamily Tachyporinae itself were in fact challenged long ago (reviewed in Thayer 2016). Even when Lawrence & Newton (1982) suggested the tachyporine group of subfamilies, they declared it to be a group of convenience. It is a predominantly predaceous assemblage with no clear synapomorphies, but lacking the synapomorphies of the other three groups (Thayer 2016). Neither previous, nor subsequent work has fully clarified sister group relationships among its constituents or between them and other groups, and the placement and ranking of the subgroups have varied widely. Non-monophyletic Tachyporinae also appeared in the molecular-based phylogeny of McKenna *et al.* (2014) (Figure 6 B).

The oxytelinae group of subfamilies was assigned as a monophyletic clade. This is congruent with morphology-based studies conducted in the past (reviewed in Thayer 2016). Interestingly, Osoriinae, which were present only in the larger dataset, were not revealed monophyletic: *Priochirus* and *Osorius* are separated several nodes apart by Oxytelinae and Apateticinae. The phylogenetically, morphologically and ecologically very diverse Osoriinae were very little studied in the past and the only putative synapomorphy of this subfamily is lacking paratergites of the abdomen. However, one cannot rule out independent loss of paratergites, which is observed in some truly remote lineages of other subfamilies, for example Steninae or Pinophilini of Paederinae. Apateticinae were resolved sister to the (*Osorius* sp. + Scaphidiinae) clade in the expanded dataset. This clade was first assigned in such

combination of subfamilies by Zhang *et al.* (2018) (Figure 6C). However, the sister group relationships of Apateticinae and Scaphidiinae are also known from morphology-based works (reviewed in Thayer 2016) and from the molecular study by McKenna *et al.* (2014) (Figure 6 B).

We resolved monophyletic Omaliinae and the omaliine group of subfamilies in both analyses (Figure 4 & 5). This is consistent with the generally well-supported monophyletic omaliine group based on a morphological analysis (Newton & Thayer 1995). Contrary to the more limited taxon sample in Zhang *et al.* (2018), where Pselaphinae and Omaliinae did not form a single clade (Figure 6C), our expanded dataset, in which the omaliine group is represented by Omaliinae, Proteininae and Pselaphinae is monophyletic (Figure 5). All species representing Oxytelinae and Scaphidiinae in our analysis formed separate clades, respectively.

The internal topology of the monophyletic Aleocharinae, represented by the same taxa in both datasets is identical and perfectly consistent with the phylogeny of aleocharines as we know it (Elven *et al.* 2012; Osswald *et al.* 2013; Yamamoto & Maruyama 2018). Another large subfamily, the Staphylininae, was recovered sister to Paederinae in both datasets, which is in agreement with the long established view on the affinity of both subfamilies, challenged in some recent phylogenetic studies including our own work on the staphylinine tribe Othiini that rendered Staphylininae paraphyletic with respect to Paederinae (Brunke *et al.* 2016; Kypke *et al.* 2018). Our genomic data reinforce a more traditional view that Staphylininae and Paederinae are sister clades. In both datasets the internal branching pattern within both subfamilies is consistent with the results of other studies that sampled more taxa and analyzed molecular (Solodovnikov & Newton 2005; Solodovnikov *et al.* 2013; Schomann & Solodovnikov 2017) or morphological (Solodovnikov & Newton 2005; Solodovnikov *et al.* 2013; Schomann & Solodovnikov 2017) data, also in consideration of stem and crown groups (Solodovnikov & Newton 2005; Solodovnikov *et al.* 2013; Schomann & Solodovnikov 2017). The only clade in our phylogeny inconsistent with the current knowledge and with little support in both analyses is (*Philonthus decorus* + *Quedius fuliginosus*). According to previous studies (Brunke *et al.* 2016; Chani-Posse *et al.* 2017) and given our taxon sampling, *Philonthus decorus* should have formed a clade together with *Ocypus* and *Staphylinus*, where *Quedius* would be the sister to them. In the tree based on the expanded dataset the genus *Paederus* is paraphyletic with respect to *Megalopaederus*, which is consistent with the well-established fact of the artificial generic boundaries in the *Paederus*-complex (Li *et al.* 2013).

Concerning the backbone of the Staphylinidae, which is the main unknown area and target of our experiment, we can highlight the following:

The Omaliinae-group (represented by Omaliinae and Proteininae) is found sister to the rest of Staphylinidae (including Silphidae). This agrees with the long-term vague consensus perception of the omaliine group as sister to the rest of Staphylinidae that retains many plesiomorphic features of a rove

Chapter 1

beetle ancestor (Thayer & Newton 1995). Such idea, however, has not been rigorously tested before. There are two major clades within the rest of Staphylinidae, one including Silphidae as sister to the oxyteline group of subfamilies; and another including Aleocharinae, Tachyporinae, Steninae, Scydmaeninae, Staphylininae and Paederinae. Although Silphidae were earlier hypothesized as a clade within Staphylinidae, its sister group there was not clear. The molecular study based on only two genes (McKenna *et al.* 2014) suggested part of Tachyporinae as a sister group for Silphidae, a hypothesis that has not been further tested. Our data are consistent with Zhang *et al.* (2018) who also recovered Silphidae as sister to the oxyteline group of Staphylinidae (Figure 6 B & C).

The idea of silphids as a sister lineage of the oxyteline group is not entirely new since earlier morphological works have suggested that Silphidae might be close to Apateticinae (Madge 1979; Hansen 1997b), one of the oxyteline group subfamilies. It is worth serious further morphological study, but the fact that Silphidae are decomposers just like the majority of other oxyteline group members makes their affinity a plausible hypothesis. Sister group relationships between Steninae and Scydmaeninae are also an interesting result of our analysis. Although the former family Scydmaenidae was convincingly shown to be nested inside Staphylinidae (Grebennikov & Newton 2009), its sister group relationships were not clear due to conflicting results of the study by McKenna *et al.* (2014). Interestingly, a recent phylogenetic analysis based on the morphology of both crown and stem groups (Żyła *et al.* 2017) placed Scydmaeninae as a sister group to (Steninae + Euaesthetinae). A result that is consistent with our findings, given that Euaesthetinae were not sampled here. Considering the rather limited taxon sampling in our analysis, the sister group relationships of (Steninae + Euaesthetinae) with *Lordithon* from the polyphyletic Tachyporinae is feasible. On the other hand, *Tachinus*, which is the second sampled lineage of Tachyporinae, is forming a clade with (Staphylininae + Paederinae), which was unexpected. Based on the many mainly morphology-based analyses (Thayer 2016) and the molecular phylogeny of Zhang *et al.* (2018) (Figure 6C), one would expect (Steninae + Euaesthetinae) to be the sister group to (Staphylininae + Paederinae). On the other hand, other molecular-based phylogenies with the hitherto most complete taxon sampling to test this hypothesis (McKenna *et al.* 2014) does not reveal that closely related Steninae and Euaesthetinae are phylogenetically close to (Staphylinidae + Paederinae). It well maybe that our genomic data signal about more complexity here, which needs to be addressed with a more inclusive taxon sample.

Advances and drawbacks of integrate -omic datasets

The database was designed with reference genomes that were intentionally chosen across Insecta (Holometabola) and sampling representative available beetle draft genomes, to target rather conserved genes. Such genes were chosen with the hope that they would also be conserved amongst the members of the Staphylinoidea and able to infer the deep-level relationships. The orthology as-

signment worked well across the various data types where on average very similar numbers were retrieved for both the transcriptomes as well as different low-coverage genomes (Supplementary Materials, Table S5). This is a very promising outlook and upon publication, this ortholog gene set can be used to address similar questions on members of Staphylinidae, for which it was specifically designed, but also Staphylinoidea and Coleoptera with some limitations. Furthermore, it can even be used as the basis to design baits for target capture, if a different sequencing approach is preferred.

The newly generated low-coverage genomes seemed to generally perform well. We found almost equal amounts of orthologs in all genomes and transcriptomes of comparable length. This was not necessarily expected since the genomes were sequenced without prior genome size estimation and knowing that beetle genomes can vary substantially in size even amongst very close relatives. In fact, most genomes were sequenced to such shallow coverage that their genome size could not be estimated with k-mer-based methods (Supplementary Materials, Table S3). Even where estimates could be obtained for the new genomes, with < 100 Mb (except for *Gymnusa* sp.) they appear to be extremely small. The smallest estimated genomes registered in the Animal Genome Size Database are *Tribolium audax* Halstead, 1969 and *Oryzaephilus surinamensis* (Linnaeus, 1758) of 160 Mb (Alvarez-Fuster *et al.* 1991; Sharaf *et al.* 2010). It cannot entirely be excluded that the genomes are of such small size, however, the genome coverage plots (Jellyplots) (Supplementary Materials, Figure S1) did not show a distinctive peak that is necessary for a reliable genome size estimation. Nonetheless, the strategy to blindly sequence extracted DNA without prior genome estimation worked for our purposes, i.e. for the assignment of orthologous sequences from the genome assemblies and subsequent phylogenetic analysis. Otherwise we intended to sequence more based on the same library to increase the coverage.

Once single-copy orthologs had been assigned in their respective sequences of the target species, the extensive pipeline filtered out a total of 74% of them, shortening the total alignment length at the amino-acid level from 2,269,610 sites down to 494,743 sites. Since initially 3,812 OGs had been assigned, this begs the question whether this was really necessary. In fact, (Tan *et al.* 2015) criticize filtering methods of MSAs that mask data blocks in alignments, independent of the algorithms those methods are based on. In their study empirical as well as randomly generated data performed better when unfiltered and for what is worse, in some cases 'optimized' datasets lead to incorrect topologies with higher branch support. This might lead to the conclusion that filtering datasets is generally bad. However, with a high amount of variable quality in the sequences, due to the type of data we used, and no better available filtering methods, as the authors also point out, masking the MSAs seemed more appropriate than inappropriate. An advantage of the specific software we chose is that it uses parametric Monte Carlo resampling within a sliding window, which is softer and less arbitrary than other software that rely on cut-off values defined by the user (Kück *et al.* 2010). If

Chapter 1

alignment masking can lead to a higher proportion of unresolved and well supported but erroneous branches (Tan *et al.* 2015), then inferred phylogenies of masked and unmasked MSAs could be compared. We decided against this strategy for the following reason: Both masked and unmasked datasets could lead to wrong topologies and potentially with high support. So unless there is a phylogenetic tree based on a different data type that these could be compared to, there is no way to choose one topology over the other. In our case, there is no phylogeny that we could reliably compare the datasets to. Instead, we propose to add further analyses, some of which we already applied here, to assess how well the dataset can perform and where its limitations lie.

This provides additional information next to branch support values that might be unreliable. Of course this approach can be applied to both masked and unmasked alignments and every time the dataset has been processed further, i.e. in our case after removing non-informative sites, after keeping only OGs with sequences belonging to specified groups, after removing sequences creating long branches and after removing sequences shorter than 200 bp. Because this study was conducted to get a first notion about how useful it is to combine the various genomic data sources, we refrained from such an extensive analysis at this point. Furthermore, testing of various supermatrices has been done by Misof *et al.* (2014), and the supermatrix where random and non-informative sites had been removed and where they applied the same grouping approach performed better than less treated datasets. Contrary to our pipeline, they removed short sequences right after the ortholog assignment step and did not conduct single-gene tree analyses to remove falsely identified orthologs in OGs (Yang & Smith 2014).

The tests we ran to assess the dataset itself showed that especially the extended supermatrix 2 was not in an optimal condition for a phylogenetic analysis. In general, the different data types were heterogeneous with regard to their composition and the missing data were unevenly distributed. This violates model assumptions during the phylogenetic analysis that usually assume that all sequences have evolved under globally stationary, reversible and homogeneous (SRH) conditions (Song *et al.* 2010). This does not necessarily lead to erroneous phylogenies, but it increases the likelihood and might form clades of unrelated species with similar sequence composition (Jermiin *et al.* 2004). Whether or not compositional heterogeneity was problematic in the conducted analyses should definitely be assessed prior to publishing these results. One option is to conduct matched-pairs tests of homogeneity (Ababneh *et al.* 2006), where sequence pairs are compared to assess whether they evolved under the same evolutionary conditions or not. Sequences with compositional bias can then be assessed further, i.e. it is possible that only a specific part of the sequence introduces inappropriate levels of heterogeneity, or entirely be removed. This test, and the phylogenetic analysis in general, should also be conducted on the nucleotide level. The removal of specific codon positions might lead to a more homogeneous composition of the dataset and hence to more trustworthy results.

The fact that we found twenty different topologies in the phylogenetic analysis of the expanded dataset (supermatrix 2), even though some of them are less likely than others and that we identified 4 rogue taxa in conjunction with lower bootstrap support and without bootstrap convergence, lowers our confidence in some clades of this tree. Since the datasets are identical except for the 40 genes of 24 additional species, it is possible that the extensive gaps in the superalignments of those sequences might have contributed largely to those problems. On the other hand, many parts in the phylogeny are identical to the omic-only dataset (supermatrix 1), which has full support on all but one branch. Non-parametric bootstrapping is a measure of the robustness of the original tree analysis, but it does not provide any information about how single species, e.g. rogue taxa, influenced the tree reconstruction (Holmes 2003). This can be tested with Four-cluster Likelihood Mapping (FcLM) (Strimmer & von Haeseler 1997). This approach assesses quartets of sequence sets and evaluates the suitability of the sequences for phylogenetic reconstruction. It can therefore be used to test different phylogenetic hypotheses for example if more than one topology was found in the ML analysis, as was the case for supermatrix 2, due to the identified rogue taxa. Additionally, FcLM can be used to assess if the composition of a sequence influences the phylogenetic analysis by permuting those sequences (Misof *et al.* 2014). For instance, the large amount of missing data in the sequences of the primer-based species has led to compositional heterogeneity in the dataset, which might have violated SRH conditions. In separate FcLM analyses the empirical sequence of a primer-based taxon can be replaced with a sequence in which all ambiguous amino-acids are permuted without changing the distribution of missing data. This would specifically eliminate the phylogenetic signal in the sequence. If the FcLM analysis of the permuted sequence leads to the same result as the one using the original sequence, then it was the sequence composition that confounded the reconstruction of the tree. Should such sequences be identified then the analysis should be repeated without them.

Prospects for the future

Encouraged by the results of this initial study, we are in the process of repeating the analysis with the following changes and additions:

- 1) Sequence 24 additional genomes using the same protocol described here. The species for the additional genomes will be sampled around poorly resolved and questionable clades, i.e. around the paraphyletic tachyporines, presumably polyphyletic osoriines and within the oxyteline and staphylinine groups. These additional species might enable us to firmly assign the Silphidae as a member of the rove beetles if the sister group relationships can be confidently resolved, and implement respective taxonomic change. Furthermore, we hope to be able to identify if the usually monophyletic staphylinine group is an artifact of the current taxon sampling or a true signal and in conjunction to that, if some Tachyporinae (monophyletic or not)

Chapter 1

are part of the staphylinine group. Sister group relationships of various poorly understood subfamilies such as Oxyporinae, Megalopsidiinae, Trigonurinae, Piestinae and others, not sampled here, should be elucidated, too.

- 2) Repeat the orthology assignment with the fixed database, i.e. in which the sequences of *N. vespilloides* are part of the OGs. Even though overall sequences in the target species were assigned to almost all OGs (3,812), the maximum number per species was 3,480 OGs out of possible 3,822. Adding the sequences of the closest relative to the ortholog set might improve the ortholog assignment during the reciprocal BLAST search.
- 3) Conduct additional analyses with the overall aim to increase our confidence in the topology of the best tree. This would include running the phylogenetic analysis on the nucleotide level in addition to the amino-acid level and trying different partition schemes. Furthermore, to assess the observed compositional heterogeneity in the dataset with matched-pairs tests of homogeneity and FcLM analyses in conjunction with permutation tests.
- 4) Conduct a combined analysis with the addition of morphological characters as a separate partition, with the extant taxa only, as well as with the extant and carefully chosen fossils. The latter dataset can then also be used for Bayesian time-calibration analysis to understand the evolutionary history of the group.

Conclusions

Our results provide the first elucidation of the evolutionary relationships of the family Staphylinidae based on a phylogenomic analysis of a comprehensive molecular dataset and reasonable taxon sampling. A new understanding of the internal relationships of Staphylinidae can be drawn from our results. Major highlights includes:

- 1) members of the monophyletic omaliine group are sister to all remaining rove beetles;
- 2) the oxyteline group is monophyletic including Silphidae, and
- 3) Tachyporinae are polyphyletic and nested within the staphylinine group of subfamilies in more than one place. It confirms the placement of Scydmaeninae as a member of the staphylinine group as currently defined and casts doubt on the monophyly of Osoriinae.

Our analysis is based on a newly designed database of 3,822 orthologs that can easily be re-used in any larger dataset. Although the database was specifically designed for the rove beetles, it includes sequences of all beetle draft genomes and can be used by others to conduct phylogenomic studies of other beetles. Our results indicate a possibility of using heterogeneous genomic data for robust phylogenetic inference given a carefully thought-out bioinformatic pipeline. This study shows a path to an accelerated phylogenetic exploration of species-rich non-model organisms.

Author Contributions

JLK, HE and AS designed and coordinated the study. JLK extracted DNA for the newly generated genomes. MN & PK assembled the transcriptomic reads and under supervision of AP. JLK assembled the low-coverage genomes under the supervision of HE. AP assessed assembled genomes and transcriptomes with BUSCO. HE generated the ortholog gene set. JLK analyzed the data under guidance from HE, beginning with identification of OGs until the end of the pipeline. JLK, AS, AP & HE drafted the manuscript.

Acknowledgements

First of all, we would like to thank all members of the 1KITE-Coleoptera team and CSIRO (Adam Slipinski, Andreas Zwick and David Yeates) for providing us with the unpublished data. We also thank Professors Oliver Niehuis and Tom Gilbert for earnest discussions regarding different options to obtain single-copy orthologs using NGS methods. Furthermore, we thank Robert Waterhouse for his guidance while obtaining the ortholog gene set for *N. vespilloides*. We are grateful for the discussions with members of the Big4-project that helped us to turn problems into opportunities. Finally, we want to greatly thank Karen Meusemann for her always very timely advice regarding the processing and the final analysis of the data and for her comments on this thesis chapter.

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 642241. This material reflects only the author's view, and the Research Executive Agency is not responsible for any use that may be made of the information it contains.

References

- Ababneh, F., Jermini, L. S., Ma, C. & Robinson, J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, 22, 1225–1231.
- Aberer, A. J., Krompass, D. & Stamatakis, A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic biology*, 62, 162–166.
- Ahn, K.-J., Cho, Y.-B., Kim, Y.-H., Yoo, I.-S. & Newton, A. F. 2017. Checklist of the Staphylinidae (Coleoptera) in Korea. *Journal of Asia-Pacific Biodiversity*, 10, 279–336.
- Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS computational biology*, 8, e1002514.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology*, 215, 403–410.
- Alvarez-Fuster, A., Juan, C. & Petitpierre, E. 1991. Genome size in *Tribolium* flour-beetles: inter- and intraspecific variation. *Genetical research*, 58, 1.
- Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- Auguie, B. 2017. gridExtra: Miscellaneous Functions for 'Grid' Graphics. Available online at: <https://CRAN.R-project.org/package=gridExtra>.

Chapter 1

- Bank, S., Sann, M., Mayer, C., Meusemann, K., Donath, A., Podsiadlowski, L., Kozlov, A., Petersen, M., Krogmann, L., Meier, R., *et al.* 2017. Transcriptome and target DNA enrichment sequence data provide new insights into the phylogeny of vespid wasps (Hymenoptera: Aculeata: Vespidae). *Molecular phylogenetics and evolution*, **116**, 213–226.
- Beck, R. M. D., Bininda-Emonds, O. R. P., Cardillo, M., Liu, F.-G. & Purvis, A. 2006. *BMC Evol Biol*, **6**, 93.
- Beutel, R. & Molenda, R. 1997. Comparative morphology of selected larvae of Staphylinioidea (Coleoptera, Polyphaga) with phylogenetic implications. *Zoologischer Anzeiger - A Journal of Comparative Zoology*, **236**, 37–67.
- Beutel, R. G. & Leschen, R. A. B. 2005. Phylogenetic analysis of Staphyliniformia (Coleoptera) based on characters of larvae and adults. *Systematic entomology*, **30**, 510–548.
- Beutel, R. G. & Molenda, R. 1997. Comparative morphology of selected larvae of Staphylinioidea with phylogenetic implications. *Zoologischer Anzeiger*, **236**, 37–67.
- Blaimer, B. B., Brady, S. G., Schultz, T. R., Lloyd, M. W., Fisher, B. L. & Ward, P. S. 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC evolutionary biology*, **15**, 271.
- Blum, P. 1979. Zur Phylogenie und ökologischen Bedeutung der Elytrenreduktion und Abdomenbeweglichkeit der Staphylinidae (Coleoptera). Vergleichend- und funktionsmorphologische Untersuchungen. *Zoologische Jahrbücher. Abteilung für Anatomie und Ontogenie der Tiere*, **102**, 533–582.
- Bocak, L., Barton, C., Crampton-Platt, A., Chesters, D., Ahrens, D. & Vogler, A. P. 2014. Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. *Systematic entomology*, **39**, 97–110.
- Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
- Boussau, B., Walton, Z., Delgado, J. A., Collantes, F., Beani, L., Stewart, I. J., Cameron, S. A., Whitfield, J. B., Johnston, J. S., Holland, P. W. H., *et al.* 2014. Strepsiptera, phylogenomics and the long branch attraction problem. *PloS one*, **9**, e107709.
- Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., Gates, M. W., Kula, R. R. & Brady, S. G. 2017. Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. *Current biology: CB*, **27**, 1019–1025.
- Brunke, A. J., Chatzimanolis, S., Schillhammer, H. & Solodovnikov, A. 2016. Early evolution of the hyper-diverse rove beetle tribe Staphylinini (Coleoptera: Staphylinidae: Staphylininae) and a revision of its higher classification. *Cladistics: the international journal of the Willi Hennig Society*, **32**, 427–451.
- Brunke, A. J., Chatzimanolis, S., Metscher, B. D., Wolf-Schwenninger, K. & Solodovnikov, A. 2017. Dispersal of thermophilic beetles across the intercontinental Arctic forest belt during the early Eocene. *Scientific reports*, **7**, 12972.
- Cai, C., Huang, D., Thayer, M. K. & Newton, A. F. 2012. Glypholomatine Rove Beetles (Coleoptera: Staphylinidae): a Southern Hemisphere Recent Group Recorded from the Middle Jurassic of China. *Journal of the Kansas Entomological Society*, **85**, 239–244.
- Cai, C., Huang, D., Newton, A. F. & Thayer, M. K. 2014. *Mesapatetica aenigmatica*, a New Genus and Species of Rove Beetles (Coleoptera, Staphylinidae) from the Middle Jurassic of China. *Journal of the Kansas Entomological Society*, **87**, 219–224.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. *BMC bioinformatics*, **10**, 421.
- Cameron, S. L., Lambkin, C. L., Barker, S. C. & Whiting, M. F. 2007. A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. *Systematic entomology*, **32**, 40–59.
- Caterino, M. S., Hunt, T. & Vogler, A. P. 2005. On the constitution and phylogeny of Staphyliniformia (Insecta: Coleoptera). *Molecular phylogenetics and evolution*, **34**, 655–672.
- Chani-Posse, M. R., Brunke, A. J., Chatzimanolis, S., Schillhammer, H. & Solodovnikov, A. 2017. Phylogeny of the hyper-diverse rove beetle subtribe Philonthina with implications for classification of the tribe Staphylinini (Coleoptera: Staphylinidae). *Cladistics: the international journal of the Willi Hennig Society*, **34**, 1–40.
- Chatzimanolis, S. 2018. A Review of the Fossil History of Staphylinioidea. In: Betz, O., Irmeler, U. & Klimaszewski, J. (eds) *Biology of Rove Beetles (Staphylinidae): Life History, Evolution, Ecology and Distribution*. Springer International Publishing, Cham, 27–45.
- Chatzimanolis, S., Grimaldi, D. a., Engel, M. S. & Fraser, N. C. 2012. The Earliest Staphyliniform Beetle, from the Late Triassic of Virginia (Coleoptera: Staphylinidae). *American Museum novitates*, **3761**, 1–28.
- Che, L.-H., Zhang, S.-Q., Li, Y., Liang, D., Pang, H., Ślipiński, A. & Zhang, P. 2017. Genome-wide survey of nuclear protein-coding markers for beetle phylogenetics and their application in resolving both deep and shallow-level divergences. *Molecular ecology resources*, **17**, 1342–1358.
- Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS one*, **2**, e383.
- Cho, S., Zwick, A., Regier, J. C., Mitter, C., Cummings, M. P., Yao, J., Du, Z., Zhao, H., Kawahara, A. Y., Weller, S., *et al.* 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for

- the advanced moths and butterflies (Hexapoda: Lepidoptera)? *Systematic biology*, **60**, 782–796.
- Coiffait, H. 1972. Coléoptères Staphylinidae de la Région Paléarctique Occidentale - Sous-familles: Xantholininae et Lep-
totyphlinae. *Publications de la Nouvelle Revue d'Entomologie*, **2**, 115–626.
- Compeau, P. & Pevzner, P. 2015. Profile HMMs for Sequence Alignment. YouTube video available online at:
https://www.youtube.com/watch?v=vO_6xfLwGao; accessed 15/November/2018.
- Cunningham, C. B., Ji, L., Wiberg, R. A. W., Shelton, J., McKinney, E. C., Parker, D. J., Meagher, R. B.,
Benowitz, K. M., Roy-Zokan, E. M., Ritchie, M. G., Brown, S. J., Schmitz, R. J. & Moore, A. J. 2015.
The genome and methylome of a beetle with complex social behavior, *Nicrophorus vespilloides* (coleoptera: Silphidae).
Genome biology and evolution, **7**, 3383–3396.
- Dell'Ampio, E., Meusemann, K., Szucsich, N. U., Peters, R. S., Meyer, B., Borner, J., Petersen, M., Aber-
er, A. J., Stamatakis, A., Walz, M. G., et al. 2014. Decisive data sets in phylogenomics: lessons from studies
on the phylogenetic relationships of primarily wingless insects. *Molecular biology and evolution*, **31**, 239–249.
- Dobler, S. & Müller, J. K. 2000. Resolving phylogeny at the family level by mitochondrial cytochrome oxidase se-
quences: phylogeny of carrion beetles (Coleoptera, Silphidae). *Molecular phylogenetics and evolution*, **15**, 390–402.
- Ebersberger, I., Strauss, S. & von Haeseler, A. 2009. HaMStR: profile hidden markov model based search for
orthologs in ESTs. *BMC evolutionary biology*, **9**, 157.
- Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics Review*, **14**, 755–763.
- Eddy, S. R. 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, **7**, e1002195.
- Elven, H., Bachmann, L. & Gusarov, V. I. 2012. Molecular phylogeny of the Athetini--Lomechusini--Ecitocharini
clade of aleocharine rove beetles (Insecta). *Zoologica scripta*, **41**, 617–636.
- Emms, D. & Kelly, S. 2018. STAG: Species Tree Inference from All Genes. *bioRxiv*, 267914, doi: 10.1101/267914.
- Espeland, M., Breinholt, J., Willmott, K. R., Warren, A. D., Vila, R., Toussaint, E. F. A., Maunsell, S. C.,
Aduse-Poku, K., Talavera, G., Eastwood, R. et al. 2018. A Comprehensive and Dated Phylogenomic Analysis
of Butterflies. *Current biology: CB*, **28**, 770–778.e5.
- Fenn, J. D., Song, H., Cameron, S. L. & Whiting, M. F. 2008. A preliminary mitochondrial genome phylogeny of
Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. *Mo-
lecular phylogenetics and evolution*, **49**, 59–68.
- Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic zoology*, **19**, 99–113.
- Gabaldón, T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome biology*, **9**, 235.
- Ganglbauer, L. 1895. *Die Käfer von Mitteleuropa. Die Käfer Der österreichisch-Ungarischen Monarchie, Deutschlands, Der
Schweiz, Sowie Des Französischen Und Italienischen Alpengebietes*. Druck und Verlag von Carl Gerold's Sohn, 1. Theil:
Staphylinidae, Pselaphidae, 881 pp.
- Geraci, N. S., Spencer Johnston, J., Paul Robinson, J., Wikel, S. K. & Hill, C. A. 2007. Variation in genome
size of argasid and ixodid ticks. *Insect biochemistry and molecular biology*, **37**, 399–408.
- Ge, S.-Q., Beutel, R. G. & Yang, X.-K. 2007. Thoracic morphology of adults of Derodontidae and Nosodendridae
and its phylogenetic implications (Coleoptera). *Systematic entomology*, **32**, 635–667.
- Gilbert, M. T. P., Moore, W., Melchior, L. & Worobey, M. 2007. DNA Extraction from Dry Museum Beetles
without Conferring External Morphological Damage Hofreiter, M. (ed.). *PLoS one*, **2**, e272.
- Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E. & Matsuda, G. 1979. Fitting the Gene
Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences.
Systematic biology, **28**, 132–163.
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L.,
Raychowdhury, R., Zeng, Q., et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a ref-
erence genome. *Nature biotechnology*, **29**, 644–652.
- Grebennikov, V. V. & Newton, A. F. 2009. Good-bye Scydmaenidae, or why the ant-like stone beetles should be-
come megadiverse Staphylinidae *sensu latissimo* (Coleoptera). *European journal of entomology*, **106**, 275–301.
- Grebennikov, V. V. & Newton, A. F. 2012. Detecting the basal dichotomies in the monophylum of carrion and rove
beetles (Insecta: Coleoptera: Silphidae and Staphylinidae) with emphasis on the Oxytelina group of subfamilies. *Ar-
thropod systematics & phylogeny*, **70**(3), 133–165.
- Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies.
Bioinformatics, **29**, 1072–1075.
- Gusarov, V. I. 2018. Phylogeny of the Family Staphylinidae Based on Molecular Data: A Review. In: Betz, O., Irmeler, U. &
Klimaszewski, J. (eds) *Biology of Rove Beetles (Staphylinidae): Life History, Evolution, Ecology and Distribution*. Springer In-
ternational Publishing, Cham, 7–25.
- Haddad, S., Shin, S., Lemmon, A. R., Lemmon, E. M., Svacha, P., Farrell, B., Ślipiński, A., Windsor, D. &
Mckenna, D. D. 2018. Anchored hybrid enrichment provides new insights into the phylogeny and evolution of
longhorned beetles (Cerambycidae): Cerambycidae phylogeny. *Systematic entomology*, **43**, 68–89.
- Hammond, P. M. 1979. Wing-folding Mechanisms of Beetles, with Special Reference to Investigations of Adephagan
Phylogeny (Coleoptera). In: Erwin, T. L., Ball, G. E., Whitehead, D. R. & Halpern, A. L. (eds) *Carabid Beetles: Their Evolu-
tion, Natural History, and Classification*. Springer Netherlands, Dordrecht, 113–180.
- Hansen, M. 1997. Phylogeny and classification of the staphyliniform beetle families (Coleoptera). *Biologiske Skrifter, Det
Kongelige Danske Videnskabernes Selskab*, **48**, 1–339.

Chapter 1

- Hara, Y., Yamaguchi, K., Onimaru, K., Kadota, M., Koyanagi, M., Keeley, S. D., Tatsumi, K., Tanaka, K., Motone, F., Kageyama, Y., *et al.* 2018. Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nature ecology & evolution*, **2**, 1761–1771.
- Hare, E. E. & Johnston, J. S. 2011. Genome Size Determination Using Flow Cytometry of Propidium Iodide-Stained Nuclei. In: Orgogozo, V. & Rockman, M. V. (eds) *Molecular Methods for Evolutionary Genetics*. Humana Press, Totowa, NJ, 3–12.
- He, K., Lin, K., Wang, G. & Li, F. 2016. Genome Sizes of Nine Insect Species Determined by Flow Cytometry and k-mer Analysis. *Frontiers in physiology*, **7**, 569.
- Holmes, S. 2003. Bootstrapping Phylogenetic Trees: Theory and Methods. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **18**, 241–255.
- Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., *et al.* 2015. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome research*, **25**, 445–458.
- Hua, J., Li, M., Dong, P., Cui, Y., Xie, Q. & Bu, W. 2009. Phylogenetic analysis of the true water bugs (Insecta: Hemiptera: Heteroptera: Nepomorpha): evidence from mitochondrial genomes. *BMC evolutionary biology*, **9**, 134.
- Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O. S., Wild, R., Hammond, P. M., Ahrens, D., Balke, M., Caterino, M. S., *et al.* 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science*, **318**, 1913–1916.
- i5K Consortium. 2013. The i5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment. *The Journal of heredity*, **104**, 595–600.
- IUCN. 2018. IUCN Red List version 2018-1, Table 1: Numbers of threatened species by major groups of organisms (1996-2018). Available online at: http://cmsdocs.s3.amazonaws.com/summarystats/2018-1_Summary_Stats_Page_Documents/2018_1_RL_Stats_Table_1.pdf.
- Jałoszyński, P., Brunke, A. J., Yamamoto, S. & Takahashi, Y. 2018. Evolution of Mastigitae: Mesozoic and Cenozoic fossils crucial for reclassification of extant tribes (Coleoptera: Staphylinidae: Scydmaeninae). *Zoological journal of the Linnean Society*, **184**, 623–652.
- Jeffery, N. W. & Gregory, T. R. 2014. Genome size estimates for crustaceans using Feulgen image analysis densitometry of ethanol-preserved tissues: Densitometry Analysis of Ethanol-Preserved Tissues. *Cytometry*, **85**, 862–868.
- Jermiin, L., Ho, S. Y., Ababneh, F., Robinson, J. & Larkum, A. W. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic biology*, **53**, 638–643.
- Junier, T. & Zdobnov, E. M. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, **26**, 1669–1670.
- Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., *et al.* 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, **24**, 1384–1395.
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, **14**, 587–589.
- Katoh, K. & Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**, 772–780.
- Kawahara, A. Y. & Breinholt, J. W. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. *Proceedings. Biological sciences / The Royal Society*, **281**, 20140970.
- Keeling, C. I., Yuen, M. M. S., Liao, N. Y., Roderick Docking, T., Chan, S. K., Taylor, G. A., Palmquist, D. L., Jackman, S. D., Nguyen, A., Li, M., *et al.* 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome biology*, **14**, R27.
- Keller, O., Kollmar, M., Stanke, M. & Waack, S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27**, 757–763.
- King, J. E., Riegler, M., Thomas, R. G. & Spooner-Hart, R. N. 2015. Phylogenetic placement of Australian carrion beetles (Coleoptera: Silphidae): Phylogeny of Australian Silphidae. *Austral Entomology*, **54**, 366–375.
- Koonin, E. V. 2005. Orthologs, Paralogs, and Evolutionary Genomics. *The Annual Review of Genetics*, **39**, 309–338.
- Kristensen, D. M., Wolf, Y. I., Mushegian, a. R. & Koonin, E. V. 2011. Computational methods for Gene Orthology inference. *Briefings in bioinformatics*, **12**, 379–391.
- Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simão, F. A., Pozdnyakov, I. A., Ioannidis, P. & Zdobnov, E. M. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic acids research*, **43**, D250–D256.
- Kück, P. & Meusemann, K. 2010. FASconCAT: Convenient handling of data matrices. *Molecular phylogenetics and evolution*, **56**, 1115–1118.
- Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B. M., Wägele, J. W. & Misof, B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in zoology*, **7**, 10.
- Kukalová-Peck, J. & Lawrence, J. F. 1993. Evolution of the hind wing in Coleoptera. *The Canadian entomologist*, **125**, 181–258.
- Kumar, V., Lammers, F., Bidon, T., Pfenninger, M., Kolter, L., Nilsson, M. A. & Janke, A. 2017. The evolu-

- tionary history of bears is characterized by gene flow across species. *Scientific reports*, **7**, 46487.
- Kusy, D., Motyka, M., Bocek, M., Vogler, A. P. & Bocak, L. 2018a. Genome sequences identify three families of Coleoptera as morphologically derived click beetles (Elateridae). *Scientific reports*, **8**, 17084.
- Kusy, D., Motyka, M., Andujar, C., Bocek, M., Masek, M., Sklenarova, K., Kokas, F., Bocakova, M., Vogler, A. P. & Bocak, L. 2018b. Genome sequencing of *Rhinorhipus* Lawrence exposes an early branch of the Coleoptera. *Frontiers in zoology*, **15**, 21.
- Kypke, J. L., Solodovnikov, A., Brunke, A., Yamamoto, S. & Żyła, D. 2018. The past and the present through phylogenetic analysis: the rove beetle tribe Othiini now and 99 Ma. *Systematic entomology*, **10**, 279.
- Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. 2017. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular biology and evolution*, **34**, 772–773.
- Lawrence, J. F. & Newton, A. F. 1982. Evolution and Classification of Beetles. *Annual review of ecology and systematics*, **13**, 261–290.
- Lawrence, J. F. & Newton, A. F. 1995. *Families and Subfamilies of Coleoptera (With Selected Genera, Notes, References and Data on Family-Group Names)*. Pakaluk, J. & Ślipiński, A. (eds). Muzeum i Instytut Zoologii PAN, Warszawa pp.
- Lawrence, J. F., Ślipiński, A., Seago, A. E., Thayer, M. K., Newton, A. F. & Marvaldi, A. E. 2011. Phylogeny of the Coleoptera Based on Morphological Characters of Adults and Larvae. *Annales zoologici / Polska Akademia Nauk, Instytut Zoologiczny*, **61**, 1–217.
- Letsch, H. & Simon, S. 2013. Insect phylogenomics: new insights on the relationships of lower neopteran orders (Polyneoptera): Phylogenomics of Polyneoptera. *Systematic entomology*, **38**, 783–793.
- Li, H., Shao, R., Song, N., Song, F., Jiang, P., Li, Z. & Cai, W. 2015. Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. *Scientific reports*, **5**, 8527.
- Linard, B., Arribas, P., Andújar, C., Crampton-Platt, A. & Vogler, A. P. 2016. Lessons from genome skimming of arthropod-preserving ethanol. *Molecular ecology resources*, **16**, 1365–1377.
- Linard, B., Crampton-Platt, A., Moriniere, J., Timmermans, M. J., Andujar, C., Arribas, P., Miller, K. E., Lipecki, J., Favreau, E., Hunter, A., et al. 2018. The contribution of mitochondrial metagenomics to largescale data mining and phylogenetic analysis of Coleoptera. *bioRxiv*, 280792, doi: 10.1101/280792.
- Li, X.-Y., Zhou, H.-Z. & Solodovnikov, A. 2013. Five New Species of the Genus *Paederus* From Mainland China, With a Review of the Chinese Fauna of the Subtribe Paederina (Coleoptera: Staphylinidae: Paederinae). *Annals of the Entomological Society of America*, **106**, 562–574.
- López-López, A. & Vogler, A. P. 2017. The mitogenome phylogeny of Adephaga (Coleoptera). *Molecular phylogenetics and evolution*, **114**, 166–174.
- Lopez, P., Casane, D. & Philippe, H. 2002. Heterotachy, an important process of protein evolution. *Molecular biology and evolution*, **19**, 1–7.
- Madge, R. B. 1979. Taxonomic notes on *Apatetica* Westwood (Coleoptera: Silphidae), with a review of the species with black elytra. *Oriental insects*, **13**, 311–321.
- Malé, P.-J. G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A. & Chave, J. 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular ecology resources*, **14**, 966–975.
- Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F. & Durbin, R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature ecology & evolution*, **2**, 1940–1955.
- Marçais, G. & Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
- Maruyama, M. & Parker, J. 2017. Deep-Time Convergence in Rove Beetle Symbionts of Army Ants. *Current biology: CB*, **27**, 920–926.
- Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Liere, K., Wägele, J.-W., et al. 2016. BaitFisher: A software package for multi-species target DNA enrichment probe design. *Molecular biology and evolution*, **33**(7), 1875–86.
- McKenna, D. D., Farrell, B. D., Caterino, M. S., Farnum, C. W., Hawks, D. C., Maddison, D. R., Seago, A. E., Short, A. E. Z., Newton, A. F. & Thayer, M. K. 2014. Phylogeny and evolution of Staphyliniformia and Scarabaeiformia: forest litter as a stepping stone for diversification of nonphytophagous beetles. *Systematic entomology*, **40**, 35–60.
- McKenna, D. D., Wild, A. L., Kanda, K., Bellamy, C. L., Beutel, R. G., Caterino, M. S., Farnum, C. W., Hawks, D. C., Ivie, M. A., Jameson, M. L., et al. 2015. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution: Phylogeny and evolution of Coleoptera (beetles). *Systematic entomology*, **40**, 835–880.
- McKenna, D. D., Scully, E. D., Pauchet, Y., Hoover, K., Kirsch, R., Geib, S. M., Mitchell, R. F., Waterhouse, R. M., Ahn, S.-J., Arsala, D., et al. 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**, 227.
- Misof, B. & Misof, K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence align-

Chapter 1

- ments: a more objective means of data exclusion. *Systematic biology*, **58**, 21–34.
- Misof, B., Meyer, B., von Reumont, B. M., Kück, P., Misof, K. & Meusemann, K. 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC bioinformatics*, **14**, 348.
- Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., *et al.* 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.
- Montiel, E. E., Manrique-Poyato, M. I., Rocha-Sánchez, S. M., López-León, M. D., Cabrero, J., Perfectti, F. & Camacho, J. P. M. 2012. Nucleolus size varies with sex, ploidy and gene dosage in insects. *Physiological entomology*, **37**, 145–152.
- Nabhan, A. R. & Sarkar, I. N. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in bioinformatics*, **13**, 122–134.
- Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., *et al.* 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science*, **347**, 1258522.
- Newton, A. F. 1997. Review of Agyrtidae (Coleoptera), with a new genus and species from New Zealand. *Annales Zoologici (Warszawa)*, **47**, 111–156.
- Newton, A. F. & Thayer, M. K. 1995. Protopselaphinae new subfamily for *Protopselaphus* new genus from Malaysia, with a phylogenetic analysis and review of the Omaliine Group of Staphylinidae including Pselaphidae (Coleoptera). In: Pakaluk, J. & Slipinski, S. A. (eds) *Biology, Phylogeny, and Classification of Coleoptera: Papers Celebrating the 80th Birthday of Roy A. Crowson*. Muzeum i Instytut Zoologii PAN, Warszawa, 219–320.
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, **32**, 268–274.
- Osswald, J., Bachmann, L. & Gusarov, V. I. 2013. Molecular phylogeny of the beetle tribe Oxypodini (Coleoptera: Staphylinidae: Aleocharinae): Molecular phylogeny of the Oxypodini. *Systematic entomology*, **38**, 507–522.
- Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E. & Stamatakis, A. 2010. How many bootstrap replicates are necessary? *Journal of computational biology: a journal of computational molecular cell biology*, **17**, 337–354.
- Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R. S., Podsiadlowski, L., Vasiliakopoulos, A., Zhou, X., Misof, B. & Niehuis, O. 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC bioinformatics*, **18**, 111.
- Peters, R. S., Meusemann, K., Petersen, M., Mayer, C., Wilbrandt, J., Ziesmann, T., Donath, A., Kjer, K. M., Aspöck, U., Aspöck, H., *et al.* 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. *BMC evolutionary biology*, **14**, 52.
- Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., *et al.* 2017. Evolutionary History of the Hymenoptera. *Current biology: CB*, **27**, 1013–1018.
- Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC evolutionary biology*, **5**, 50.
- Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.-Y., Lin, H., Lin, J.-W. & Hackett, K. 2015. The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic acids research*, **43**, D714–D719.
- Pons, J., Ribera, I., Bertranpetit, J. & Balke, M. 2010. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Molecular phylogenetics and evolution*, **56**, 796–807.
- Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M. & Lemmon, A. R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, **526**, 569–573.
- Pryszcz, L. P. & Gabaldón, T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic acids research*, **44**, e113–e113.
- Pyron, R. A. & Wiens, J. J. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular phylogenetics and evolution*, **61**, 543–583.
- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, Available online at: <http://www.R-project.org/>.
- Richards, S., Gibbs, R. A., Weinstock, G. M., Brown, S. J., Denell, R., Beeman, R. W., Gibbs, R., Beeman, R. W., Brown, S. J., Bucher, G., *et al.* 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature*, **452**, 949–955.
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., *et al.* 2010. *De novo* assembly and analysis of RNA-seq data. *Nature methods*, **7**, 909–912.
- Schomann, A. M. & Solodovnikov, A. 2017. Phylogenetic placement of the austral rove beetle genus *Hyperomma* triggers changes in classification of Paederinae (Coleoptera: Staphylinidae). *Zoologica scripta*, **46**, 336–347.
- Sharaf, K., Horová, L., Pavlíček, T., Nevo, E. & Bureš, P. 2010. Genome size and base composition in *Oryzaephilus surinamensis* (Coleoptera: Sylvanidae) and differences between native (feral) and silo pest populations in Israel. *Journal of stored products research*, **46**, 34–37.

- Shin, S., Clarke, D. J., Lemmon, A. R., Moriarty Lemmon, E., Aitken, A. L., Haddad, S., Farrell, B. D., Marvaldi, A. E., Oberprieler, R. G. & McKenna, D. D. 2018. Phylogenomic Data Yield New and Robust Insights into the Phylogeny and Evolution of Weevils. *Molecular biology and evolution*, **35**, 823–836.
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
- Simon, S. & Hadrys, H. 2013. A comparative analysis of complete mitochondrial genomes among Hexapoda. *Molecular phylogenetics and evolution*, **69**, 393–403.
- Slater, G. S. C. & Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, **6**, 31.
- Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research*, **26**, 1134–1144.
- Solodovnikov, A., Yue, Y., Tarasov, S. & Ren, D. 2013. Extinct and extant rove beetles meet in the matrix: Early Cretaceous fossils shed light on the evolution of a hyperdiverse insect lineage (Coleoptera: Staphylinidae: Staphylininae). *Cladistics: the international journal of the Willi Hennig Society*, **29**, 360–403.
- Solodovnikov, A. Y. U. & Newton, A. F. 2005. Phylogenetic placement of Arrowwini trib. n. within the subfamily Staphylininae (Coleoptera: Staphylinidae), with revision of the relict South African genus *Arrowinus* and description of its larva. *Systematic entomology*, **30**, 398–441.
- Song, F., Li, H., Jiang, P., Zhou, X., Liu, J., Sun, C., Vogler, A. P. & Cai, W. 2016. Capturing the Phylogeny of Holometabola with Mitochondrial Genome Data and Bayesian Site-Heterogeneous Mixture Models. *Genome biology and evolution*, **8**, 1411–1426.
- Song, H., Sheffield, N. C., Cameron, S. L., Miller, K. B. & Whiting, M. F. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Systematic entomology*, **35**, 429–448.
- Song, J.-H. & Ahn, K.-J. 2018. Species trees, temporal divergence and historical biogeography of coastal rove beetles (Coleoptera: Staphylinidae) reveal their early Miocene origin and show that most divergence events occurred in the early Pliocene along the Pacific coasts. *Cladistics: the international journal of the Willi Hennig Society*, **34**, 313–332.
- Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
- Strimmer, K. & von Haeseler, A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 6815–6819.
- Suyama, M., Torrents, D. & Bork, P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, **34**, W609–W612.
- Talavera, G. & Vila, R. 2011. What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. *BMC evolutionary biology*, **11**, 315.
- Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M. & Dessimoz, C. 2015. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic biology*, **64**, 778–791.
- Teeling, E. C., Springer, M. S., Madsen, O., Bates, P., O'Brien, S. J. & Murphy, W. J. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, **307**, 580–584.
- Thayer, M. K. 2016. Staphylinidae Latreille, 1802. In: Beutel, R. G. & Leschen, R. A. B. (eds) *Handbook of Zoology. Coleoptera, Beetles - Volume 1: Morphology and Systematics (Archostemata, Adephaga, Myxophaga, Polyphaga Partim)*. Walter de Gruyter GmbH, Berlin/ Boston, 394–442.
- Thomas, G. W. C., Dohmen, E., Hughes, D. S. T., Murali, S. C., Poelchau, M., Glastad, K., Anstead, C. A., Ayoub, N. A., Bellair, M., Binford, G. J., et al. 2018. The Genomic Basis of Arthropod Diversity. *bioRxiv Genomics*, doi: 10.1101/382945.
- Thompson, S. D., Prahalad, S. & Colbert, R. A. 2016. Chapter 5 - Integrative Genomics. In: Petty, R. E., Laxer, R. M., Lindsley, C. B. & Wedderburn, L. R. (eds) *Textbook of Pediatric Rheumatology (Seventh Edition)*. W.B. Saunders, Philadelphia, 43–53.
- Thomsen, P. F., Elias, S., Gilbert, M. T. P., Haile, J., Munch, K., Kuzmina, S., Froese, D. G., Sher, A., Holdaway, R. N. & Willerslev, E. 2009. Non-Destructive Sampling of Ancient Insect DNA. *PloS one*, **4**, e5048.
- Timmermans, M. J. T. N., Barton, C., Haran, J., Ahrens, D., Culverwell, C. L., Ollikainen, A., Dodsworth, S., Foster, P. G., Bocak, L. & Vogler, A. P. 2015. Family-Level Sampling of Mitochondrial Genomes in Coleoptera: Compositional Heterogeneity and Phylogenetics. *Genome biology and evolution*, **8**, 161–175.
- Tsutsui, N. D., Suarez, A. V., Spagna, J. C. & Johnston, J. S. 2008. The evolution of genome size in ants. *BMC evolutionary biology*, **8**, 64.
- Van Dam, M. H., Lam, A. W., Sagata, K., Gewa, B., Laufa, R., Balke, M., Faircloth, B. C. & Riedel, A. 2017. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-weevils. *PloS one*, **12**, e0188044.
- Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J. & Schatz, M. C. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, **33**, 2202–2204.
- Wang, J. 2004. SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic acids research*, **33**, D399–D402.
- Wickham, H. 2011. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, **40**, 1–29.
- Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*, Available online at:

Chapter 1

<http://ggplot2.org>.

Wickham, H. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. Available online at: <https://CRAN.R-project.org/package=tidyverse>.

Wyder, S., Kriventseva, E. V., Schröder, R., Kadowaki, T. & Zdobnov, E. M. 2007. Quantification of ortholog losses in insects and vertebrates. *Genome biology*, **8**, R242.

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

Yamamoto, S. 2016a. The first fossil of dasycerine rove beetle (Coleoptera: Staphylinidae) from Upper Cretaceous Burmese amber: Phylogenetic implications for the omaliine group subfamilies. *Cretaceous Research*, **58**, 63–68.

Yamamoto, S. 2016b. The oldest tachyporine rove beetle in amber (Coleoptera, Staphylinidae): A new genus and species from Upper Cretaceous Burmese amber. *Cretaceous Research*, **65**, 163–171.

Yamamoto, S. & Maruyama, M. 2018. Phylogeny of the rove beetle tribe Gymnusini sensu n. (Coleoptera: Staphylinidae: Aleocharinae): implications for the early branching events of the subfamily: Phylogeny of Gymnusini rove beetles. *Systematic entomology*, **43**, 183–199.

Yang, Y. & Smith, S. A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular biology and evolution*, **31**, 3081–3092.

Ye, C., Ma, Z., Cannon, C. H., Pop, M. & Yu, D. W. 2012. Exploiting sparseness in *de novo* genome assembly. *BMC bioinformatics*, **13**, S1.

Young, A. D., Lemmon, A. R., Skevington, J. H., Mengual, X., Ståhls, G., Reemer, M., Jordaens, K., Kelso, S., Lemmon, E. M., Hauser, M., De Meyer, M., Misof, B. & Wiegmann, B. M. 2016. Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC evolutionary biology*, **16**, 143.

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

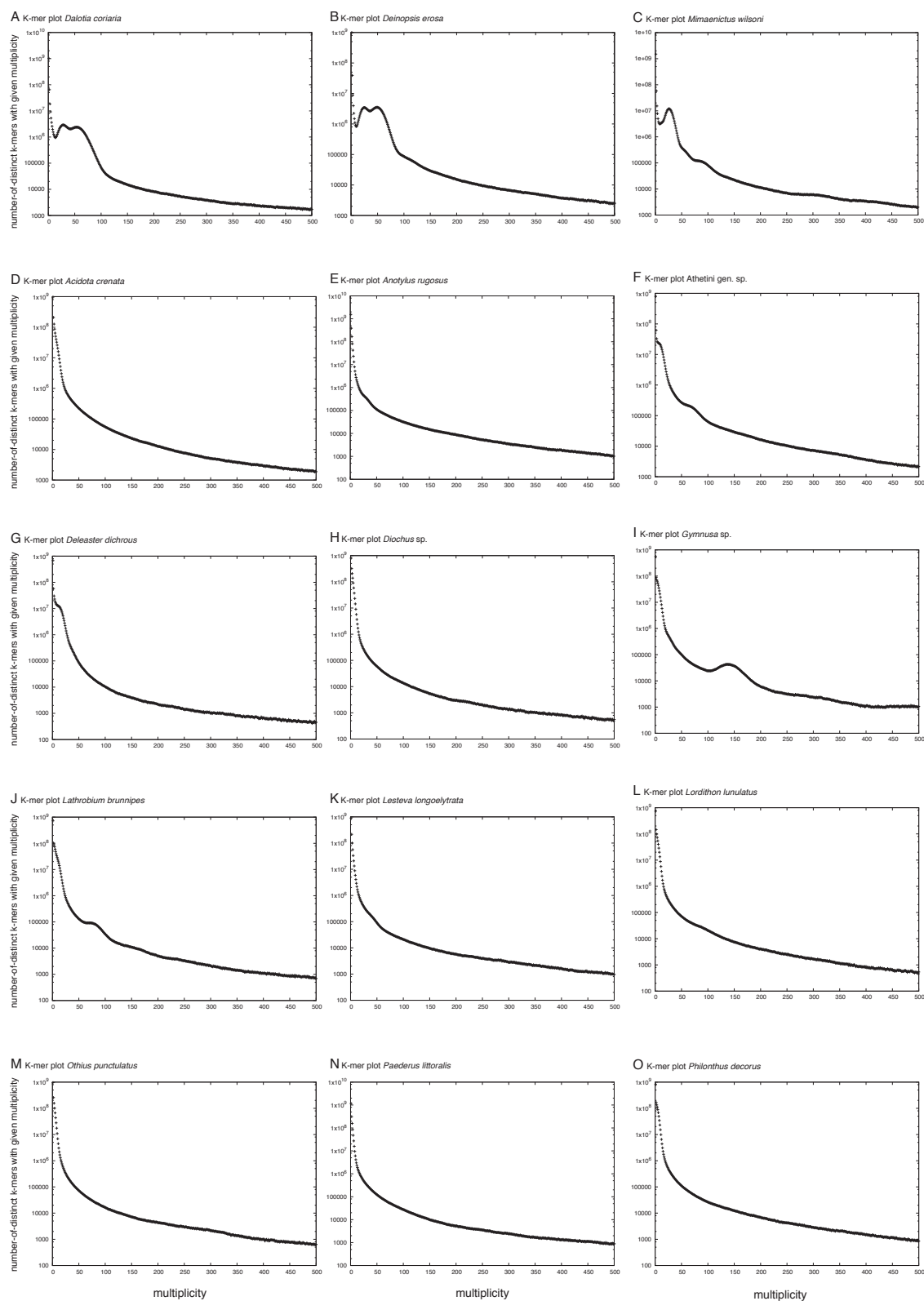
Zhang, S. Q., Che, L. H., Li, Y., Dan, L., Pang, H., Ślipiński, A. & Zhang, P. 2018. Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nature communications*, **9**, doi: 10.1038/s41467-017-02644-4.

Zhang, X. & Zhou, H. 2018. Aedeagus evolution promotes speciation? A primary pattern in rove beetle phylogeny. *Zoological Systematics*, **43**, 125–138.

Zhang, X. & Zhou, H.-Z. 2013. How old are the rove beetles (Insecta: Coleoptera: Staphylinidae) and their lineages? Seeking an answer with DNA. *Zoological science*, **30**, 490–501.

Żyła, D., Yamamoto, S., Wolf-Schwenninger, K. & Solodovnikov, A. 2017. Cretaceous origin of the unique prey-capture apparatus in mega-diverse genus: stem lineage of Steninae rove beetles discovered in Burmese amber. *Scientific reports*, **7**, 45904.

Supplementary Material



Chapter 1

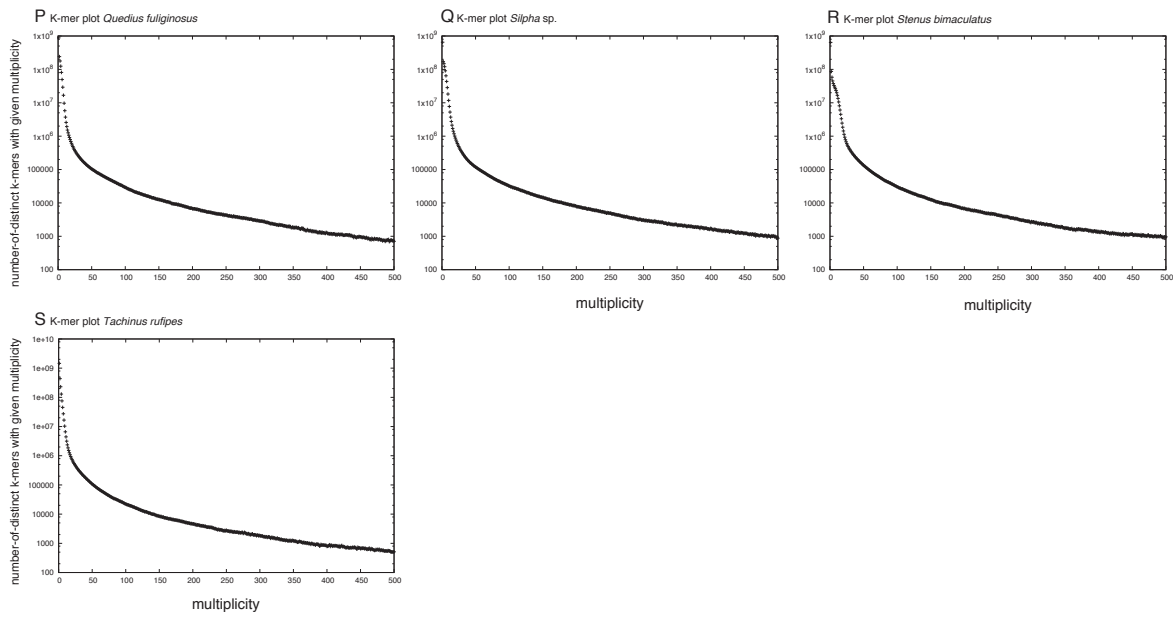
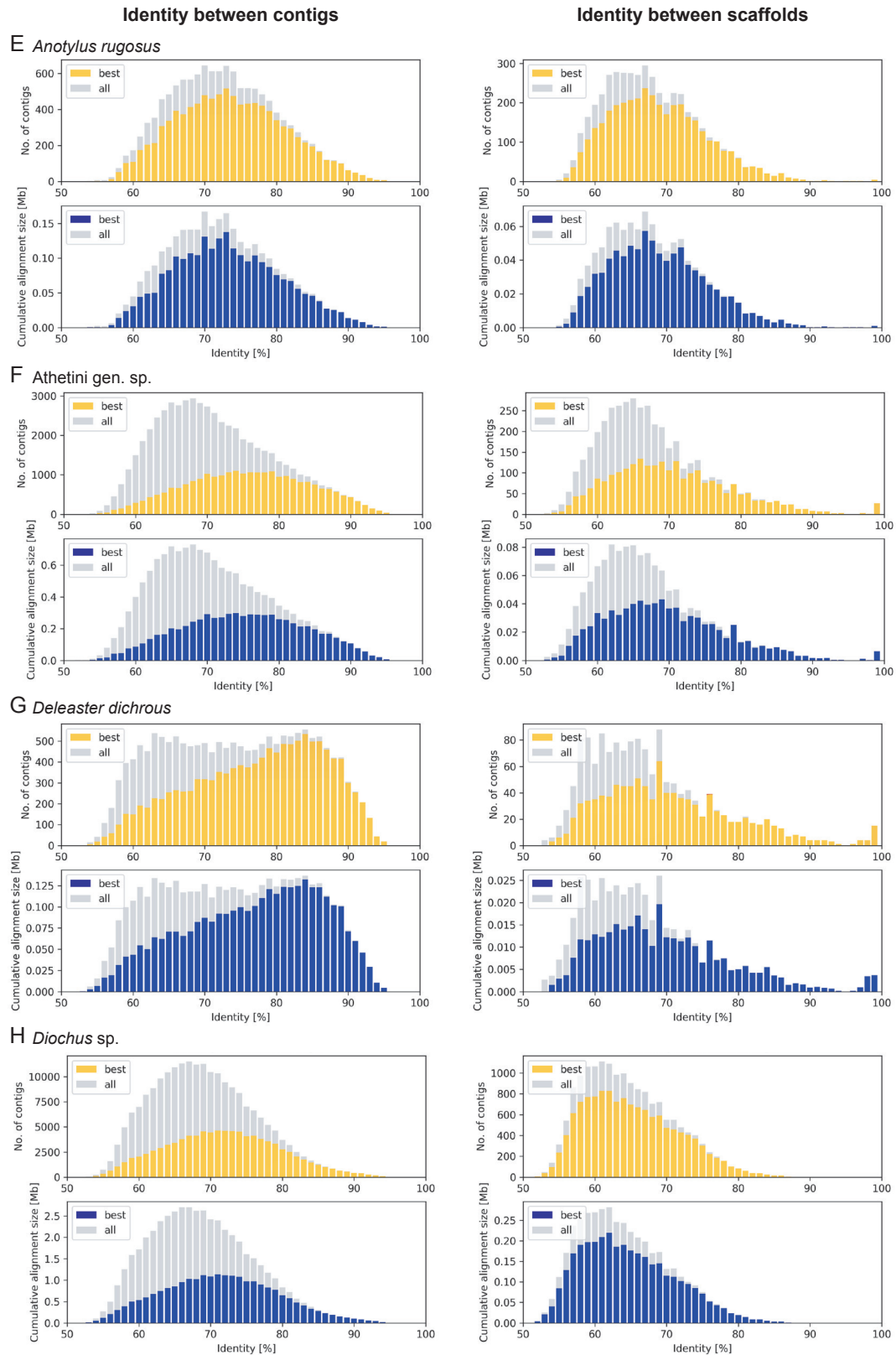
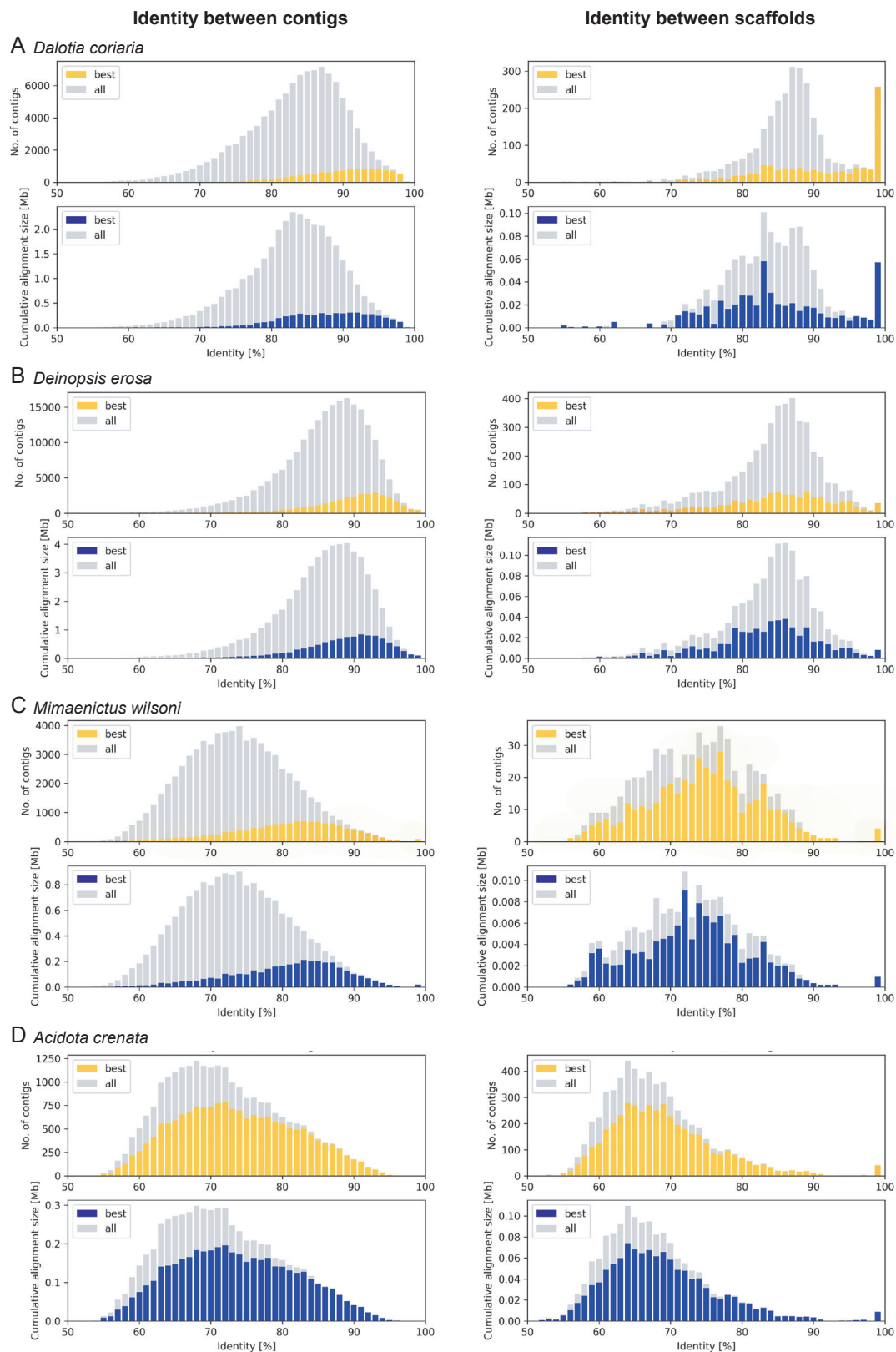
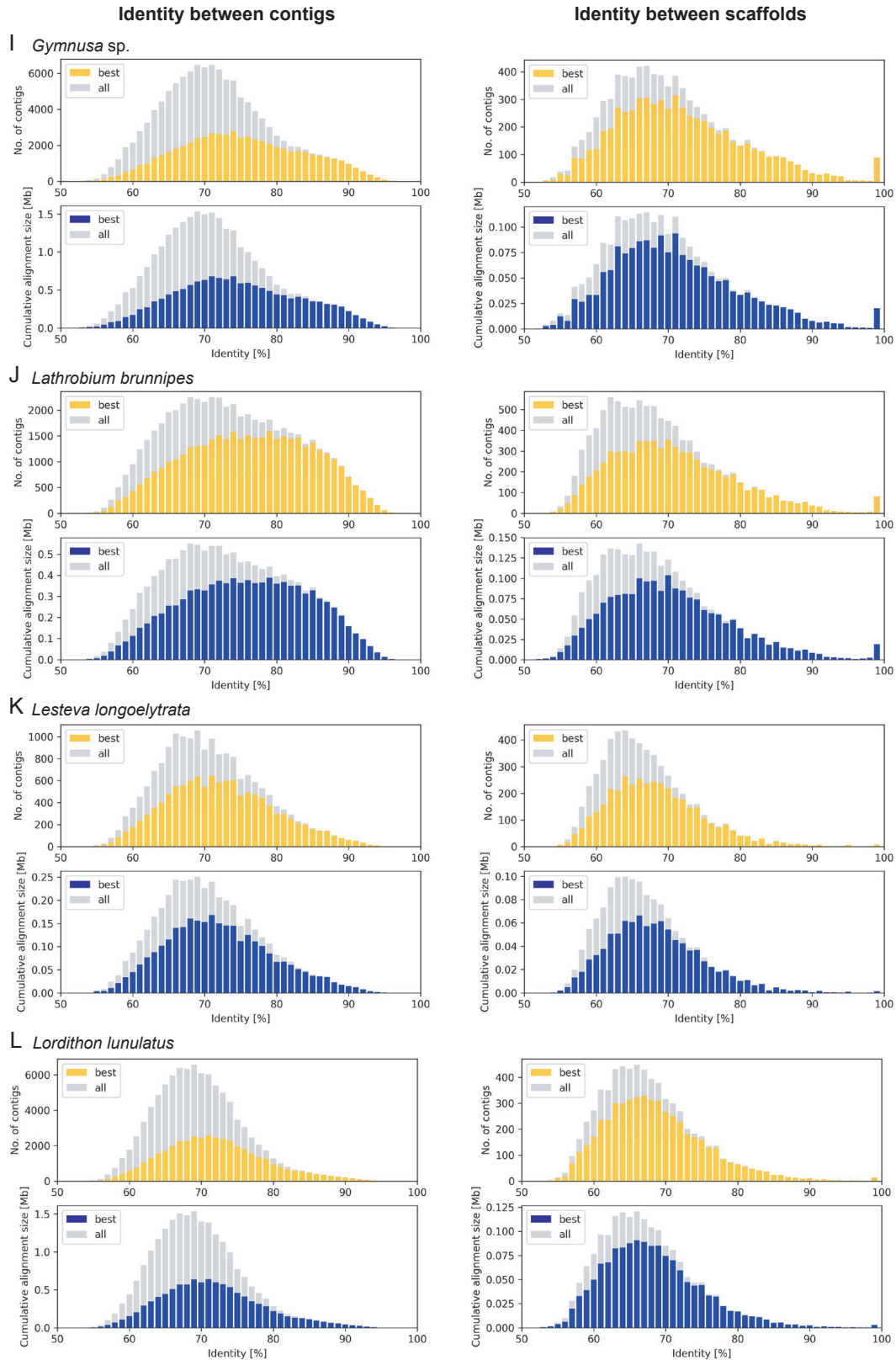


Figure S1. K-mer frequency plots generated by Jellyfish on the raw sequence reads for A-C: aleocharine genomes (published by J. Parker); and D-S: sixteen newly generated genomes. Peaks represent the mean coverage and used as input to estimate individual genome sizes.

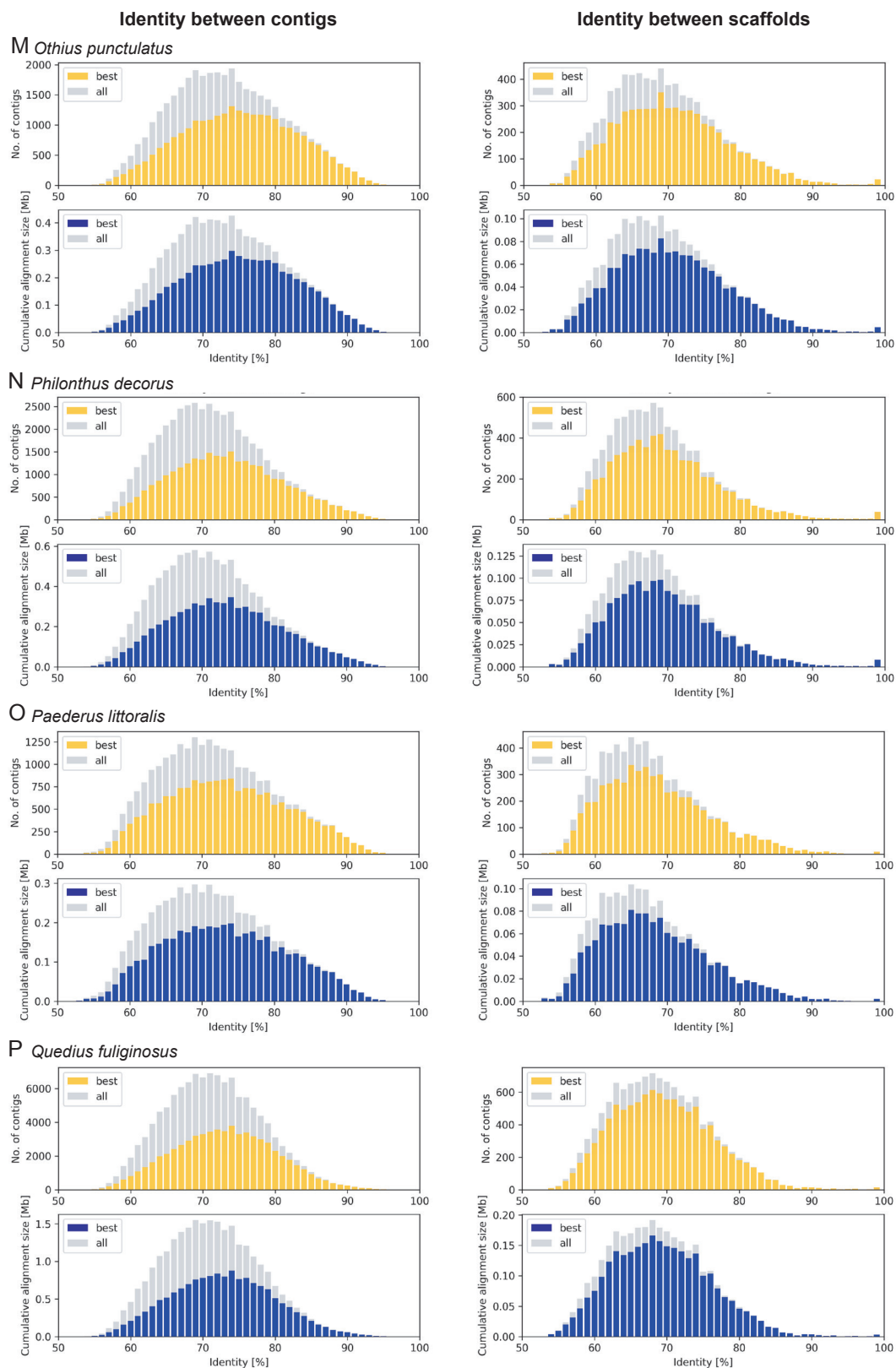


Chapter 1





Chapter 1



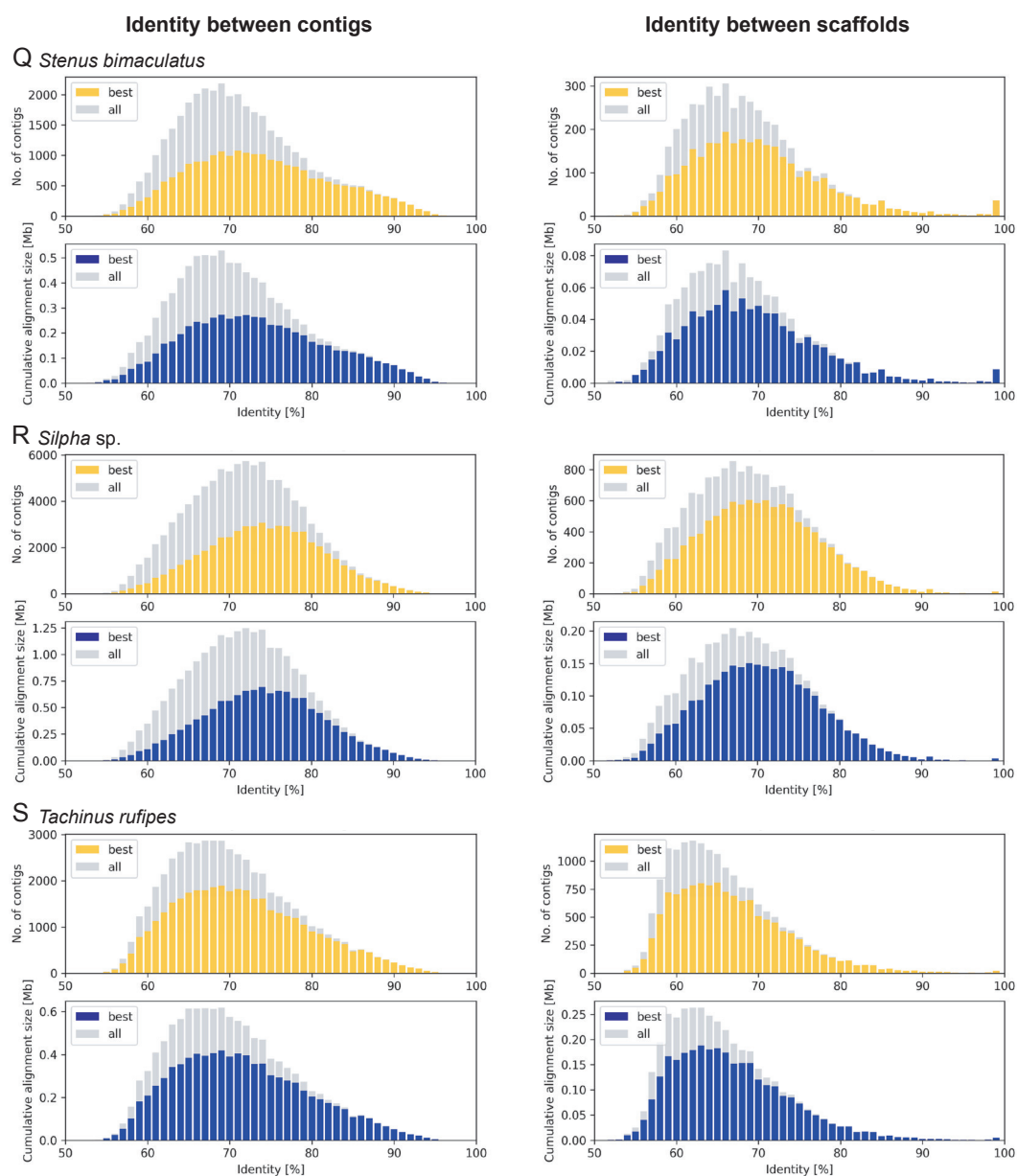


Figure S2. Identity graphs during the reduction step of contigs (left hand side) and scaffolds (right hand side) to generate the final assemblies for species A-S using Redundans. Identity (%) of a contig/ scaffold with the assembly. Upper graphs: frequency; Lower graphs: cumulative alignment size in Megabases (Mb); Grey: entirety that a contig/ scaffold was aligned to a location in the assembly; Yellow/ blue: 'best' matching contig/ scaffold that will be retained before reduction step.

Chapter 1

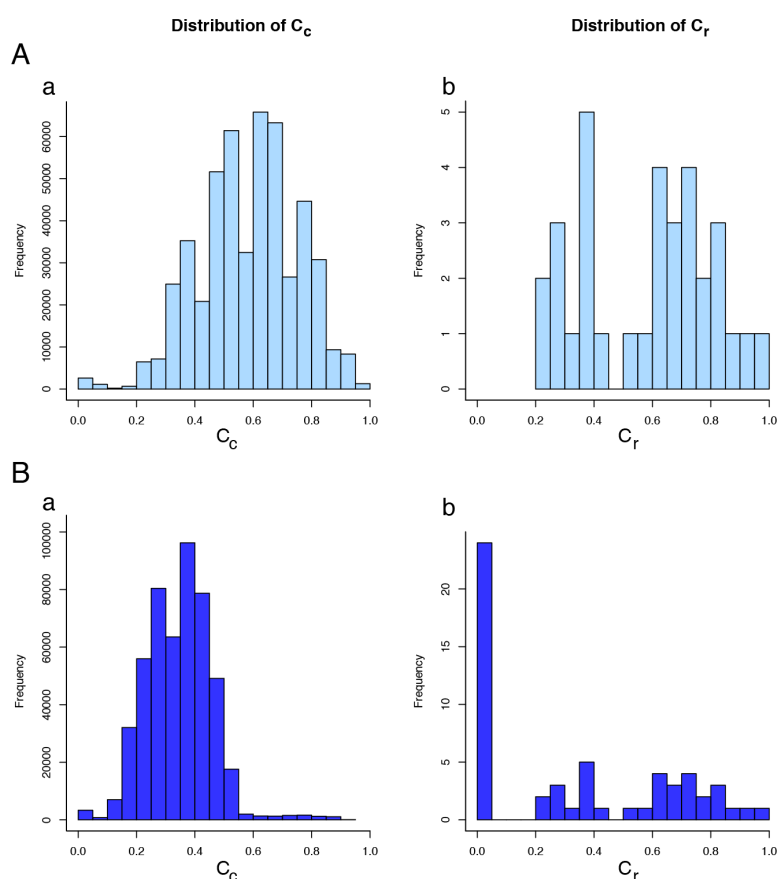


Figure S4. Distribution of completeness (C)
scores for individual sequences in super-alignments of A: supermatrix 1, i.e. the omic-only dataset (light blue); and B: supermatrix 2, i.e. the expanded = omic & primer-based dataset (dark blue); a: scores calculated by columns (number of unambiguous characters in the column / number of sequences); b: scores calculated by rows (number of unambiguous characters in the sequence / alignment length).

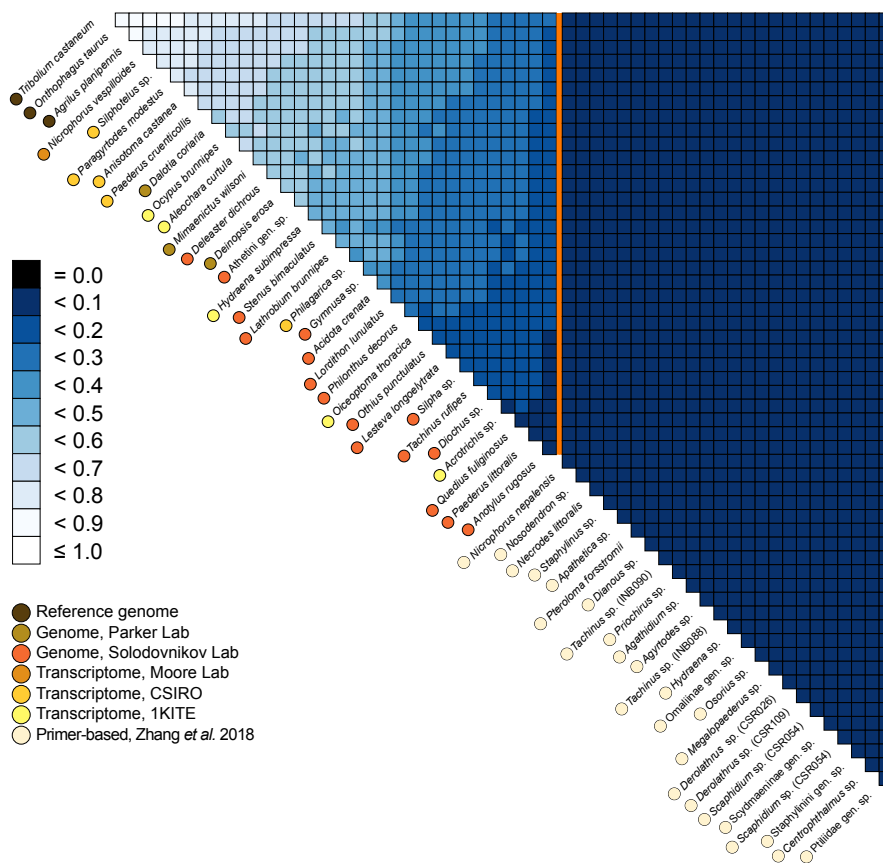
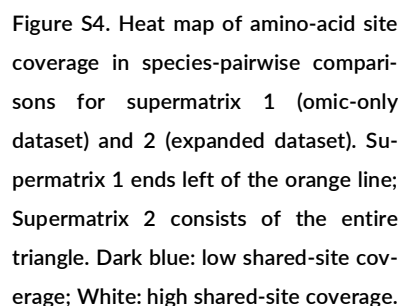


Table S1. FastQC output of WGS raw sequence reads (before assembly), and Quast output after assembling untrimmed reads with SparseAssembler and Redundans of *Diochus* sp. and *Gymnusa* sp.

	<i>Diochus</i> sp.	<i>Gymnusa</i> sp.
FastQC output		
Total Sequences (bp)	23,263,798	21,805,858
# Sequences flagged as poor quality	0	0
Sequence length (bp)	150	150
GC (%)	34	33
Quast output		
# contigs (>= 0 bp)	2,789,578	947,150
# contigs (>= 200 bp)	417,955	309,617
# contigs (>= 500 bp)	31,543	158,313
# contigs (>= 1000 bp)	1,629	48,003
# contigs (>= 5000 bp)	11	223
# contigs (>= 10000 bp)	1	78
# contigs (>= 25000 bp)	0	30
# contigs (>= 50000 bp)	0	14
Total length (>= 0 bp)	396,457,588	272,932,513
Total length (>= 200 bp)	135,958,692	201,712,074
Total length (>= 500 bp)	20,879,055	150,216,536
Total length (>= 1000 bp)	2,298,993	74,616,400
Total length (>= 5000 bp)	78,041	3,501,912
Total length (>= 10000 bp)	12,651	2,540,495
Total length (>= 25000 bp)	0	1,798,513
Total length (>= 50000 bp)	0	1,236,485
# contigs	31,543	158,313
Largest contig	12,651	210,997
Total length	20,879,055	150,216,536
GC (%)	35.25	32.49
N50	626	995
N75	552	710
L50	12,663	48,497
L75	21,586	93,449
# N's per 100 kbp	532	1,245

Table S 2. Gene identifier (ID) in correspondence to the global eucaryote ortholog group identifier (EOG ID) of each reference species. Gene descriptions are provided for *N. vespilloides*. Accessible online:

<https://drive.google.com/open?id=1PDA7OnBJieoEMZWYxAA-n5-5-TuHbCHf>

Chapter 1

Table S3. Summary statistics reported by various software in the process of assembling 19 low-coverage genomes generated using the whole genome re-sequencing approach. Listed are the quality checks of the raw sequence reads before and after trimming (FastQC output); genome size estimates generated by Jellyfish and GenomeScope; assembly statistics generated by SparseAssembler; and summary statistics assessing the quality of the final assemblies (after removing redundant sequences with Redundans).

	<i>Dalotia coriaria</i>	<i>Deinopsis erosa</i>	<i>Mimaenictus wilsoni</i>	<i>Acidota crenata</i>	<i>Anotylus rugosus</i>
FastQC output (raw data)					
Total Sequences (bp)	87,210,670	58,037,941	57,519,823	34,199,519	23,159,772
# Sequences flagged as poor quality	0	0	0	0	0
Sequence length (bp)	151	151	151	150	150
GC (%)	32 / 33	29	39 / 40	31	36
Jellyfish output (k=31)					
Total # nucleotides	13,168,811,170	8,763,729,091	8,685,493,273	no peak in Jellyplot	no peak in Jellyplot
Estimated coverage	34.94	34.94	32.45	-	-
Estimated genome size (Mb)	376	251	268	-	-
Genome scope output (k=31)					
Heterozygosity (min / max %)	1.13 / 1.14	0.92 / 0.93	0.25	-	-
Genome haploid length (min / max Mb)	114	152	257 / 258	-	-
Model fit (min / max %)	96.13 / 97.30	96.7 / 98.9	95.8 / 99.3	-	-
Read error rate (min / max %)	0.61	0.45	0.73	-	-
FastQC output (raw data trimmed)					
Total Sequences (bp)	76,634,155	50,223,449	43,730,931	29,240,404	19,047,418
# Sequences flagged as poor quality	0	0	0	0	0
Sequence length (bp)	40-146	40-146	40-146	40-145	40-145
GC (%)	31 / 32	29	38	31	35
SparseAssembler					
Average length (bp)				136	133
Total # nucleotides				8568391011	5546645705
Coverage				28	18
Estimated # of k-mers				2052625585	1419495231
Total # nodes				75610811	84146360
Total # edges				170300612	161933302
# contigs				12087872	11129268
Quast output*					
# contigs (>= 0 bp)	9,838	48,152	60,677	285084	249,624
# contigs (>= 1000 bp)	5,928	30,617	34,362	39,399	4,034
# contigs (>= 5000 bp)	3,516	6,951	8,350	497	178
# contigs (>= 10000 bp)	2,499	2,111	3,539	38	71
# contigs (>= 25000 bp)	1,211	317	688	1	11
# contigs (>= 50000 bp)	505	78	90	0	1
Total length (>= 0 bp)	103,583,837	137,688,861	172,310,691	175,524,596	85,658,984
Total length (>= 1000 bp)	102,152,293	130,333,675	160,607,482	64,515,096	7,647,322
Total length (>= 5000 bp)	95,975,530	72,478,055	102,381,239	3,438,652	1,976,037
Total length (>= 10000 bp)	88,663,821	39,334,318	68,809,468	515,234	1,241,403
Total length (>= 25000 bp)	67,817,606	13,915,785	25,794,075	48,248	377,560
Total length (>= 50000 bp)	43,378,309	5,913,219	5,899,928	0	54,583
# contigs	6,793	36,079	43,591	131,677	38,222
Largest contig	275,233	153,054	116,605	48,248	54,583
Total length	102,773,260	134,384,034	167,484,779	128,842,277	29,360,507
GC (%)	37.54	29.24	37.24	32	34.03
N50	39,630	5,553	7,390	970	680
N75	18,061	2,838	2,965	699	573
L50	686	5,947	5,287	38,911	10,625
L75	1,647	14,553	14,428	76,446	20,271
# N's per 100 kbp	3.92	5.52	122.91	788	2,318.21
* All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g. '# contigs (>= 0 bp)' and 'Total length (>= 0 bp)' include all contigs).					
	<i>Athetini</i> gen. sp.	<i>Deleaster dichrous</i>	<i>Dioclius</i> sp.	<i>Gymnusa</i> sp.	<i>Lathrobium brunripes</i>
FastQC output (raw data)					
Total Sequences (bp)	40,200,886	26,362,739	23,263,789	21,805,858	25,074,630
# Sequences flagged as poor quality	0	0	0	0	0
Sequence length (bp)	150	150	150	150	150
GC (%)	33 / 34	31 / 32	34	33	33 / 34
Jellyfish output (k=31)					
Total # nucleotides	6,030,132,900	no peak in Jellyplot	no peak in Jellyplot	3,270,878,700	3,761,194,500
Estimated coverage	90	-	-	172.5	80
Estimated genome size (Mb)	67	-	-	190	47
Genome scope output (k=31)					
Heterozygosity (min / max %)	2.17 / 2.32	-	-	-	-
Genome haploid length (min / max Mb)	35 / 36	-	-	models did not converge	models did not converge
Model fit (min / max %)	81.32 / 93.42	-	-	-	-
Read error rate (min / max %)	3.19	-	-	-	-
FastQC output (raw data trimmed)					
Total Sequences (bp)	33,770,275	21,509,820	20,931,551	18,996,723	21,102,720
# Sequences flagged as poor quality	0	0	0	0	0
Sequence length (bp)	40-140	40-140	40-145	40-145	40-145
GC (%)	33	31	34	30 / 33	33
SparseAssembler					
Average length (bp)	128	128	128	131	134
Total # nucleotides	9386148787	6062137493	6062137493	5292772594	6094929927
Coverage	31	20	20	17	20
Estimated # of k-mers	2206083663	1529731395	1529731395	1352599498	1546587293
Total # nodes	37735707	30957653	30957653	46783126	62738417
Total # edges	90088496	67872641	67872641	97118662	136984511
# contigs	6939473	2868552	2868552	3869530	6727382
Quast output					
# contigs (>= 0 bp)	80,125	73,537	704,002	256,588	209,065
# contigs (>= 1000 bp)	40,765	37,063	8,951	70,113	62,126
# contigs (>= 5000 bp)	4,741	4,840	21	1,463	2,357
# contigs (>= 10000 bp)	864	1,262	0	135	202
# contigs (>= 25000 bp)	55	95	0	28	7
# contigs (>= 50000 bp)	13	4	0	12	1
Total length (>= 0 bp)	138,606,712	128,779,919	253,125,701	220,789,669	204,548,036
Total length (>= 1000 bp)	120,189,791	112,638,286	11,406,198	139,911,379	141,995,000
Total length (>= 5000 bp)	38,810,354	43,894,029	128,346	11,415,714	16,345,720
Total length (>= 10000 bp)	13,029,052	19,768,843	0	3,121,329	2,913,102
Total length (>= 25000 bp)	2,306,016	3,083,573	0	1,627,635	281,002
Total length (>= 50000 bp)	975,683	225,815	0	1,076,012	104,794
# contigs	56,197	49,953	109,843	133,400	108,837
Largest contig	119,530	59,540	9,176	187,699	104,794
Total length	131,360,738	122,023,013	76839930	184819489	174,874,669
GC (%)	37.55	34.09	35	33	33.67
N50	3,081	3,167	678	1,694	2,099
N75	1,804	1,709	573	1,016	1,177
L50	11,919	9,516	42,896	33,555	25,698
L75	25,762	22,734	73,850	68,827	52,675
# N's per 100 kbp	162.74	354.9	12.32	383.14	428.56
* All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g. '# contigs (>= 0 bp)' and 'Total length (>= 0 bp)' include all contigs).					

	<i>Lesteva longoelytrata</i>	<i>Lordithon lunulatus</i>	<i>Othius punctulatus</i>	<i>Paederus littoralis</i>	<i>Philonthus decorus</i>
FastQC output (raw data)					
Total Sequences (bp)	16,553,050	20,396,214	22,205,266	20,453,976	24,318,503
# Sequences flagged as poor quality	0	0	0	0	0
Sequence length (bp)	150	150	150	150	150
GC (%)	34 / 35	39 / 40	33 / 34	36	31
Jellyfish output (k=31)					
Total # nucleotides	2,482,957,500	-	-	-	-
Estimated coverage	105	-	-	-	-
Estimated genome size (Mb)	24	-	-	-	-
Genome scope output (k=31)					
Heterozygosity (min / max %)	2.50 / 2.76	-	-	-	-
Genome haploid length (min / max Mb)	21 / 23	-	-	-	-
Model fit (min / max %)	78.63 / 92.95	-	-	-	-
Read error rate (min / max %)	4.38	-	-	-	-
FastQC output (raw data trimmed)					
Total Sequences (bp)	13,360,776	15,269,126	18,392,915	16,691,691	20,160,522
# Sequences flagged as poor quality	0	0	0	0	0
Sequence length (bp)	40-140	40-140	40-145	40-145	40-145
GC (%)	34	37	33	35	31
SparseAssembler					
Average length (bp)	131	119	133	134	134
Total # nucleotides	3868087736	4181609618	5336921564	4889338650	5858521731
Coverage	12	13	17	16	19
Estimated # of k-mers	1043011410	1080819330	1357301051	1297269834	1484257928
Total # nodes	49113355	47391461	72747332	73401012	76087345
Total # edges	99777262	98074793	146698943	146483287	165702199
# contigs	7507444	4656217	7969977	10082165	10532830
Quast output					
# contigs (>= 0 bp)	157,796	316,997	421,232	300,543	445,789
# contigs (>= 1000 bp)	7,786	40,009	27,195	7,886	40,974
# contigs (>= 5000 bp)	235	49	97	88	128
# contigs (>= 10000 bp)	74	2	51	4	20
# contigs (>= 25000 bp)	3	0	23	1	2
# contigs (>= 50000 bp)	0	0	6	0	0
Total length (>= 0 bp)	68,698,919	180,117,544	191,800,287	111,694,690	228,458,367
Total length (>= 1000 bp)	13,484,047	60,302,478	39,736,364	12,260,494	58,126,490
Total length (>= 5000 bp)	2,215,127	296,477	1,817,430	579,515	1,022,600
Total length (>= 10000 bp)	1,069,989	23,134	1,510,816	59,917	310,128
Total length (>= 25000 bp)	82,695	0	1,053,168	26,239	57,864
Total length (>= 50000 bp)	0	0	457,841	0	0
# contigs	44,330	134,257	130,581	61,086	171,564
Largest contig	29,230	11,840	120,487	26,239	32,144
Total length	37,572,989	125,602,883	109,390,188	46,636,526	147,972,051
GC (%)	34.17	35.46	34.81	35.21	32.34
N50	790	948	815	702	852
N75	612	687	626	581	651
L50	11,996	42,298	40,994	18,563	56,124
L75	23,810	80,108	76,287	34,197	103,036
# N's per 100 kbp	1,721.62	484.72	1,057.26	1,868.42	828.09

* All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g. '# contigs (>= 0 bp)' and 'Total length (>= 0 bp)' include all contigs).

	<i>Quedius fuliginosus</i>	<i>Silpha sp.</i>	<i>Stenus bimaculatus</i>	<i>Tachinus rufipes</i>
FastQC output (raw data)				
Total Sequences (bp)	21,835,452	26,143,518	23,025,792	24,657,085
# Sequences flagged as poor quality	0	0	0	0
Sequence length (bp)	150	150	150	150
GC (%)	33 / 34	37	28	30
Jellyfish output (k=31)				
Total # nucleotides	-	-	-	-
Estimated coverage	-	-	-	-
Estimated genome size (Mb)	-	-	-	-
Genome scope output (k=31)				
Heterozygosity (min / max %)	-	-	-	-
Genome haploid length (min / max Mb)	-	-	-	-
Model fit (min / max %)	-	-	-	-
Read error rate (min / max %)	-	-	-	-
FastQC output (raw data trimmed)				
Total Sequences (bp)	17,828,495	21,345,876	19,209,246	20,681,634
# Sequences flagged as poor quality	0	0	0	0
Sequence length (bp)	40-145	40-140	40-140	40-145
GC (%)	33	37 / 36	28	30
SparseAssembler				
Average length (bp)	133	129	127	134
Total # nucleotides	5187337302	6048324326	5299402712	5985525993
Coverage	17	20	17	19
Estimated # of k-mers	1357301051	1532649599	1342752149	1484257928
Total # nodes	77162200	72983888	43470512	113123506
Total # edges	154582878	157308153	95289571	214811459
# contigs	7185123	9709005	4707963	11397571
Quast output				
# contigs (>= 0 bp)	569,759	365,668	148,100	523,375
# contigs (>= 1000 bp)	26,155	21,683	55,385	32,075
# contigs (>= 5000 bp)	113	10	2,068	101
# contigs (>= 10000 bp)	49	0	124	16
# contigs (>= 25000 bp)	7	0	0	0
# contigs (>= 50000 bp)	0	0	0	0
Total length (>= 0 bp)	244,915,312	162,566,088	165,909,862	228,872,582
Total length (>= 1000 bp)	36,202,324	29,713,260	123,616,118	45,741,370
Total length (>= 5000 bp)	1,392,379	64,458	13,827,104	747,510
Total length (>= 10000 bp)	959,478	0	1,551,054	204,805
Total length (>= 25000 bp)	279,211	0	0	0
Total length (>= 50000 bp)	0	0	0	0
# contigs	168,255	110,996	90,269	152,738
Largest contig	49,969	9,743	23,121	19,919
Total length	130,828,537	89,777,993	148,555,490	126,810,984
GC (%)	33.31	33.21	30.72	28.96
N50	744	795	2,063	807
N75	603	618	1,223	622
L50	57,314	38,075	22,515	47,266
L75	102,859	69,146	45,310	87,229
# N's per 100 kbp	831.75	544.38	406.12	1,469.77

* All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g. '# contigs (>= 0 bp)' and 'Total length (>= 0 bp)' include all contigs).

Chapter 1

Table S4. BUSCO assessment of the assembled transcriptomes (T) and genomes (G) in comparison to 2,442 Endopterygota genes. Listed are the number (#) of paired-end (PE) reads, individual read length, complete, single-copy, duplicated, fragmented, and missing BUSCOs.

Data type	Species	# PE reads	Read length	BUSCO complete	BUSCO single-copy	BUSCO duplicated	BUSCO fragmented	BUSCO missing
T	<i>Acrotrichis</i> sp.	4,645,862	150	935	519	416	531	976
T	<i>Aleochara curtula</i>	9,950,251	150	1,891	1,303	588	355	196
T	<i>Anisotoma castanea</i>	43,121,683	150	2,239	1,359	880	149	54
T	<i>Hydraena subimpressa</i>	6,293,432	150	1,582	1,087	495	553	307
T	<i>Nicrophorus vespilloides</i>	713,929,590	90	2,223	1,085	1,138	175	44
T	<i>Ocyopus brunnipes</i>	8,572,322	150	1,789	1,244	545	385	268
T	<i>Oiceoptoma thoracica</i>	8,762,442	150	670	543	127	768	1,004
T	<i>Paederus cruenticollis</i>	43,722,256	150	1,908	612	1,296	436	98
T	<i>Paragyrtonotus modestus</i>	42,964,840	150	2,200	1,357	843	156	86
T	<i>Philagrica</i> sp.	43,239,611	150	2,023	973	1,050	292	127
T	<i>Silphotus</i> sp.	39,385,834	150	2,139	1,337	802	150	153
G	<i>Acidota crenata</i>	34,199,519	150	913	908	5	885	644
G	<i>Anotylus rugosus</i>	23,159,772	150	60	50	10	129	2,253
G	Athetini gen. sp.	40,200,886	150	1,829	1,823	6	461	152
G	<i>Dalotia coriaria</i>	87,210,670	150	2,332	2,322	10	74	36
G	<i>Deinopsis erosa</i>	58,037,941	150	1,945	1,933	12	362	135
G	<i>Deleaster dichrous</i>	26,362,739	150	2,063	2,047	16	275	104
G	<i>Diochus</i> sp.	23,263,789	150	89	89	0	211	2,142
G	<i>Gymnusa</i> sp.	21,805,858	150	1,054	1,046	8	853	535
G	<i>Lathrobium brunnipes</i>	25,074,630	150	1,312	1,302	10	738	392
G	<i>Lesteva longoelytrata</i>	16,553,050	150	231	209	22	477	1,734
G	<i>Lordithon lunulatus</i>	20,396,214	150	346	346	0	861	1,235
G	<i>Mimaenictus wilsoni</i>	57,519,823	150	2,171	2,153	18	183	88
G	<i>Othius punctulatus</i>	22,205,266	150	230	230	0	618	1,594
G	<i>Paederus littoralis</i>	20,453,976	150	59	59	0	395	1,988
G	<i>Philonthus decorus</i>	24,318,503	150	257	256	1	855	1,330
G	<i>Quedius fuliginosus</i>	21,835,452	150	58	58	0	367	2,017
G	<i>Silpha</i> sp.	26,143,518	150	164	164	0	546	1,732
G	<i>Stenus bimaculatus</i>	23,025,792	150	1,488	1,474	14	669	285
G	<i>Tachinus rufipes</i>	24,657,085	150	161	161	0	504	1,777

Table S5. Orthograph output summary that lists: the total number (#) of orthologs (OGs); total number of amino-acids (AA); number of 'X's and stop codons in the sequences; N50; minimum (min), median, average, and maximum (max) length of amino-acid sequences of each target species. Target species were sorted alphabetically and by data type: low-coverage genome

Family	Species	Data type	Total # genes	Total # AA	# X	# stop codons	N50	min length	median length	average length	max length
Staphylinidae	<i>Acidota crenata</i>	G	3,447	786,423	245	49	286	8	185	228	1,816
Staphylinidae	<i>Anotylus rugosus</i>	G	2,525	270,282	199	20	129	9	87	107	736
Staphylinidae	<i>Dalotia coriaria</i>	G	3,331	1,002,404	8	78	395	13	238	300	2,425
Staphylinidae	<i>Deinopsis erosa</i>	G	3,423	980,855	9	89	373	11	229	286	4,194
Staphylinidae	<i>Deleaster dichrous</i>	G	3,480	982,815	50	105	369	8	225	282	2,536
Staphylinidae	<i>Diochus sp.</i>	G	3,201	414,585	28	21	152	6	111	129	1,289
Staphylinidae	Athetini gen. sp.	G	3,467	973,049	196	50	366	14	227	280	2,020
Staphylinidae	<i>Gymnusa sp.</i>	G	3,480	826,377	334	96	297	25	200	237	1,774
Staphylinidae	<i>Lathrobium brunnipes</i>	G	3,460	883,875	144	83	326	6	207	255	1,900
Staphylinidae	<i>Lesteva longoelytrata</i>	G	2,882	471,003	579	26	207	21	132	163	994
Staphylinidae	<i>Lordithon lunulatus</i>	G	3,360	623,180	243	34	231	21	155	185	1,183
Staphylinidae	<i>Mimaenictus wilsoni</i>	G	3,417	999,458	21	52	390	10	230	292	3,011
Staphylinidae	<i>Othius punctulatus</i>	G	3,196	514,652	476	34	200	9	135	161	915
Staphylinidae	<i>Paederus littoralis</i>	G	2,790	323,467	397	28	141	9	100	115	807
Staphylinidae	<i>Philonthus decorus</i>	G	3,356	565,609	383	26	207	6	143	168	964
Staphylinidae	<i>Quedius fuliginosus</i>	G	3,100	409,352	336	66	159	7	114	132	899
Staphylinidae	<i>Stenus bimaculatus</i>	G	3,448	909,982	167	78	334	6	216	263	2,783
Staphylinidae	<i>Tachinus rufipes</i>	G	3,177	486,248	498	86	187	11	134	153	957
Silphidae	<i>Silpha sp.</i>	G	3,202	502,539	283	186	190	15	134	156	1,447
Hydraenidae	<i>Hydraena subimpressa</i>	T	3,337	915,302	0	17	336	6	235	274	1,318
Leiodidae	<i>Anisotoma castanea</i>	T	3,550	1,203,303	0	12	426	9	284	338	2,396
Leiodidae	<i>Paragyrtonotus modestus</i>	T	3,411	1,175,325	0	9	431	15	288	344	2,723
Ptiliidae	<i>Acrotichis sp.</i>	T	2,322	429,029	0	6	228	13	155	184	981
Ptiliidae	<i>Philagarica sp.</i>	T	3,327	928,968	0	8	351	6	232	279	1,789
Silphidae	<i>Nicrophorus vespilloides</i>	T	3,608	1,244,924	0	13	429	20	284	345	2,863
Silphidae	<i>Oiceoptoma thoracicum</i>	T	2,778	504,259	0	21	217	13	162	181	810
Staphylinidae	<i>Aleochara curtula</i>	T	3,283	993,766	0	11	379	12	256	302	2,086
Staphylinidae	<i>Ocyopus brunnipes</i>	T	3,131	986,999	0	7	398	8	260	315	2,297
Staphylinidae	<i>Paederus cruenticollis</i>	T	3,453	1,105,871	0	6	400	7	266	320	2,415
Staphylinidae	<i>Silphotelus sp.</i>	T	3,290	1,117,049	0	6	427	10	283	339	2,704
Agyrtidae	<i>Pteroloma forstromii</i>	PB	38	8,810	0	0	248	115	217	231	479
Hydraenidae	<i>Hydraena sp.</i>	PB	35	8,107	0	0	246	116	216	231	476
Jacobsoniidae	<i>Derolathrus sp.</i>	PB	28	6,040	0	0	254	69	217	215	356
Jacobsoniidae	<i>Derolathrus sp.</i>	PB	26	5,937	0	0	250	69	228	228	472
Leiodidae	<i>Agathidium sp.</i>	PB	35	8,425	0	0	265	40	220	240	485
Leiodidae	<i>Agyrtodes sp.</i>	PB	37	8,425	0	0	245	75	221	227	471
Nosodendridae	<i>Nosodendron sp.</i>	PB	39	9,584	0	0	265	124	221	245	609
Ptiliidae	gen. sp.	PB	8	1,429	0	0	194	96	193	178	252
Silphidae	<i>Nicrodes littoralis</i>	PB	37	9,339	0	0	265	115	238	252	495
Silphidae	<i>Nicrophorus nepalensis</i>	PB	39	9,999	0	0	274	135	233	256	609
Staphylinidae	<i>Apatetica sp.</i>	PB	34	8,854	0	0	278	124	242	260	629
Staphylinidae	<i>Centrophthalmus sp.</i>	PB	12	2,550	0	0	243	119	208	212	419
Staphylinidae	<i>Dianous sp.</i>	PB	35	8,822	0	0	256	123	222	252	637
Staphylinidae	Omaliinae gen. sp.	PB	31	7,645	0	0	258	133	221	246	600
Staphylinidae	Scydmaeninae gen. sp.	PB	25	5,146	0	0	249	80	185	205	478
Staphylinidae	Staphylinini gen. sp.	PB	17	3,382	0	1	217	113	212	198	258
Staphylinidae	<i>Megalopaederus sp.</i>	PB	30	6,853	0	0	256	110	203	228	496
Staphylinidae	<i>Osorius sp.</i>	PB	32	7,661	0	0	258	115	213	239	602
Staphylinidae	<i>Priochirus sp.</i>	PB	35	8,458	0	0	258	93	229	241	486
Staphylinidae	<i>Scaphidium sp.</i>	PB	22	4,877	0	0	244	101	202	221	416
Staphylinidae	<i>Scaphidium sp.</i>	PB	27	5,839	0	0	236	92	199	216	474
Staphylinidae	<i>Staphylinus sp.</i>	PB	36	8,834	0	0	259	91	227	245	494
Staphylinidae	<i>Tachinus sp.</i>	PB	34	8,380	0	3	258	124	220	246	609
Staphylinidae	<i>Tachinus sp.</i>	PB	34	8,686	0	0	266	124	234	255	634

Chapter 1

Table S6. The number (#) of sequences removed from individual MSAs by species and data type. RG: reference genome; G: low-coverage genome; T: transcriptome; PB: primer-based genes.

Family	Species	Data type	# Removed
Agrilinae	<i>Tribolium castaneum</i>	RG	3
Scarabaeinae	<i>Onthophagus taurus</i>	RG	16
Tenebrioninae	<i>Agrilus planipennis</i>	RG	35
Staphylinidae	<i>Dalotia coriaria</i>	G	1
Staphylinidae	<i>Acidota crenata</i>	G	2
Staphylinidae	<i>Deinopsis erosa</i>	G	4
Staphylinidae	<i>Lathrobium brunnipes</i>	G	4
Staphylinidae	Athetini gen. sp.	G	5
Staphylinidae	<i>Gymnusa sp.</i>	G	6
Staphylinidae	<i>Philonthus decorus</i>	G	7
Staphylinidae	<i>Lordithon lunulatus</i>	G	8
Staphylinidae	<i>Deleaster dichrous</i>	G	10
Staphylinidae	<i>Stenus bimaculatus</i>	G	12
Staphylinidae	<i>Lesteva longoelytrata</i>	G	15
Staphylinidae	<i>Othius punctulatus</i>	G	16
Staphylinidae	<i>Paederus littoralis</i>	G	16
Staphylinidae	<i>Diochus sp.</i>	G	19
Staphylinidae	<i>Tachinus rufipes</i>	G	19
Silphidae	<i>Silpha sp.</i>	G	22
Staphylinidae	<i>Quedius fuliginosus</i>	G	26
Staphylinidae	<i>Anotylus rugosus</i>	G	74
Staphylinidae	<i>Silphotelus sp.</i>	T	1
Staphylinidae	<i>Paederus cruenticollis</i>	T	2
Staphylinidae	<i>Ocypus brunnipes</i>	T	4
Staphylinidae	<i>Aleochara curtula</i>	T	6
Silphidae	<i>Nicrophorus vespilloides</i>	T	6
Leiodidae	<i>Paragyrtonodes modestus</i>	T	6
Hydraenidae	<i>Hydraena subimpressa</i>	T	9
Leiodidae	<i>Anisotoma castanea</i>	T	11
Silphidae	<i>Oiceoptoma thoracicum</i>	T	13
Ptiliidae	<i>Philagarica sp.</i>	T	139
Ptiliidae	<i>Acrotrichis sp.</i>	T	208
Staphylinidae	<i>Megalopaederus sp.</i>	PB	1
Staphylinidae	<i>Osorius sp.</i>	PB	1
Staphylinidae	<i>Tachinus sp.</i> (INB088)	PB	1

Table S7. Descriptive statistics of amino-acid (aa) site coverage of pairwise species comparisons of supermatrices 1 (omic-only) and 2 (omic & primer-based). Ca: completeness (C) score of the alignment (total number of unambiguous characters / (number of sequences * length of alignment)); Cr: Completeness score of individual sequences in the alignment, i.e. rows (Number of unambiguous characters in the sequence / alignment length); Cc: completeness score for individual sites, i.e. columns (Number of unambiguous characters in the column / number of sequences); Cij: completeness score for each pair of sequences in an alignment; i.e. overlap between sequences in the alignment (Number of columns in which the corresponding characters of both i-th and j-th sequence are unambiguous / length of alignment). Cij = 1, when i=j; Minimum (min), maximum (max) and average (avg) values are provided.

	Omic-only	Omic & primer-based
# species	33	57
# aa sites	494,743	
Ca	0.59	0.35
Cr_min	0.21	0.003
Cr_max	0.98	0.98
Cc_min	0	0
Cc_max	1	0.95
Cij_min	0.06	0
Cij_max	0.9	0.9
Pij_min	0.08	0.003
Pij_avg	0.26	0.17
Pij_max	0.37	0.37

Chapter 1

Table S8. Summary statistics of maximum likelihood (ML) analyses of supermatrices 1 (omic-only) and 2 (omic & primer-based). Model: best partition scheme, listing the type of model and how often it was used in the analysis. The 993 partitions were merged where possible into 828 (supermatrix 1) and 400 (supermatrix 2) partitions; LogLikelihood: log-likelihood values of 20 tree searches, each beginning with a random starting tree. Lowest absolute log-likelihood of the best ML tree in bold.

	Supermatrix 1	Supermatrix 2
Model		
Dayhoff	2	0
DayhoffF	5	2
DCMut	2	1
DCMutF	0	1
JTT	84	20
JTTDCMut	35	11
JTTDCMutF	19	11
JTTF	79	53
LG	53	19
LG4M	5	3
LG4X	437	199
LGF	88	65
PMB	0	1
VT	7	3
VTF	10	10
WAGF	2	1
logLikelihood		
Tree 1	-7,748,997.77	-8,262,233.99
Tree 2	-7,748,997.78	-8,262,354.82
Tree 3	-7,748,959.54	-8,262,207.80
Tree 4	-7,748,974.63	-8,262,185.19
Tree 5	-7,748,994.17	-8,262,189.12
Tree 6	-7,748,974.37	-8,262,189.80
Tree 7	-7,748,933.11	-8,262,313.03
Tree 8	-7,748,989.30	-8,262,213.07
Tree 9	-7,748,960.00	-8,262,352.98
Tree 10	-7,748,932.44	-8,262,323.79
Tree 11	-7,748,934.64	-8,262,364.37
Tree 12	-7,748,944.42	-8,262,161.87
Tree 13	-7,748,980.05	-8,262,237.58
Tree 14	-7,748,994.12	-8,262,195.10
Tree 15	-7,748,933.31	-8,262,145.86
Tree 16	-7,748,929.79	-8,262,161.28
Tree 17	-7,748,901.04	-8,262,166.76
Tree 18	-7,748,994.18	-8,262,158.90
Tree 19	-7,748,933.10	-8,262,167.50
Tree 20	-7,748,925.23	-8,262,230.75

Chapter 2

The past and the present through phylogenetic analysis: the rove beetle tribe Othiini now and 99 Ma

J.L. Kypke, A. Solodovnikov, A. Brunke, S. Yamamoto, D. Żyła (2018) *Systematic Entomology*, DOI: 10.1111/syen.12305.

The past and the present through phylogenetic analysis: the rove beetle tribe Othiini now and 99 Ma

JANINA L. KYPKE¹, ALEXEY SOLODOVNIKOV¹ ,
ADAM BRUNKE², SHŪHEI YAMAMOTO^{3,4}  and DAGMARA ŻYŁA¹ 

¹Natural History Museum of Denmark, Biosystematics Section, Zoological Museum, Copenhagen, Denmark, ²Canadian National Collection of Insects, Arachnids and Nematodes, Agriculture and Agri-Food Canada, Ottawa, ON, Canada, ³The Kyushu University Museum, Fukuoka, Japan and ⁴Integrative Research Center, Field Museum of Natural History, Chicago, IL, U.S.A.

Abstract. In order to classify and taxonomically describe the first two fossil Othiini (Coleoptera: Staphylinidae: Staphylininae) species from three well-preserved specimens in Cretaceous Burmese amber, a phylogenetic analysis was conducted, combining extant and extinct taxa. A dataset of 76 morphological characters scored for 33 recent species across the subfamilies Staphylininae and Paederinae was analysed using maximum parsimony and Bayesian inference methods. The many differing phylogenetic hypotheses for higher-level relationships in the large rove beetle subfamilies Staphylininae and Paederinae were summarized and their hitherto known fossil record was reviewed. Based on the analyses, the new extinct genus *Vetatrecus* **gen.n.** is described with two new species: *V. adelfiae* **sp.n.** and *V. secretum* **sp.n.** Both species share character states that easily distinguish them from all recent Othiini and demonstrate a missing morphological link between subfamilies Staphylininae and Paederinae. This is the first morphology-based evidence for the paraphyly of Staphylininae with respect to Paederinae, suggested earlier by two independent molecular-based phylogenies of recent taxa. Our newly discovered stem lineage of Othiini stresses the importance of fossils in phylogenetic analyses conducted with the aim of improving the natural classification of extant species. It also suggests that the definitions of Staphylininae and Paederinae, long-established family-group taxa, may have to be reconsidered.

This published work has been registered in ZooBank, <http://zoobank.org/urn:lsid:zoobank.org:pub:817F39C4-F36B-4FD9-96CD-5F8FB064C39E>.

Introduction

The phylogenetic systematics of beetles (Coleoptera) is undergoing an exciting transformation: paleontological evidence is increasingly integrated at the data analysis stage rather than in posthoc discussions. The wealth of evolutionary theory demonstrating the importance of fossils in understanding sister-group relationships of extant taxa (Patterson, 1981; Donoghue *et al.*, 1989; Smith, 2009; Wiens & Morrill, 2011; Pyron, 2015) has finally trickled down to the systematic entomology of this group, the largest among insects. Rove beetles (Staphylinidae) are no exception. With 63 137 extinct and extant described species

(Ahn *et al.*, 2017) and many more still undescribed, this ubiquitous beetle family displays a truly spectacular evolutionary radiation (Thayer, 2016), within which relationships are still very poorly understood. A great majority of the past systematic work in Staphylinidae, even when phylogeny-based, has been done without the consideration of fossils. However, the use of fossils in rove beetle systematics is now becoming a trend, which is showing promising results (Solodovnikov *et al.*, 2013; Jałoszyński, 2015; Parker, 2016; Yamamoto *et al.*, 2016; Yamamoto & Maruyama, 2017; Żyła *et al.*, 2017).

For example, resolving the phylogenetic relationships among and within the tribes of the ‘typical rove beetles’, the subfamilies Staphylininae and related Paederinae, based on recent taxa alone, has been quite problematic even though the systematics of Staphylininae has been studied more extensively compared with other rove beetle subfamilies (Chatzimanolis *et al.*, 2010a;

Correspondence: Janina L. Kypke, Natural History Museum of Denmark, Biosystematics Section, Zoological Museum, Universitetsparken 15, DK-2100 Copenhagen, Denmark. E-mail: j.kypke@snm.ku.dk

Chatzimanolis, 2012, 2014; Brunke *et al.*, 2016; Chani-Posse *et al.*, 2018). After Aleocharinae (16 500 described species) (Cai *et al.*, 2017a) and Pselaphinae (10 000 described species) (Yin *et al.*, 2017b), Staphylininae are the third most speciose rove beetle subfamily (7972 described species) (Thayer, 2016), currently divided into seven extant tribes: Staphylinini, Platypsopini, Arrowinini, Diochini, Xantholinini, Maorothiini and Othiini. There is little consensus among the published morphology- and molecular-based phylogenetic analyses regarding sister-group relationships of those tribes. Even the long-assumed monophyly of the subfamily Staphylininae was recently placed in doubt based on phylogenetic analyses of six genes and five genes, respectively (Brunke *et al.*, 2016; Schomann & Solodovnikov, 2017). While Brunke *et al.* (2016) focused on Staphylininae with special interest in the tribe Staphylinini, Schomann & Solodovnikov (2017) sampled Paederinae more densely with the goal of placing the genus *Hyperomma* Fauvel phylogenetically. However, despite their different sampling and study goals, both phylogenies rejected the monophyly of Staphylininae.

The hypothesis of a paraphyletic Staphylininae was in fact partly substantiated in an earlier study by Solodovnikov *et al.* (2013), who included Staphylininae fossils from the Lower Cretaceous Yixian Formation in a phylogenetic analysis together with recent taxa. One of the outcomes of that study was the discovery of a new stem lineage of Staphylininae, the tribe Thayeralini. Interestingly, members of this tribe showed character combinations transitional between recent tribes of Staphylininae and even between that subfamily and Paederinae. As very often occurring in palaeontology, a limitation of Solodovnikov *et al.* (2013) was the partially poor preservation of the rock (compression) fossils, preventing a detailed examination of some character systems.

In this light, it is very exciting to find two well-preserved fossil species from Cretaceous Burmese amber, which also combine characters of Staphylininae and Paederinae and overall resemble members of Othiini, one of the less diverse but phylogenetically understudied tribes of Staphylininae. Due to a burst of recent studies of its well preserved inclusions, Burmese amber is becoming one of the world's most important sources of Staphylinidae fossil specimens, with representatives of most subfamilies already discovered: Aleocharinae (Cai & Huang, 2015a; Yamamoto *et al.*, 2016; Yamamoto & Maruyama, 2017; Cai *et al.*, 2017a), Dasycerinae (Yamamoto, 2016a), Euaesthetinae (Clarke & Chatzimanolis, 2009), Megalopsidiinae (Yamamoto & Solodovnikov, 2016), Micropeplinae (Cai & Huang, 2014), Osoriinae (Cai & Huang, 2015b), Oxytelinae (Lü *et al.*, 2017), Oxyporinae (Yamamoto, 2017; Cai *et al.*, 2017b), Proteininae (Cai *et al.*, 2016), Pselaphinae (Parker, 2016), Scydmaeninae (Chatzimanolis *et al.*, 2010b; Cai & Huang, 2016; Jałoszyński *et al.*, 2016, 2017a, 2017b; Yin *et al.*, 2017a), Solieriinae (Thayer *et al.*, 2012), Steninae (Żyła *et al.*, 2017), and Tachyporinae (Yamamoto, 2016b). Interestingly, these two presumed Othiini are the first known representatives of the Staphylininae found in Burmese amber. Their resemblance to recent Othiini necessitates a focused systematic study of this tribe. In fact, the monophyly of the tribe Othiini has never

been thoroughly explored, and the most recent and significant contribution (Assing, 2000) towards this goal was matrix-based but noncomputational. Assing (2000) moved the New Zealand species of *Othius* Stephens into a new genus that formed its own tribe, Maorothiini. Even after that, much uncertainty remains regarding the composition and sister-group relationships of Othiini, as well as the monophyly of Staphylininae and phylogenetic relationships of its tribes. These questions, combined with the unusual character combinations of the newly found fossils, prompted a morphological phylogenetic analysis, combining extinct and extant Staphylininae and Paederinae. We anticipate that these newly found fossil species will play a significant role in future phylogenetic work on Staphylinidae.

To provide context for our phylogenetic analyses and resulting identity of the newly discovered fossils, we briefly review the systematic controversies of the 'typical rove beetles' and their fossil record.

Composition and rank of Othiini

Othiini is a markedly north-temperate group (Fig. 1) presently consisting of 142 species in four genera as follows: *Othius* Stephens, 1829 (126 species), *Atrecus* Jacquelin du Val, 1856 (13 species), *Parothius* Casey, 1906 (two species) and *Caecolinus* Jeannel, 1923 (one species). The majority of species belongs to the widespread Eurasian genus *Othius*, of which some Palearctic species were introduced to North America (Klimaszewski *et al.*, 2013; Rood *et al.*, 2015). The genus *Atrecus* has fewer species but it naturally and broadly occurs in the Holarctic. Both species of *Parothius* are endemic to the North American west coast (Smetana, 1982; Bousquet *et al.*, 2013) and a single species of *Caecolinus* has only been found in the Bihor Mountains of Romania in Europe (Jeannel, 1922).

Erichson (1839) was the first person to group *Xantholinus* Dejean, 1821, *Othius*, *Diochus* Erichson, 1839, *Platyprosopus* Mannerheim, 1830 and four other genera into Xantholinina (a subtribe that he named 'Xantholinini'), nested within the tribe Staphylinini and a close group to the subtribe Staphylinina genuina ('Staphylinini genuini'). Jacquelin du Val (1856) described and added *Atrecus* to the subtribe Xantholinina ('Xantholinini'). Around the same time, Thomson (1859), based on the Scandinavian fauna, created the subtribe Othiina ('Othiides') for *Othius* (the type genus) and *Gyrophypnus* Leach, 1819, which he considered closely affiliated to the subtribe Xantholinina ('Xantholinides') containing *Xantholinus* and *Leptacinus* Erichson, 1839. However, other systematists did not adopt this classification but rather followed the original concept by Erichson. LeConte's (1861) subtribe Xantholinina ('Xantholinini') for North America included *Xantholinus*, *Diochus*, *Othius* and a few other genera, including species of *Parothius* (then within *Othius*). The main difference between subsequent works was that Diochini and Platypsopini were considered as separate tribes by Casey (1906), while Reitter (1908) included them in the tribe Othiini, comprising *Othius*, *Diochus*, *Platyprosopus* and *Atrecus* (= *Baptolinus* Kraatz, 1857). Finally, Jeannel (1922) described *Caecolinus endogaeus* and placed it in Xantholinini, close to

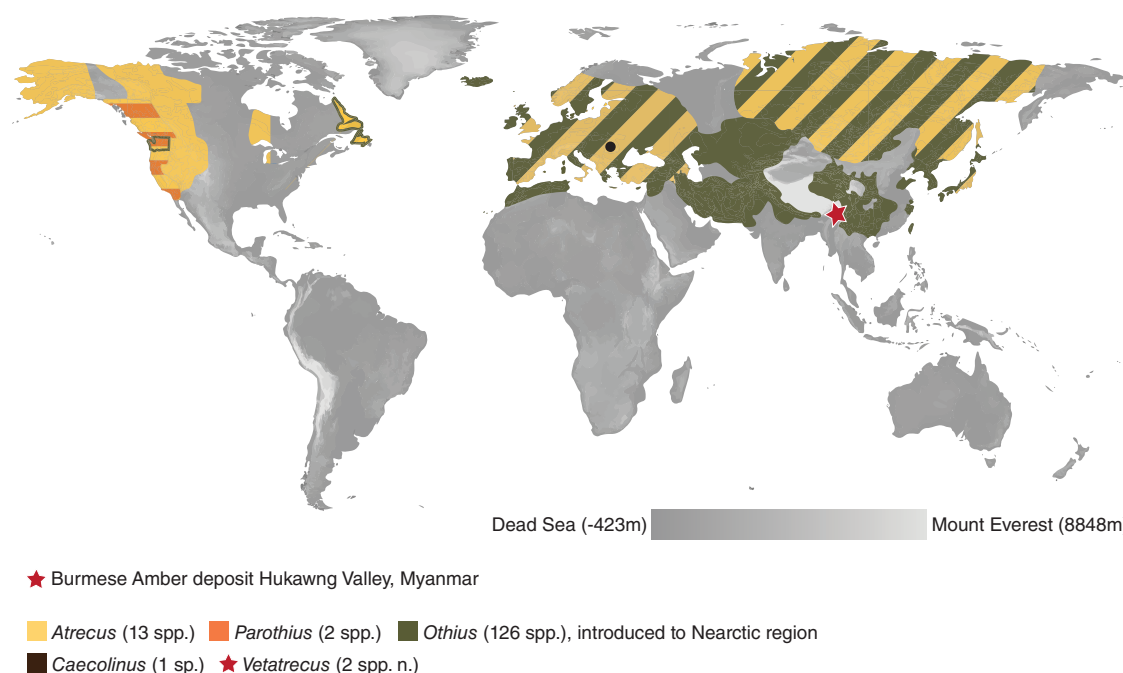


Fig. 1. Geographic distribution of the tribe Othiini (summarized from the literature). World map after Huffman & Patterson (2013) licensed under <http://creativecommons.org/licenses/by-nc-nd/3.0/> [Colour figure can be viewed at wileyonlinelibrary.com].

Atrecus. For North America, Moore & Legner (1973, 1975, 1979) used the classification of Moore (1964), who raised Xantholinini, Diochini and Platyprosopini to subfamily level. Smetana (1982), for North America, and Coiffait (1972), for the Western Palearctic, followed Casey (1906) in using Xantholinini, Diochini, Platyprosopini and Othiini as tribes, the latter with the genera *Caecolinus*, *Othius* and *Atrecus*. The current tribal rank and composition of Othiini has been stabilized in Newton & Thayer (1992) to comply with the International Code of Zoological Nomenclature.

Phylogeny and tribal system of Staphylininae and allies

Clearly, misleading similarities in habitus and conflicting evaluations of characters, especially when studies were limited to local faunas, have led to much controversy over the concept of Othiini and allied taxa. In addition to Maorothiini (Assing, 2000), described after a thorough character evaluation using a global taxon sample, two new tribes, the extant Arrowinini (Solodovnikov & Newton, 2005) and extinct Thayeralinini (Solodovnikov *et al.*, 2013), have been erected within Staphylininae based on phylogenetic analyses, the latter including fossils. Additionally, a fossil-integrated phylogeny revealed the stem group of the tribe Arrowinini (*Paleowinus* Solodovnikov & Yue). The first molecular-based phylogeny within Staphylininae (Chatzimanolis *et al.*, 2010a) was based on four genes and focused primarily on the tribe Staphylinini. Interestingly, it placed Othiini and Xantholinini within Staphylinini, but the former two tribes were very poorly sampled and were probably subject to long-branch attraction.

This unusual topology was not recovered by other molecular (McKenna *et al.*, 2014) or morphological studies (Solodovnikov & Newton, 2005; Solodovnikov *et al.*, 2013), which, despite minor differences (Fig. 2A, B), agree on the monophyly of Staphylinini. A similar study of Staphylinini by Brunke *et al.* (2016), using six genes and a broader taxon sampling, confirmed the monophyly of most staphylinine tribes but Staphylininae was rendered paraphyletic because of the [Arrowinini + (Paederinae + Platyprosopini)] clade nested within Staphylininae (Fig. 2C). In a molecular-based phylogeny targeting Paederinae (Schomann & Solodovnikov, 2017), and therefore based on a different taxon sampling, Staphylininae were again recovered as paraphyletic with respect to Paederinae (Fig. 2C). These molecular studies, as well as recently discovered stem lineages such as Thayeralinini or the genus *Apticax* Schomann & Solodovnikov, 2012, displaying extinct character combinations transitional between the subfamilies Staphylininae and Paederinae, are worth serious consideration and further exploration. It may well be that the current composition and sister-group relationships of Paederinae and Staphylininae assumed by conventional systematics and supported by the morphology-based phylogenetic analyses of recent taxa are an artefact. More extensive molecular data and discovery of stem lineages through exploration of fossils should shed light on this problem.

The fossil record of Staphylininae and Paederinae

The number of described fossil staphylinid species is steadily growing and currently totals 501 species in 24 subfamilies (Alroy *et al.*, 2017). Fossils assigned to the subfamily

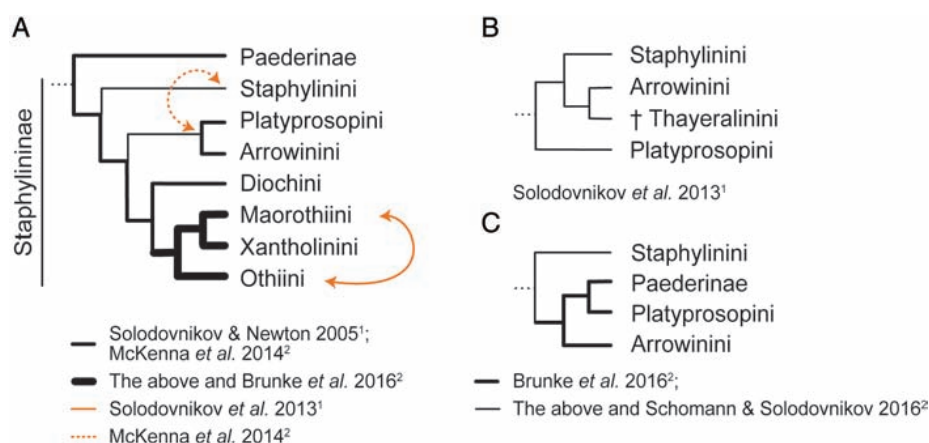


Fig. 2. Summary of previous phylogenetic studies with comparable datasets and Pseudopsinae (not shown) as an outgroup. (A) Compatible topologies from all studies except Schomann & Solodovnikov (2017), mostly congruent except two branch-swapping events indicated by respective arrows; (B) topological differences to A in Solodovnikov et al. (2013); (C) topologies from Brunke et al. (2016) and Schomann & Solodovnikov (2017) which differ from (A). 1, studies based on morphological data; 2, studies based on molecular data. Branch lengths changed. [Colour figure can be viewed at wileyonlinelibrary.com].

Staphylininae (82) constitute the majority of them (Alroy et al., 2017). Some rove beetle fossils cannot be assigned to any subfamily, either because they might represent an ancestor to recent taxa or because they are poorly preserved.

The oldest fossil recorded as Staphylininae (Staphylinini) is the impression fossil *Philonthus kneri* Giebel from the Lulworth Formation (Purbeck Group), from the Middle Berriasian of the Early Cretaceous dating back 145.5–140.2 Ma (Alroy et al., 2017). However, as for many other fossils described back then, the identity of this fossil is in need of revision. For example, the genus *Laostaphylinus* Zang with two species from impression fossils found in the Laiyang Formation in Laiyang, Shandong, China (125.5–122.5 Ma), could be suspected to be Staphylininae but does not even belong to the ‘Staphylininae–Paederinae’ lineage as revealed by Solodovnikov et al. (2013). Its systematic position within Staphylinidae remains uncertain.

Yixian and Laiyang in Northeast China are younger Lower Cretaceous formations where most of the hitherto described Staphylininae fossils originate. They can be dated to the Barremian to Aptian, c. 130–125 Ma (Swisher et al., 1999; Wang et al., 2001; Chang et al., 2009). Impression fossils from these formations have been described recently, most of them following a phylogenetic analysis (Solodovnikov et al., 2013). They included impressions of the extant tribe Staphylinini with *Durothorax creticus* Solodovnikov & Yue (Solodovnikov et al., 2013) and *Quedius cretaceus* Cai & Huang (in our view congeneric with *Cretoquedius* contrary to Cai & Huang, 2013a), and Arrowinini with five species of the genus *Paleowinus*. Furthermore, five species from these formations formed the genus *Thayeralinus* Solodovnikov & Yue placed in its own extinct tribe Thayeralinini. Four staphylinines from the Yixian formation remain incertae sedis within this subfamily either due to the poor state of preservation in *Cretoprosopus problematicus* Solodovnikov & Yue and *Paleothius gracilis* Solodovnikov & Yue, or due to the puzzling morphology in the two species of *Megolisthaerus*

Solodovnikov & Yue (Yue et al., 2010; Cai & Huang, 2013b). *Megolisthaerus* was described as incertae sedis even within Staphylinidae (Yue et al., 2010); however, it was later placed in the subfamily Staphylininae (Cai & Huang, 2013b). After a gap in geological history, there is a good sample of fossil Staphylininae from the Cenozoic, most of which have been described by Scudder (1900) as Xantholinini. Their generic and even subfamily assignment must be revised. These are five species of *Leptacinus* Erichson and *Xantholinus tenebrarius* Scudder, all from the 37.2–33.9 Ma Florissant Formation in Colorado (U.S.A.). Additionally, there is the Oligocene (28.4–23 Ma) impression fossil *Xantholinus westwoodianus* Heer from the Niveau du gypse d’Aix Formation, France, and *Neoxantholinus apolithomenus* Chatzimanolis & Engel in Miocene Dominican amber (20.4–13.7 Ma) (Heer, 1856; Chatzimanolis & Engel, 2013). Interestingly, so far only two species of Staphylininae, *Diochus electus* Chatzimanolis & Engel (of the tribe Diochini) and *Acylophorus hoffeinsorum* (Żyła & Solodovnikov, 2017) (of the tribe Staphylinini, subtribe Acylophorina), have been described from Eocene (37.2–33.9 Ma) Baltic amber, one of the largest sources of well-preserved insect fossils in the world (Chatzimanolis & Engel, 2011). However, we are aware of several other undescribed species of Staphylininae from Baltic amber, namely more species of *Diochus*, *Cyrtoquediina* and *Acylophorina* from Staphylinini, as well as species of Xantholinini (Brunke et al., 2017; Żyła & Solodovnikov, 2017; unpublished data). *Bembicidiones inaequicollis* Schauffuss, a Baltic amber fossil, which has recently been reported to be a staphylinine of unknown tribal level, originally by Chatzimanolis & Engel (2011) and followed by Alekseev (2013), should not be considered as such but of subfamily incertae sedis, as in Schauffuss (1888) and Herman (2001). The original description of *B. inaequicollis* is not informative for a subfamily (and even family) placement, and there has been no further study of this fossil. Currently, there are no described fossils for the tribes Platypsopini, Maorothiini and Othiini. Although without description, we are aware of a picture

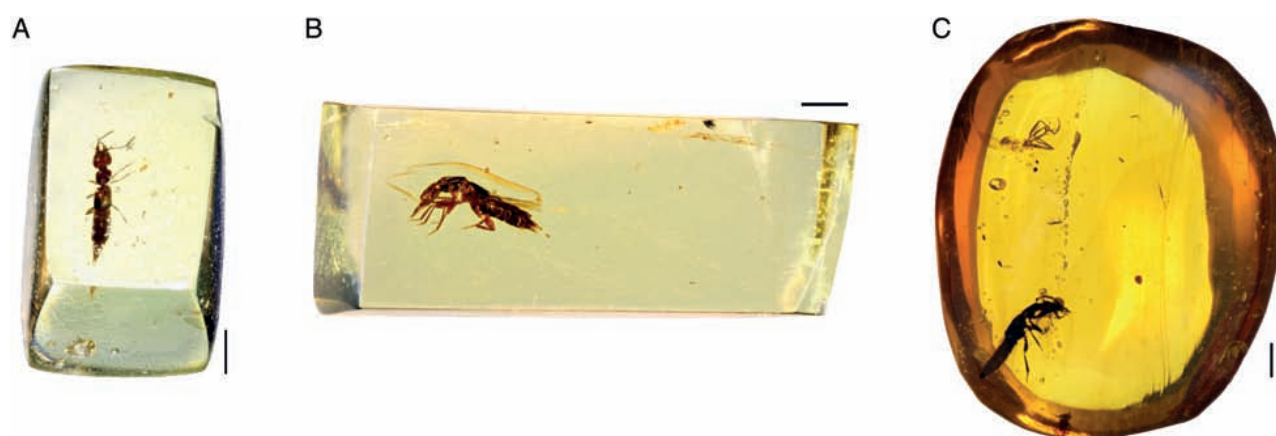


Fig. 3. Examined Burmese amber pieces, each with one specimen. (A) *Vetatrecus adelfiae*, holotype (NHMD-190040); (B) *Vetatrecus secretum*, holotype (NHMD-190041); (C) *Vetatrecus secretum*, paratype (NHMD-190042). Scale bars = 1 mm. [Colour figure can be viewed at wileyonlinelibrary.com].

showing an unnamed impression fossil in Lower Cretaceous (120 Ma) limestone from Brazil that possibly belongs to Othiini (Grimaldi & Engel, 2005).

The fossil record of Paederinae is significantly less well known than that of Staphylininae and has recently been reviewed in Schomann & Solodovnikov (2012) in connection with the discovery of *Apticax*, an Early Cretaceous genus transitional between both subfamilies and more recently in Bogri *et al.* (2018).

Materials and methods

Examination and deposition of extant taxa

For morphological study, recent specimens were boiled for 10–15 min in 10% KOH to soften and bleach them, which allowed for their dissection and better observation of exoskeletal structures. They were then rinsed in distilled water, disarticulated when necessary, and stored/examined in small Petri dishes with glycerine. Specimens of *Platyprosopus mexicanus* Sharp, *Maorothius volans* Assing, *Nudobius arizonicus* (Casey), *Atracus americanus* (Casey) and *Parothius punctatus* Smetana have been deposited in the Canadian National Collection of Insects, Arachnids and Nematodes in Ottawa (CNC). All other specimens are kept at the Natural History Museum of Denmark in Copenhagen (NHMD).

Examination and deposition of fossil taxa

Our studied material consisted of three pieces of Burmese amber, each with a single specimen (Fig. 3). Two of them, obtained by Shûhei Yamamoto, are deposited at the NHMD and each is supplied with a unique inventory number NHMD-190040 and NHMD-190041. One piece (NHMD-190042) is part of Anders Leth Damgaard's (Copenhagen) private collection which will be deposited at NHMD.

The preservation ranges from rather distorted and deformed (specimen NHMD-190042) to well-preserved (specimen NHMD-190040), allowing for detailed examination.

Microscopy and illustrations

Both recent and fossil specimens were examined using Leica M205 C and Leica M125 stereoscopes. Drawings were made freehand and digitally inked in Adobe ILLUSTRATOR CS6. All pictures were taken with a Canon EOS 6D camera (Japan) attached to a Leica M205 C stereoscope (Germany) with the help of EOS UTILITY 3.4.30.0 software (U.K.). Photomontage was accomplished using ZERENE STACKER (Zerene Systems LLC, 2012) and photos were edited in Adobe PHOTOSHOP CS6. The reconstruction of the extinct species *Vetatrecus secretum* **gen. et sp.n.** (using NHMD-190041) in Fig. 8D was made based on a freehand pencil drawing and further treated in Adobe PHOTOSHOP and ILLUSTRATOR CS6. All measurements are in mm and were taken using IMAGEJ 1.50i (Schneider *et al.*, 2012). They are abbreviated as follows: HL, head length (from apex of clypeus to neck constriction); HW, maximal head width; PL, length of pronotum (along medial line); PW, maximal pronotum width; EL, elytral length (from tip of scutellum to the level of most distal extension of elytral apical margin); EW, combined width of both elytra; FL, forebody length (calculated as the sum of HL + PL + EL); TBL, total body length (sum of FL and length of abdomen).

Taxon sampling and outgroup for phylogenetic analysis

The main focus of the present study was to resolve sister-group relationships of our studied fossil taxa as a basis for their classification and taxonomic description. Preliminary examination of the target fossils revealed two morphospecies that could be either members of, or related to, the tribe Othiini of Staphylininae. Therefore, we included representatives of all genera currently

classified as Othiini, except for the derived genus *Caecolinus*, which is known only from one species that was unavailable to us. Given the uncertainty associated with the sister-group relationships of Othiini and overall phylogeny of Staphylininae, our taxon sampling broadly covers that entire subfamily as well. Finally, in view of the recently emerged hypothesis of the paraphyly of Staphylininae with respect to Paederinae based on molecular data (Brunke *et al.*, 2016; Schomann & Solodovnikov, 2017), we included various lineages of Paederinae. Overall, the taxon sampling consisted of 33 recent species across the subfamilies Staphylininae and Paederinae in addition to our two target fossil species from Burmese amber. The species *Pseudopsis subulata* Herman, from the subfamily Pseudopsinae, was added as an outgroup as in Solodovnikov & Newton (2005) and Grebennikov & Newton (2009). In order to check the accuracy of our dataset, we also conducted separate analyses without fossil taxa. We did not include other Cretaceous compression fossils of Staphylininae or Paederinae in the analysis here, because they seem only remotely related to our target fossils and, at the same time, are significantly more poorly preserved.

Phylogenetic analyses

The data matrix included 76 characters (numbered 1–76) scored for 36 taxa and was constructed with MESQUITE 3.2 (Maddison & Maddison, 2017). Unknown character states were coded with '?' and inapplicable states with '-'. The nexus file containing the character matrix is provided in Appendix S1. Most characters and their states in our matrix are derived from previously published datasets such as Solodovnikov & Newton (2005), Brunke & Solodovnikov (2013) and Chani-Posse *et al.* (2018). However, the majority of the previously published characters have been modified to align more closely with the significantly different taxon sampling and character states observed in our target fossils. Additionally, published molecular phylogenies with comparable taxon sampling were used as an orientation for choosing morphological characters likely bearing phylogenetic signal. Therefore, we treat our matrix here as new and do not trace our characters through the listed matrices or indicate proposed changes compared with previously published data.

The maximum parsimony (MP) analyses were conducted in TNT 1.5 (Goloboff & Catalano, 2016) using the 'traditional search' option to find the most parsimonious trees (MPTs) under the following parameters: memory set to hold 1 000 000 trees; tree bisection–reconnection (TBR) branch-swapping algorithm with 1000 replications saving ten trees per replicate; zero-length branches collapsed after the search. All character states were treated as unordered and equally weighted. Autapomorphic characters were deactivated in the parsimony analysis. Bremer support was calculated using the TNT Bremer function, using suboptimal trees up to 20 steps longer. Character mapping was done in WINCLADA v1.00.08 (Nixon, 2002) using unambiguous optimization and trees were annotated in Adobe ILLUSTRATOR CS6.

Bayesian inference (BI) was conducted in MRBAYES v3.2.6 (Ronquist *et al.*, 2012) running on the CIPRES Science Gateway v3.3. (phylo.org), using the Mkv model and default settings

for priors. All analyses used four chains (one cold and three heated) and two runs of 40 million generations. Autapomorphic characters were included in the dataset, and the analyses were conducted using a gamma distribution. Convergence of the two runs was visualized in TRACER v1.6 (Rambaut *et al.*, 2013), and by examining potential scale reduction factor (PSRF) values and the average standard deviation of split frequencies in the MRBAYES output. Nodes with posterior probability (PP) > 0.95 were considered strongly supported; with PP = 0.90–0.94 moderately supported, and PP = 0.80–0.89 weakly supported. Nodes with PP < 0.80 were considered unsupported.

Characters included in the analysis

1. Antennae, form: (0) non-geniculate; (1) geniculate.
2. Antennae, insertion, base of antennomere 1: (0) partly concealed; (1) fully visible.
3. Antennae, base of antennomere 1, position: (0) on top of head; (1) concealed under a 'shelf' (Figs 6D, 7E).
4. Antennae, base of antennomere 1, position: (0) laterally; (1) dorsally.
5. Antennae, antennomere 3, dense pubescence: (0) absent; (1) present.
6. Antennae, antennomere 1–3, shape: (0) quadrangular; (1) elongate.
7. Antennae, antennomere 4–10, shape: (0) quadrangular; (1) (at least some antennomeres) elongate.
8. Head, shape: (0) oval; (1) quadrate.
9. Head, frontoclypeal (epistomal) suture: (0) present; (1) absent.
10. Head, ventral basal ridge, development: (0) absent; (1) present, strongly projecting anteriad; (2) present, extending parallel to ventral portion of postoccipital suture; (3) present, incomplete, only lateral portion.
11. Head, postgenal ridge: (0) absent; (1) present.
12. Head, infraorbital ridge: (0) absent; (1) present.
13. Head, nuchal ridge: (0) absent; (1) present.
14. Head, labrum, development: (0) entire; (1) bilobed to different extent.
15. Head, labrum, shape (from above): (0) quadrate to only slightly transverse (less than twice as wide as long); (1) transverse (more than twice as wide as long).
16. Head, mandibles, when closed: (0) projected anteriad; (1) laterally.
17. Head, mandibles, width: (0) stout; (1) thin.
18. Head, mandibular base, narrow edge with microtrichia: (0) present; (1) absent.
19. Head, mandibular prosthema, development: (0) present as a brush of long setae attached to inner margin of mandible, without well-developed supporting structure; (1) present with a more or less well-developed lanceolate supporting structure; (2) absent.
20. Head, maxillary palpomere 4 (apical), shape: (0) fusiform; (1) aciculate or conical (Figs 6D, F, 7E, F); (2) truncate, or nipple-shaped; (3) strongly modified, enlarged; (4) narrower at base, widened towards apex.

21. Head, maxillary palpomere 4, length: (0) equal to or longer than 3; (1) significantly shorter than 3.
22. Head, maxillary palpomere 4, width: (0) almost as wide as tip of penultimate palpomere; (1) half of width of tip of penultimate palpomere.
23. Head, maxillary palpomere 3, setation: (0) with at most few macrosetae; (1) heavily setose.
24. Head, labial palpomere 3, width: (0) nearly or quite as wide as penultimate; (1) half or less as wide as than penultimate.
25. Head, ligula, development: (0) bilobed; (1) entire; (2) reduced (Fig. 6F).
26. Head, development of neck constriction: (0) without distinct neck constriction; (1) with neck constriction fully developed; (2) with very weak neck constriction.
27. Neck, width: (0) as wide as head, or only slightly narrower; (1) broad, more than half of head width; (2) narrow, one-third of head width or narrower.
28. Head versus pronotum, width: (0) narrower; (1) equal to or wider.
29. Prothorax, antesternal plates, development: (0) absent (Fig. 6F); (1) sclerotized membrane; (2) present.
30. Prothorax, antesternal plates, shape: (0) triangular; (1) rectangular.
31. Pronotum, superior marginal line of pronotal hypomeron versus anterior angles of pronotum: (0) marginal line (sometimes indistinct) developed through its whole length, not deflexed under anterior angle of pronotum; (1) marginal line developed through its whole length, deflexed under anterior angle of pronotum; (2) marginal line shorter, does not extend to anterior edge of pronotum.
32. Pronotum, superior marginal line versus inferior marginal line of hypomeron: (0) inferior line not meeting superior line (inferior line sometimes very obsolete, almost indistinct and tracked by the inferior margin of pronotal hypomeron only); (1) inferior line subcontiguous or fused to superior line posterior to anterior angles of pronotum.
33. Pronotum, front angles, shape: (0) not produced beyond (anteriad) anterior margin of prosternum; (1) produced beyond (anteriad) anterior margin of prosternum (Fig. 6F).
34. Pronotum, hypomeron, development: (0) not inflexed (i.e. visible in lateral view of prothorax); (1) inflexed (i.e. not visible in lateral view of prothorax).
35. Pronotum, postcoxal process of hypomeron, development: (0) well developed and sclerotized similarly to rest of hypomeron (Fig. 7G, H); (1) absent.
36. Prosternum, pronotosternal suture, development: (0) well developed, clearly visible; (1) indistinct or completely fused at the middle.
37. Prosternum, sharp longitudinal carina: (0) present (Fig. 6F); (1) absent.
38. Prosternum, basisternum, setation: (0) more or less even, without conspicuous macrosetae, or glabrous; (1) with macrosetae.
39. Mesoventrite, medial carina: (0) present; (1) absent.
40. Mesoventrite, sternopleural suture, position and shape: (0) straight, transverse; (1) straight, oblique (i.e. medial end of suture situated anterior to its lateral end); (2) slightly curved so that medial part of suture is more longitudinal and lateral part more transverse.
41. Mesoventrite, posterior carina of prepectus, development: (0) straight or very gradually curved, parallel to anterior edge of mesoventrite, not angulate; (1) distinctly angulate at middle, forming obtuse to sharp angle.
42. Mesoventrite, mesocoxal cavities: (0) contiguous or narrowly separated by mesocoxal process; (1) widely separated (to different extent) by elevated part of metaventrite.
43. Mesothorax, scutellum, ridge(s): (0) with only one (posterior) scutellar ridge (Fig. 6D, E); (1) with anterior and posterior scutellar ridges; (2) absent.
44. Mesothorax, elytral epipleural ridge, development: (0) nearly complete, extending from apex past humerus towards base; (1) incomplete, extending from humerus towards apex; (2) absent.
45. Mesothorax, elytra, overlapping: (0) absent; (1) present.
46. Pronotum and elytra, costae: (0) present; (1) absent.
47. Tarsi, empodium, setation: (0) with one pair of setae (Figs 6D, 7E, F); (1) glabrous.
48. Tarsi, empodial setae, length: (0) about as long as or longer than claws; (1) much shorter (about half or less as long as) claws.
49. Protibiae, ventral setae, development: (0) not formed into distinct transverse or oblique comb-like rows (except a single row near tibial apex); (1) formed into several to many transverse or oblique comb-like rows.
50. Protarsi, spatulate setae: (0) absent, or present only in male; (1) present in both sexes.
51. Protarsi, tarsomeres 1–4, width: (0) not widened; (1) widened; (2) extremely widened.
52. Hind coxae, shape, in ventral view: (0) wider than longer, with laterodorsal portion widely exposed; (1) about as long as wide or elongate, with laterodorsal portion moderately to scarcely exposed; (2) about as long as wide or elongate, with wide lateroventral lobe concealing laterodorsal portion.
53. Hindwing, venation, MP3 vein: (0) present; (1) absent.
54. Hindwing, venation, veins MP4 and CuA, development: (0) completely separate; (1) largely or completely fused.
55. Metanotum, prototergal glands: (0) absent; (1) present.
56. Abdominal tergite III, transverse basal carinae: (0) only anterior; (1) anterior and posterior (Fig. 6D).
57. Abdominal tergite IV, transverse basal carinae: (0) only anterior; (1) anterior and posterior (Fig. 6D).
58. Abdominal tergite V, transverse basal carinae: (0) only anterior; (1) anterior and posterior (Fig. 7D).
59. Abdominal tergite VI, transverse basal carinae: (0) only anterior, complete; (1) anterior and posterior (Fig. 7D).
60. Abdominal tergite VII, transverse basal carinae: (0) only anterior; (1) anterior and posterior, complete (Fig. 7D).
61. Abdominal tergite VIII, transverse basal carinae: (0) only anterior, complete; (1) absent; (2) anterior and posterior complete, the latter irregular.
62. Abdominal tergites III–VI, paratergites: (0) absent; (1) present.
63. Abdominal tergites III–V, number of paratergites: (0) each segment with one paratergite on each side; (1) each segment

- with two paratergites on each side, divided longitudinally (Figs 6D; 7E).
64. Abdominal tergite VII, number of paratergites: (0) one on each side; (1) two on each side.
 65. Intersegmental membrane, pattern: (0) regular, rectangular or quadrangular, brick-wall pattern; (1) triangular, rhomboidal, with vertical lines; (2) triangular, rhomboidal, without vertical lines, brick wall pattern; (3) rhomboidal, vertical lines less distinct; (4) elongated and quadrangular in different parts.
 66. Abdominal tergal sclerites IX, shape: (0) produced but rather flat, apically obtuse to sharp, sometimes with spine-like process; (1) produced into inflated, apically sharp process; (2) produced inflated, apically obtuse or rounded process.
 67. Sternite III, keel between metacoxae: (0) absent; (1) present.
 68. Sternite IV, anteromedian gland: (0) absent; (1) present.
 69. Male, sternum VIII, apex: (0) medially straight to very slightly concave; (1) with median emargination or process.
 70. Male, lateral tergal sclerite IX: (0) dorsally fused in front of tergum X; (1) dorsally separated to different extent.
 71. Male, aedeagus, parameres, development: (0) paired; (1) fused; (2) reduced.
 72. Male, aedeagus, sensory peg setae of the paramere(s): (0) absent; (1) present.
 73. Male, aedeagus, basal part of medial lobe: (0) bulbus symmetrical; (1) bulbus asymmetrical.
 74. Male, aedeagus, internal sac of median lobe: (0) without distinct flagellum; (1) with coiled flagellum.
 75. Female, modified genital segment: (0) absent; (1) present.
 76. Female, lateral tergal sclerites IX: (0) dorsally fused in front of tergum X; (1) dorsally separated to different extent.

Results

Phylogenetic analyses

The MP analysis under equal weights resulted in one most parsimonious tree with 234 steps, a consistency index (CI) of 0.37 and a retention index (RI) of 0.73 (Fig. 4A). The BI analysis converged before 40 million generations, and at the end of the run an average standard deviation of split frequencies had stabilized well below 0.01, while nearly all PSRF values were 1.000 (maximum 1.001) (Fig. 4B).

Both MP and BI resulted in similar topologies, but with some important differences (Fig. 4). Details on the evolution of characters of *Vetatrecus* and other taxa as suggested by the most parsimonious tree are shown in Fig. 5. In both phylogenies, the subfamily Staphylininae is paraphyletic with respect to Paederinae. It forms two sister clades, with the monophyletic Paederinae nested within one of them. However, the placement of Paederinae is ambiguous as the backbone topology is not well supported in either analysis and varies between them. In the MP tree, Paederinae are sister

to the [Platyprosopini + (Arrowinini + Staphylinini)] clade. While those three tribes form the same clade in the BI tree, Paederinae are closest related to either Diochini or to the [Othiini + (Maorothiini + Xantholinini)] clade (here called the MOX clade). In the MP tree, Diochini are sister to the MOX clade. Also in that tree, the species belonging to the genus *Othius* do not form a monophyletic group, in contrast to the BI where this genus is weakly supported (PP=0.81) as monophyletic. The genera *Atrecus* and *Vetatrecus* **gen.n.** form a clade in both MP (Bremer support 4) and BI (PP=0.97) trees, as do the species of *Vetatrecus* (Bremer support 2, PP = 1). The genus *Atrecus*, by contrast, was resolved as a monophylum only using MP, while in BI it formed a well-supported polytomy with species of *Vetatrecus* (PP=0.97). The relationships amongst the remaining genera belonging to Othiini are unresolved. Given our results, two systematic solutions can be proposed: to describe the fossil species in a new genus that, most likely, is sister to *Atrecus*; or to place them in *Atrecus*. As both fossil species share character states that can easily discriminate them from *Atrecus* and that are given high importance in the suprageneric classification of Staphylininae, we preferred the first solution.

Systematic palaeontology

Order Coleoptera Linnaeus, 1758.

Family Staphylinidae Latreille, 1802.

Subfamily Staphylininae Latreille, 1802.

Tribe Othiini Thomson, 1859.

Genus *Vetatrecus* Kypke, Solodovnikov et Żyła **gen.n.**

(<http://zoobank.org/urn:lsid:zoobank.org:act:51B7E282-AA15-4B92-B12E-20EC3A52AEB5>).

Type species. Vetatrecus adelfiae Kypke, Solodovnikov et Żyła **sp.n.**

Etymology. The generic name is a chimera of the Latin adjective *vetus*, which can be translated as 'old' or 'ancient' and abbreviated as 'vet'. The stem is its presumable sister genus *Atrecus*. The gender is masculine.

Diagnosis. *Vetatrecus* can be distinguished from other genera of Staphylininae and Paederinae by the following character combination: antennae inserted dorsally on frons but their insertions slightly concealed; last segment of maxillary palpi long, aciculate; postcoxal process on hypomeron of the pronotum strongly developed; antesternal plates absent; elytra with distinct, even though incomplete, elytral epipleural ridge running from humerus towards the apex of elytron without reaching it.

Description. Habitus. Small Othiini (body length up to 2.5 mm) with relatively large head, robust long and sharp mandibles, relatively long thin legs and glabrous forebody (disc of head, pronotum and elytra without visible microsculpture),

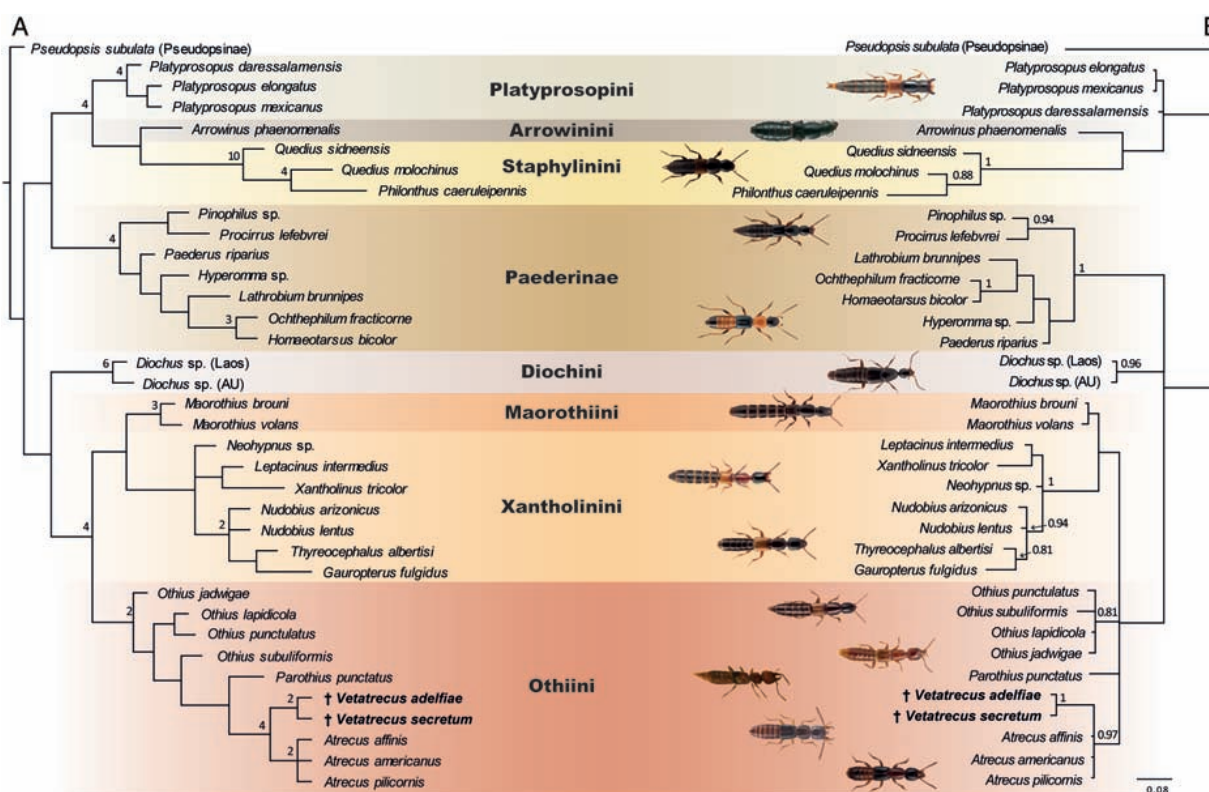


Fig. 4. Results of phylogenetic analysis. (A) The most parsimonious tree, with numbers at the corresponding nodes showing Bremer support values. (B) Fifty per cent majority-rule consensus tree from the Bayesian analysis, with posterior probabilities > 0.8 reported for the corresponding nodes. *Vetatrecus* fossil species boldfaced. Photo credits: K.V. Makarov for *Platypsopinus elongatus*, *Xantholinus tricolor* (modified from https://www.zin.ru/animalia/coleoptera/eng/staph_sf.htm); A. Kappel Hansen for *Atrecus affinis*, *Diochus* sp., *Lathrobium brunnipes*, *Maorothius brouni*, *Nudobius lentus*, *Othius punctulatus*, *Paederus riparius*, and *Quedius molochinus* (some modified from <http://www.danbiller.dk>, licensed under <https://creativecommons.org/licenses/by-nc/4.0/>). [Colour figure can be viewed at wileyonlinelibrary.com].

with few long macrosetae (Figs 6, 7). Body overall darker in coloration than paler extremities.

Head. Head capsule longer than wide with posterior angles rounded but distinct; neck about half as wide as base of head; gula well-developed, gular sutures mostly parallel to each other (Fig. 6F). Macrosetae of varying size present as in Figs 6D, 7E (also, see ‘Comment’ below); Antennae distinctly longer than head, all antennomeres setose, antennomeres 4–11 tomentose; first antennomere slightly wider and about twice as long as second antennomere; antennomeres 4–11 with distinct narrow cylindrical stem basally, starting with antennomere 6 they increase in diameter apicad; last antennomere ellipsoid with dense short pubescence at its tip. Mandibles nearly as long as head capsule, quite straight for most of their length with at least one distinct tooth internally. Labrum relatively narrow, transverse, distinctly bilobed and without teeth or serration but setose. Maxillary palps around as long as mandibles; their apical palpomere aciculate, much thinner than apex of penultimate palpomere and glabrous.

Prothorax. Pronotum longer than wide, widest around anterior angles; hypomeron visible in lateral view, with strongly developed postcoxal process; superior marginal line separating hypomeron from pronotal disc turning down beneath anterior

angles of pronotum. Antesternal plates absent. Pronotal disc with few long macrosetae anteriorly and laterally (Figs 6D, 7E).

Elytra. Moderately long, not overlapping at suture, with distinct humeri, with epipleura margined by incomplete (not reaching elytral apex) epipleural ridge.

Legs. Tibiae setose, apically with two distinct spines on posterior side and ctenidia on anterior and posterior sides. Tarsal formula 5-5-5, apical tarsomere about as long as tarsomeres 1–4 combined; protarsi with slightly broadened tarsomeres 1–4 having long, thin adhesive spatulate setae underneath; such setae distinctly shorter, sparser and thicker on more narrow middle and posterior tarsi; tarsomere 1 of all legs about half as long as tarsomere 2; tarsomeres 2–4 of approximately same length; each tarsus apically with a pair of long claws and a pair of shorter empodial setae. Middle and hind coxae contiguous.

Abdomen. About as wide as elytra, more or less parallel-sided and from segment VII distinctly tapering apicad; segments III–VII with two longitudinally divided paratergites on each side, and fine and dense setation; tergites III–VII most likely with both anterior and posterior basal carinae (in *V. secretum* can be observed only on tergites V–VII and in *V. adelfiae* only on tergites III and IV). Intersegmental membrane connecting

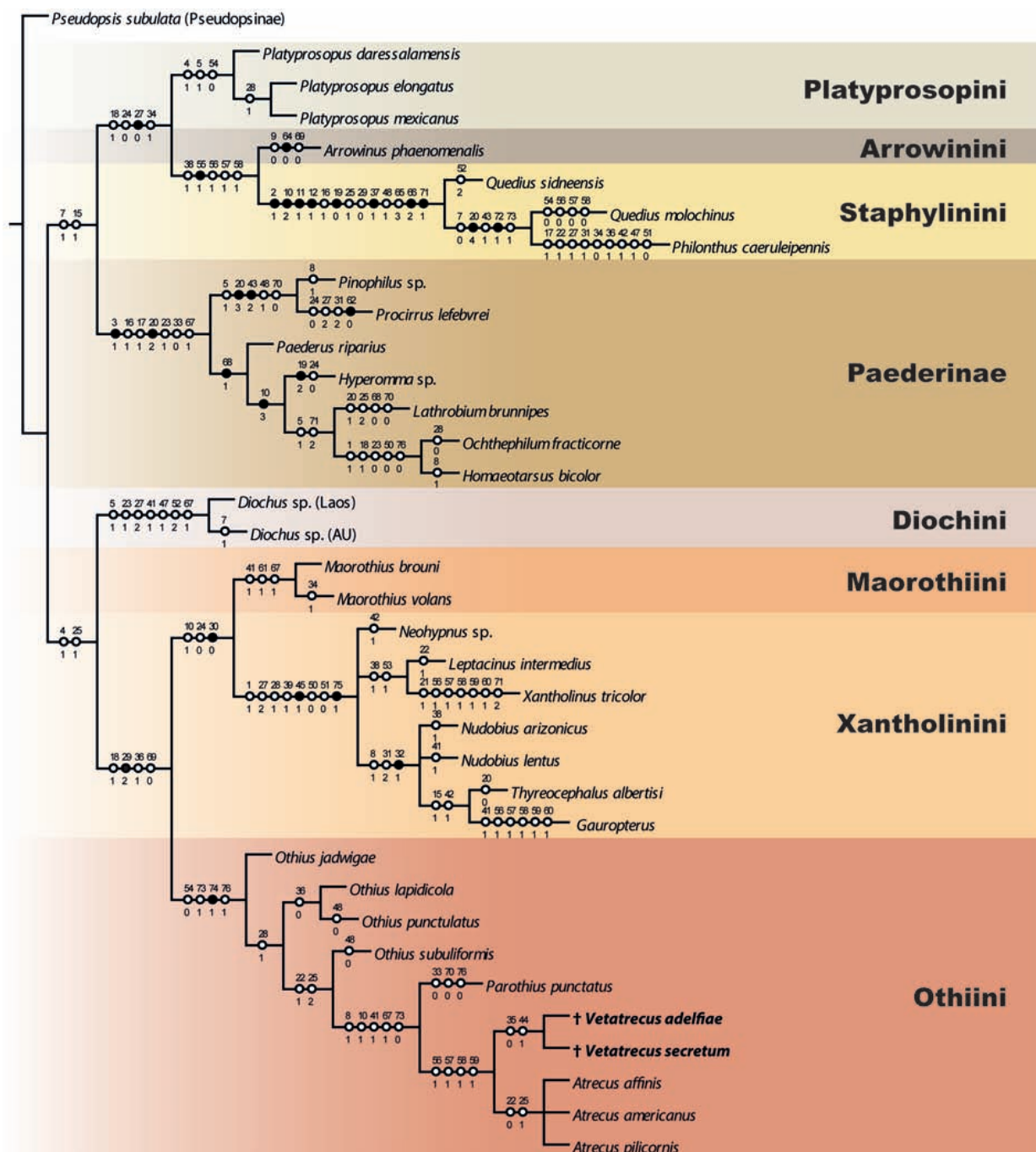


Fig. 5. Character evolution of *Vetatrecus* and other taxa as suggested by the most parsimonious tree; all character states are treated as unordered and equally weighted. Circles along the branches depict unambiguously optimized apomorphies: black circles, unique traits; white circles, homoplasious traits; numbers above the circles indicate characters, and numbers below circles indicate their states. [Colour figure can be viewed at wileyonlinelibrary.com].

abdominal segments most likely attached subapically, as observed in *A. secretum* (see paratype in ventral view; Fig. 7C) and as judged from segments too closely telescopically interlocked in other specimens. Apical margin of sterna VII and VIII straight in both sexes; segments VIII and especially VII with notable, very long and strong macrosetae.

Locality and horizon. Both the holotype and paratypes were found in Kachin, Hukawng Valley, Myanmar; Burmese amber; Upper Cretaceous, lowermost Cenomanian (Shi *et al.*, 2012).

Comment. Distinctive macrosetae on the head and pronotum of both species have been labelled in order to describe them.

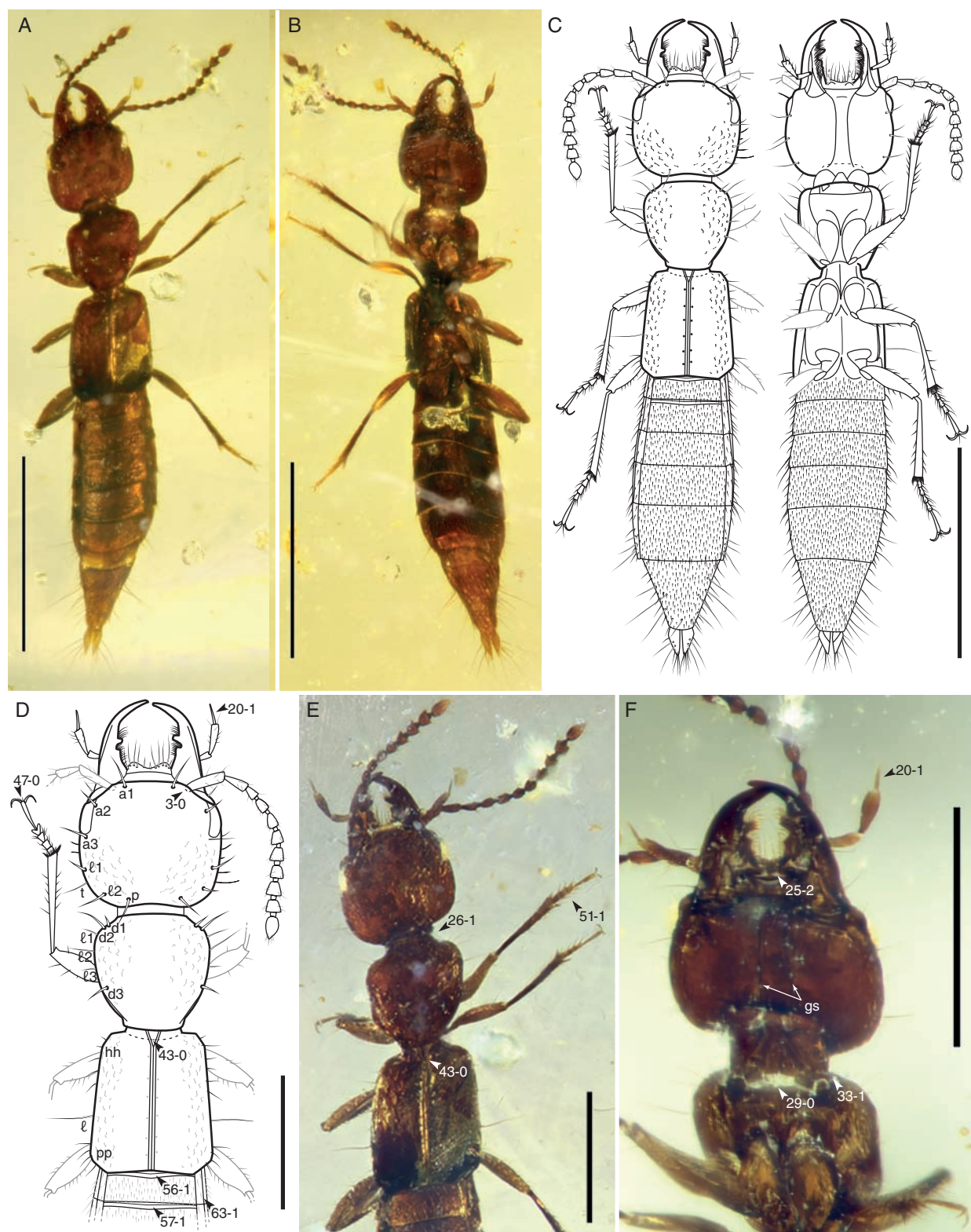


Fig. 6. Burmese amber fossil *Vetatreacus adelfiae* gen. et sp.n., holotype NHMD-190040. (A, B) Habitus photographs in dorsal (A) and ventral (B) views; (C) habitus reconstruction in dorsal (left) and ventral (right) view; (D, E) reconstruction (D) and photograph (E) of the forebody, dorsal view; (F) photograph of head and prothorax in ventral view. Abbreviations for setae: a, anterior; d, dorsal; h, humeral; ℓ - lateral; p, posterior; t, temporal; gs, gular sutures. Arrows show respective characters-states from the data matrix in S1. Scale bars: 1 mm (A–C), 0.5 mm (D–F). [Colour figure can be viewed at wileyonlinelibrary.com].

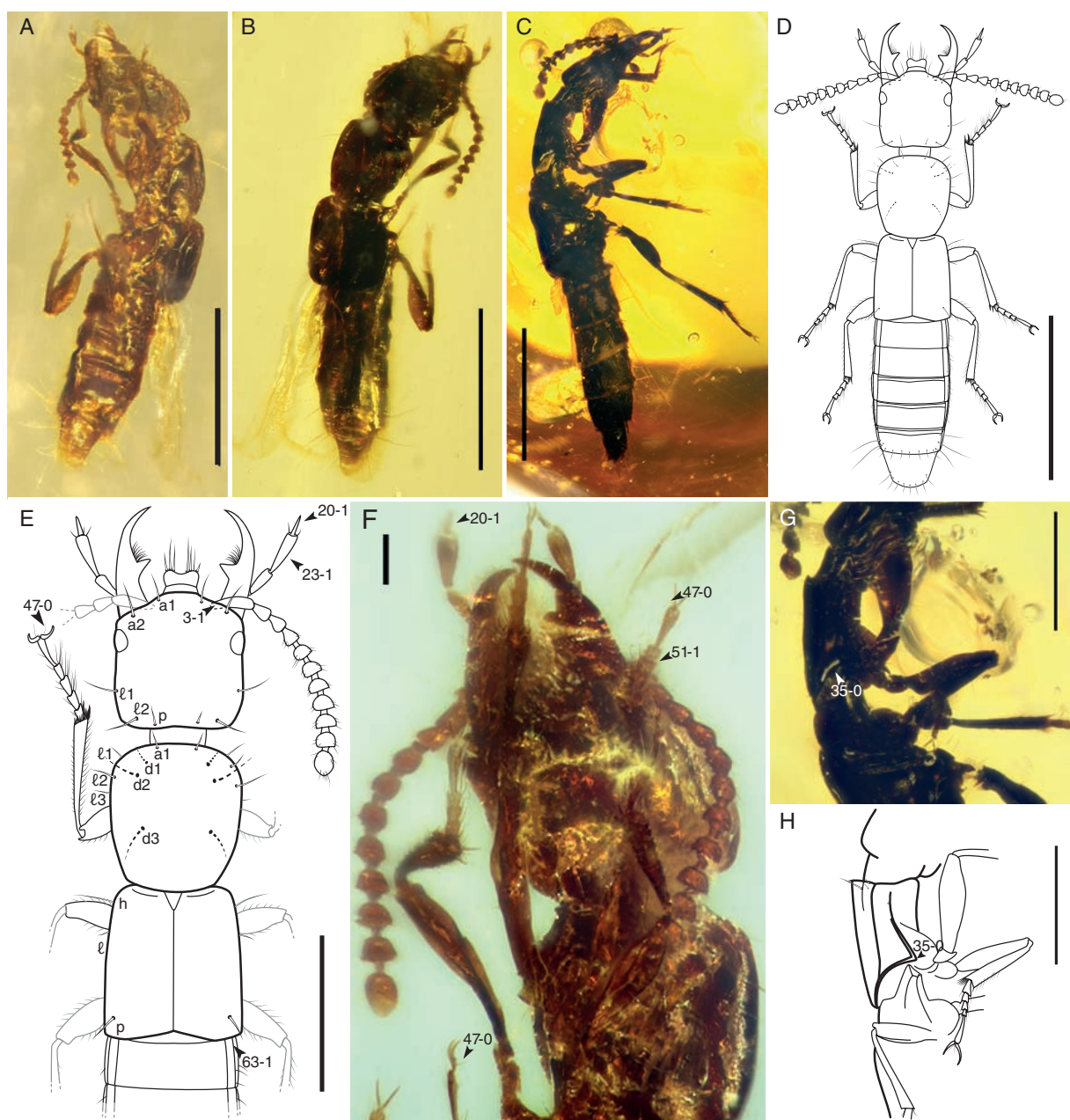


Fig. 7. Burmese amber fossil *Vetatrecus secretum* **gen. et sp.n.** (A, B, F) holotype NHMD-190041; (C, G, H), paratype NHMD-190042; (A–C) habitus photographs in ventral (A), dorsal (B) and lateral (C) views; (D) habitus reconstruction in dorsal view; (E) reconstruction of the forebody, dorsal view; (F) photograph of head and prothorax in ventral view; (G, H) photograph (G) and line drawing (H) of prothorax and adjacent body parts showing postcoxal process, lateral view. Arrows show respective character states from the data matrix in S1. Scale bars: 1 mm (A–E), 0.5 mm (G, H), 0.1 mm (F). [Colour figure can be viewed at wileyonlinelibrary.com].

Although identical labels assume a certain degree of homology, they do not assume serial homology. A more thorough study of the fossils and its closest relatives is needed to confirm these hypotheses.

***Vetatrecus adelfiae* Kypke, Solodovnikov et Żyła sp.n.**
(<http://zoobank.org/urn:lsid:zoobank.org:act:71B1E49A-4F64-4929-A3B6-9536123F3AE6>) (Fig. 6).

Type material. *Holotype*, ♀, NHMD-190040, a well-preserved specimen inside a small (c. 4 × 6 mm), very transparent prism-like piece of amber with only few impurities (Fig. 3A) (NHMD).

State of preservation. The specimen is entire and extremely well preserved, so that both dorsal and ventral sides with minor details like punctation, setation and even differences

in the degree of sclerotization of various body parts can be observed. Overall, the body shape has not been much affected by fossilization, although the legs have slightly, but noticeably, been compressed. Its labium is not visible and might have been detached, as this area is clearly visible (Fig. 6F).

Etymology. The name is dedicated to J.L. Kypke's siblings Julian Selinger, Laura Tufano, Sarah and Giorgia Ciccarese. It stems from the Greek word *αδελφία* ('adelfia'), which means siblings.

Diagnosis. *Vetatrecus adelfiae* can be distinguished from its congener *V. secretum* by the relatively larger head with more rounded hind angles and pronotum more strongly narrowing posteriad. Also, it has a slightly rugose (not smooth) surface of head and pronotum, much more densely setose abdomen, and distinct second (small) tooth basally from large tooth on inner side of mandible.

Description. Habitus. Gracile species (TBL = 2.50 mm) with large head and relatively short, roundish pronotum; head and pronotum with slightly rugose punctation, elytra and abdomen with very fine small punctation and, especially abdomen, very setose.

Head. Slightly wider than long (HW = 0.43 mm; HL = 0.38 mm), wider than pronotum (HW = 0.43 mm; PW = 0.37 mm) with parallel-sided temples. Eyes about 0.38× as long as temples, only very slightly protruding over head-contour. Visible macrosetae arranged as follows (setae listed for one side only) (Fig. 6D; also see 'Comment' after the genus description): three anterior (a1–3) of which one located laterally posterior to eye (a3); two lateral (ℓ1 and ℓ2); one temporal (t); and one posterior (p) located near neck constriction. Mentum with one pair of macrosetae. Antennomere 3 conical and slightly shorter than preceding antennomere; antennomeres 4 and 5 of same size, ellipsoid and shorter than 3; antennomere 11 about as wide as preceding antennomere but slightly longer. Mandibles mostly straight but apical third curved so that tips are crossed over in resting position; with two teeth (one larger, one smaller) on inner margin.

Pronotum. Slightly longer than wide (PL = 0.44 mm; PW = 0.37 mm), with lateral sides strongly tapering towards base, with strongly rounded and thus indistinct posterior angles. Glabrous with few macrosetae at its outer margin arranged as follows (setae listed for one side only) (Fig. 6D): three dorsal (d1–3) and three lateral (ℓ1–3).

Elytra. Slightly wider than pronotum at its widest point, 0.9× as long as pronotum (EL = 0.46 mm), densely punctured with very small punctures, with one large lateral macroseta (ℓ) and groups of smaller but distinct humeral (hh) and posterior (pp) setae, the latter close to posterior angles of elytra (Fig. 6D). Epipleural ridge developed at edge of elytral disc at base of elytra as well as close to apex of elytra but faded out in between.

Legs. Long and thin; tarsomere 5 about as long as all preceding tarsomeres combined; protarsomere 4 bilobed.

Abdomen. About as wide as elytra and parallel-sided for most of its length, tapering towards apex from segment VII; all segments densely punctured with fine setation; one pair of paratergites on segments III–VII; segments III–VI of more or less equal size, segment VII around twice as long as preceding segment, segment VIII even longer; tergum III and IV with both anterior and posterior basal carinae; tergum VIII apically with rounded convexity; apical gonocoxites widest at around the middle of their length, apically pointed.

***Vetatrecus secretum* Kypke, Solodovnikov et Żyła sp.n.**

(<http://zoobank.org/urn:lsid:zoobank.org:act:ACD84E3B-415F-4CF4-9D46-A5A45F793106>) (Fig. 7).

Type material. Holotype, presumably ♀, NHMD-190041, inside a cuboid piece of light yellow amber (c. 9×3 mm) (Fig. 3B) (NHMD).

Paratype, ♂, NHMD-190042, belonging to Anders Damgaard but deposited at the NHMD collection. The specimen can only be observed from one angle (lateroventrally) due to the round shape of the amber piece and convex surface on one side (Fig. 3C). The amber is rather dark, ranging from dark yellow to orange.

State of preservation. The holotype (NHMD-190041) is an overall well-preserved specimen, but most of the body parts are notably compressed and hence appear flattened. Hindwings are unfolded over the abdomen. Sex of the specimen is unclear, as the abdominal segments IX and X are retracted in segment XIII so that their shape is concealed. Based on structures slightly visible through segment XIII and resembling tips of gonocoxites, the specimen is presumed to be a female. The paratype (NHMD-190042) is less well preserved. However, in structures available for observation, both specimens are identical and thus assumed to be conspecific. Based on the entire apical part of sternite IX protruding from under sternite VIII, the paratype can be identified as male.

Etymology. The name stems from the Latin word *secretum*, meaning secret, seclusion or mystery, and refers to the rather difficult-to-observe postcoxal process that was traditionally considered an important character to separate Staphylininae from Paederinae.

Diagnosis. *Vetatrecus secretum* can be distinguished from its congener *V. adelfiae* in habitus features such as relatively smaller head with more pronounced hind angles and pronotum only moderately narrowing posteriad. Also, it has smooth (not slightly rugose) surface of head and pronotum, less densely setose abdomen and only one (large) tooth on inner side of mandible.

Description. Habitus. Robust species (TBL = 2.42 mm – all measurements taken from holotype, NHMD-190041) with relatively long antennae and legs, smooth glabrous forebody and sparsely setose abdomen.

Head. Approximately equal in length and width (HL = 0.44 mm; HW = 0.43 mm), almost as wide as pronotum (PW = 0.45 mm); macrosetae arranged as follows (setae listed for one side only) (Fig. 7E): two anterior (a1, a2); two lateral (ℓ 1, ℓ 2); and one smaller posterior (p) seta medially near neck constriction. Antennomeres 2–5 conical, increasing in diameter towards apex and of about same size; last antennomere as wide and twice as long as preceding antennomere. Mandibles curved in apical third but tips not strongly crossing over in resting position, apically sharply pointed, symmetrical, each with a single median tooth.

Pronotum. Longer than wide (PL = 0.53 mm; PW = 0.45 mm), widest around its poorly distinct anterior angles, narrowest at more distinct posterior angles, with sides converging posteriad, especially along posterior third of pronotal length; macrosetae (listed for one side of pronotum) arranged as in Fig. 7E: one anterior seta (a1) medially at anterior margin of pronotum; three dorsal setae (d1–3, ambiguous; these may not be setae but in fact small fractures on pronotal surface); and four lateral setae (ℓ 1–4).

Elytra. Moderately long (EL = 0.38 mm) and about as wide as pronotum (EW = 0.44 mm), with very large humeral seta (h), one smaller lateral seta behind humeral seta (ℓ) and one larger posterior seta (p) near elytral posterior margin (Fig. 7E); sub-basal ridge long, extending towards humerus, not adjacent to elytral articulation. Hindwings fully developed.

Legs. Femora distinctly wider than tibiae. Protibia with denser but thinner and longer setae than meso- and metatibiae.

Abdomen. Only slightly more narrow than elytra for most of its length and gradually tapering apicad from segment VII; generally not very setose with only few striking macrosetae on segments VII and VIII; with one pair of paratergites on segments III–VII; all visible segments (III–VIII) of about equal length, only segment VIII slightly longer; tergum of segments V–VII with both anterior and posterior basal carinae; apex of sternum VIII in both sexes medially straight; sternite IX in male apically rounded.

Discussion

There are a number of reasons for highly conflicting systematic concepts of Othiini, Xantholinini and other tribes of Staphylininae. Until the end of the 1990s, most conclusions were based on a limited set of external and aedeagal characters. This is problematic as one cannot test which characters are homoplastic and which bear the phylogenetic signal. Although molecular phylogenetic studies of Staphylinidae are progressing, they still remain rather limited in terms of breadth of taxon and depth of genetic marker sampling. Finally, as we are starting to appreciate in rove beetle systematics, the crown diversity of Staphylinidae is only a tiny fraction of the entire phylogenetic diversity of the family that ever existed. The deeper the divergence between extant groups, the more important stem groups become in phylogenetic reconstruction. It is therefore not surprising that there is little congruence among systematic hypotheses from conventional classifications of Staphylininae and

Paederinae, and new ones proposed by recent phylogenetic analyses, all largely derived from studies of recent taxa. Ideally, a holistic approach where alternative types of data are used for extant and extinct taxa together should be applied whenever possible. In the present case, a purely morphological dataset was the only practical approach to place the two new amber fossils phylogenetically, as molecular data are not yet available for several key Othiini taxa in our matrix.

In both topologies recovered by MP and BI, respectively, almost all tribes of Staphylininae and the subfamily Paederinae form relatively well-supported monophyletic groups, with the exception of Othiini, which was not monophyletic in the BI analysis. This might be due to a lack of synapomorphic characters defining this group, as currently the only unambiguous character is the presence of a coiled flagellum in the internal sac of the median lobe of the aedeagus (74-1). A similar problem exists for the genus *Othius*, which was not monophyletic in the MP analysis but may be defined by the asymmetrical bulbus of the aedeagus (73-1). The backbone nodes of both topologies are not well supported and therefore the intertribal relationships recovered here should be judged with caution until additional informative morphological characters can be added to the matrix.

We interpret the CI of only 0.37 as a sign of characters being homoplastic, rather than a result of a taxon sampling that is too small, knowing that there are characters that have been pointed out to be homoplastic in the past (Brunke *et al.*, 2016). Characters like the visibility of the hypomeron in lateral view, the development of the infraorbital ridge or the setation of maxillary palpomeres are difficult to define and most likely evolved several independently in the past. However, depending on the taxonomic level under consideration, they can become synapomorphic in combination with other characters and hence can contribute to resolving phylogenetic relationships. Arguably, this is what happened, as a striking result of this paper is that the topologies of both our analyses are closer to the molecular-based phylogenies, with comparable taxon and DNA marker sampling (Brunke *et al.*, 2016; Schomann & Solodovnikov, 2017), than to the morphological phylogenies. In congruence with the molecular analyses described earlier, the monophyly of Staphylininae is here rejected, as it is paraphyletic with respect to Paederinae. While both Brunke *et al.* (2016) and Schomann & Solodovnikov (2017) had strong support for a sister-group relationship between Paederinae and Platypsopini, the placement of Paederinae in our topologies remains incongruent and not well supported. Still, the overall congruence of the present study with the molecular-based phylogenies suggests that the addition of fossil *Vetatrecus* increased the accuracy of the phylogenetic inference, and that a subfamily-level reclassification of the Staphylininae-Paederinae lineage may be needed in the near future.

The results of our phylogenetic analyses have revealed both fossil taxa to be sister species with high support, forming the sister clade to the genus *Atrecus*. There are two character states that, according to the analysis, would present the most convincing evidence for their placement in the MOX clade: the ventral basal ridge strongly projecting anteriad (10-1) and an evenly arcuate apex of male sternite VIII (69-0). Unfortunately,

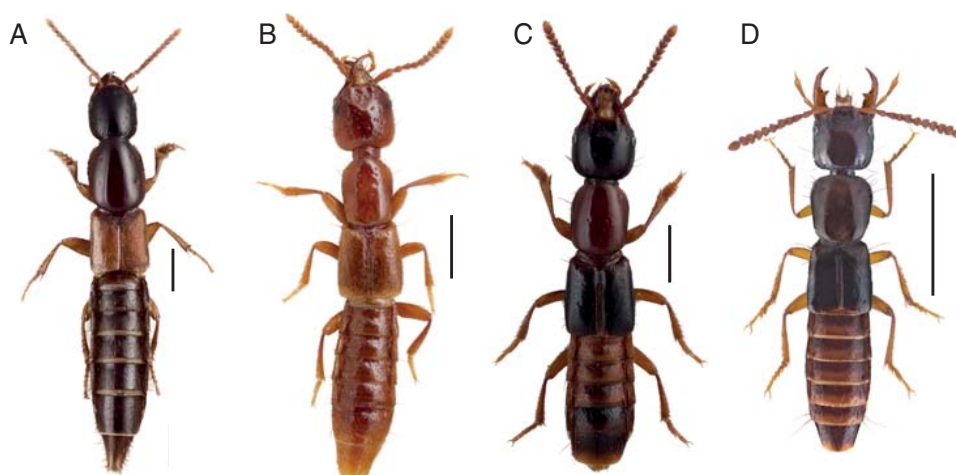


Fig. 8. Diversity of extant and extinct Othiini genera. (A) *Othius punctulatus*; (B) *Parothius punctatus*; (C) *Atrecus affinis*; (D) digital reconstruction of *Vetatrecus secretum*. Scale bars: 1 mm. Photo credits: A. Kappel Hansen for *O. punctulatus* and *A. affinis* (modified from www.danbiller.dk, licensed under <https://creativecommons.org/licenses/by-nc/4.0/>). [Colour figure can be viewed at wileyonlinelibrary.com].

the former cannot be observed in *Vetatrecus* due to poor preservation of respective body structures, although the latter is clearly observed. These are character traits to pay close attention to when any additional fossil specimens of *Vetatrecus* or similar are discovered. The first state is unique to the MOX clade and the second is additionally found in Arrowinini in our dataset (Fig. 5), although it is also known to occur in a few highly derived genera of Staphylinini (e.g. Brunke & Solodovnikov, 2013). Within the MOX clade, *Vetatrecus* was resolved as a member of Othiini and diagnostically they share plesiomorphic characters such as nonoverlapping elytra and a wide neck (at least half the width of the head) (Fig. 8). We justify placing the analysed fossils in a new genus because they differ significantly from all recent Othiini in a number of characters. Both extinct species are notably smaller than recent species of this tribe and they have several characters atypical for crown Othiini, including an incompletely developed epipleural ridge on the elytra that fades out in the middle, a lack of antesternal plates, and a well developed and sclerotized postcoxal process. The epipleural ridge is entirely absent from all Staphylininae, except for the extinct Thayeralinini, but is thought to be found in many Paederinae. However, the homology of the ‘epipleural ridge’ observed in paederines needs reassessment because it could, in fact, be a different ridge on the eplipleuron. As defined by Naomi (1989), the ‘epipleural ridge’ is the line marking where the elytral disc and the epipleuron, the lateral deflexed part of the elytron, meet. It should be on the same level as the elytral disc or very close to it. Lateral lines on the epipleuron have been coined ‘epipleural keel’ (Leschen & Newton, 2003) and ‘epipleural fold’ (Clarke & Grebennikov, 2009) but have been used synonymously with the epipleural ridge (e.g. Solodovnikov & Newton, 2005; Solodovnikov *et al.*, 2013). All extant species belonging to the staphylinine tribes of the MOX clade and Arrowinini have antesternal plates; however, they are missing in *Vetatrecus*. The only other tribe without antesternal plates in Staphylininae is Staphylinini, while Platyprosopini and Diochini

presumably have an intermediate state, a sclerotized membrane. While the postcoxal process is usually absent or only very weakly developed in Staphylininae, this structure is typical of paederines and it is surprising to observe this character in *Vetatrecus*. The latter two character states are not unique to the new genus but when combined in *Vetatrecus* they demonstrate a missing link between Staphylinini (postcoxal process and antesternal plates absent) and Paederinae (postcoxal process present and antesternal plates absent) and are thus of great value from an evolutionary perspective. In agreement with recent molecular phylogenetic work, the morphology of *Vetatrecus* combined with the phylogenetic placement of Paederinae here casts further doubt on both the monophyly of Staphylininae and the placement of the entire MOX clade to Staphylininae. Interestingly, these findings give new incentives to revisit the discussion started about 160 years ago when Thomson (1859) first hypothesized that all species belonging to the MOX clade should be placed in their own subfamily.

Author contributions

JK, DŽ and AS designed the study. JK and DŽ generated data with additions from AB and AS. DŽ performed the analyses while JK drafted the paper and drew the illustrations. JK, AS, AB and DŽ finalized the draft. SY provided important material and commented on the draft. The authors declare no conflicting financial interest.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1. Nexus file containing the morphological data matrix of 76 characters (numbered 1–76) scored for 36 taxa.

Acknowledgements

We are very thankful to Anders Leth Damgaard (Copenhagen, Denmark) for providing one of the fossils, Yui Takahashi (University of Tsukuba, Tsukuba, Japan) for preparing the amber specimens used in this study, Aslak Kappel Hansen for some illustrations, and Arn Rytter Jensen for the digital artistic reconstruction of *Vetatrecus secretum*. Members of the Solodovnikov laboratory at the Natural History Museum of Denmark (<http://www.solodovnikovlab.com>) and Volker Assing (Hannover, Germany) are thanked for fruitful discussions, even though we might agree to disagree. We are grateful to the Willi Hennig Society (cladistics.org), which made TNT freely available and the CIPRES Scientific Gateway provided access to computational resources. This work was partially supported by the Villum Foundation (block postdoctoral scholarship for DŽ) and has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 642241 (PhD fellowship of JLK).

References

- Ahn, K.-J., Cho, Y.-B., Kim, Y.-H., Yoo, I.-S. & Newton, A.F. (2017) Checklist of the Staphylinidae (Coleoptera) in Korea. *Journal of Asia-Pacific Biodiversity*, **10**, 279–336. <https://doi.org/10.1016/j.japb.2017.06.006>.
- Alekseev, V.I. (2013) The beetles (Insecta: Coleoptera) of Baltic amber: the checklist of described species and preliminary analysis of biodiversity. *Zoology and Ecology*, **23**, 5–12. <https://doi.org/10.1080/21658005.2013.769717>.
- Alroy, J., Clapham, M., Smith, D., Bottjer, D., Carrano, M. & Uhen, M. (2017) *Taxonomic Occurrences of Staphylinidae Recorded in the Paleobiology Database and ETE*. Fossilworks. [WWW document]. URL <http://fossilworks.org> [accessed on 9 March 2017].
- Assing, V. (2000) A taxonomic and phylogenetic revision of Maorothiini trib. n. from the New Zealand subregion: (Coleoptera: Staphylinidae, Staphylininae): with 2 tables, 22 figure plates, and 3 maps. *Beiträge Zur Entomologie*, **50**, 3–64.
- Bogri, A., Solodovnikov, A. & Żyła, D. (2018) Baltic amber impact on historical biogeography and palaeoclimate research: oriental rove beetle *Dysanabatum* found in the Eocene of Europe (Coleoptera: Staphylinidae: Paederinae). *Papers in Palaeontology*. <https://doi.org/10.1002/spp2.1113>.
- Bousquet, Y., Bouchard, P., Davies, A.E. & Sikes, D.S. (2013) Checklist of beetles (Coleoptera) of Canada and Alaska. Second edition. *ZooKeys*, **360**, 1–44. <https://doi.org/10.3897/zookeys.360.4742>.
- Brunke, A.J. & Solodovnikov, A. (2013) *Alesiella* gen.n. and a newly discovered relict lineage of Staphylinini (Coleoptera: Staphylinidae). *Systematic Entomology*, **38**, 689–707. <https://doi.org/10.1111/syen.12021>.
- Brunke, A.J., Chatzimanolis, S., Schillhammer, H. & Solodovnikov, A. (2016) Early evolution of the hyperdiverse rove beetle tribe Staphylinini (Coleoptera: Staphylinidae: Staphylininae) and a revision of its higher classification. *Cladistics*, **32**, 427–451. <https://doi.org/10.1111/cla.12139>.
- Brunke, A.J., Chatzimanolis, S., Metscher, B.D., Wolf-Schwenninger, K. & Solodovnikov, A. (2017) Dispersal of thermophilic beetles across the intercontinental Arctic forest belt during the early Eocene. *Scientific Reports*, **7**, 1–8. <https://doi.org/10.1038/s41598-017-13207-4>.
- Cai, C. & Huang, D. (2013a) A new species of small-eyed *Quedius* (Coleoptera: Staphylinidae: Staphylininae) from the Early Cretaceous of China. *Cretaceous Research*, **44**, 54–57. <https://doi.org/10.1016/j.cretres.2013.03.004>.
- Cai, C. & Huang, D. (2013b) *Megolisthaerus*, interpreted as staphylinine rove beetle (Coleoptera: Staphylinidae) based on new Early Cretaceous material from China. *Cretaceous Research*, **40**, 207–211. <https://doi.org/10.1016/j.cretres.2012.07.003>.
- Cai, C.-Y. & Huang, D.-Y. (2014) The oldest micropepline beetle from Cretaceous Burmese amber and its phylogenetic implications (Coleoptera: Staphylinidae). *Naturwissenschaften*, **101**, 813–817. <https://doi.org/10.1007/s00114-014-1221-z>.
- Cai, C. & Huang, D. (2015a) The oldest aleocharine rove beetle (Coleoptera, Staphylinidae) in Cretaceous Burmese amber and its implications for the early evolution of the basal group of hyper-diverse Aleocharinae. *Gondwana Research*, **28**, 1579–1584. <https://doi.org/10.1016/j.gr.2014.09.016>.
- Cai, C. & Huang, D. (2015b) The oldest osoriine rove beetle from Cretaceous Burmese amber (Coleoptera: Staphylinidae). *Cretaceous Research*, **52**(PB), 495–500. <https://doi.org/10.1016/j.cretres.2014.03.020>.
- Cai, C. & Huang, D. (2016) *Cretoleptochromus archaicus* gen. et sp. nov., a new genus of ant-like stone beetles in Upper Cretaceous Burmese amber (Coleoptera, Staphylinidae, Scydmaeninae). *Cretaceous Research*, **63**, 7–13. <https://doi.org/10.1016/j.cretres.2016.02.016>.
- Cai, C., Newton, A.F., Thayer, M.K., Leschen, R.A.B. & Huang, D. (2016) Specialized proteinine rove beetles shed light on insect–fungal associations in the Cretaceous. *Proceedings of the Royal Society B: Biological Sciences*, **283**, 20161439. <https://doi.org/10.1098/rspb.2016.1439>.
- Cai, C., Huang, D., Newton, A.F., Eldredge, K.T. & Engel, M.S. (2017a) Early evolution of specialized termitophily in Cretaceous rove beetles. *Current Biology*, **27**, 1229–1235. <https://doi.org/10.1016/j.cub.2017.03.009>.
- Cai, C., Leschen, R.A.B., Hibbett, D.S., Xia, F. & Huang, D. (2017b) Mycophagous rove beetles highlight diverse mushrooms in the Cretaceous. *Nature Communications*, **8**, 14894. <https://doi.org/10.1038/ncomms14894>.
- Casey, T.L. (1906) Observations on the staphylinid groups Aleocharinae and Xantholinini, chiefly of America. *Transactions of The Academy of Science of St. Louis*, **16**, 356–433.
- Chang, S., Zhang, H., Renne, P.R. & Fang, Y. (2009) High-precision $^{40}\text{Ar}/^{39}\text{Ar}$ age for the Jehol Biota. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **280**, 94–104. <https://doi.org/10.1016/j.palaeo.2009.06.021>.
- Chani-Posse, M.R., Brunke, A.J., Chatzimanolis, S., Schillhammer, H. & Solodovnikov, A. (2018) Phylogeny of the hyper-diverse rove beetle subtribe Philonthina with implications for classification of the tribe Staphylinini (Coleoptera: Staphylinidae). *Cladistics*, **34**, 1–40. <https://doi.org/10.1111/cla.12188>.
- Chatzimanolis, S. (2012) *Zackfalinus*, a new genus of Xanthopygina (Coleoptera: Staphylinidae: Staphylinini) with description of 20 new species. *Annals of Carnegie Museum*, **80**, 261–308. <https://doi.org/10.2992/007.080.0401>.
- Chatzimanolis, S. (2014) Phylogeny of xanthopygine rove beetles (Coleoptera) based on six molecular loci. *Systematic Entomology*, **39**, 141–149. <https://doi.org/10.1111/syen.12040>.
- Chatzimanolis, S. & Engel, M.S. (2011) A new species of *Diochus* from Baltic amber (Coleoptera, Staphylinidae, Diochini). *ZooKeys*, **73**, 65–73. <https://doi.org/10.3897/zookeys.138.1896>.
- Chatzimanolis, S. & Engel, M.S. (2013) The Fauna of Staphylininae in Dominican Amber (Coleoptera: Staphylinidae). *Annals of Carnegie Museum*, **81**, 281–294.

- Chatzimanolis, S., Cohen, I.M., Schomann, A. & Solodovnikov, A. (2010a) Molecular phylogeny of the mega-diverse rove beetle tribe Staphylinini (Insecta, Coleoptera, Staphylinidae). *Zoologica Scripta*, **39**, 436–449. <https://doi.org/10.1111/j.1463-6409.2010.00438.x>.
- Chatzimanolis, S., Engel, M.S., Newton, A.F. & Grimaldi, D.A. (2010b) New ant-like stone beetles in mid-Cretaceous amber from Myanmar (Coleoptera: Staphylinidae: Scydmaeninae). *Cretaceous Research*, **31**, 77–84. <https://doi.org/10.1016/j.cretres.2009.09.009>.
- Clarke, D.J. & Chatzimanolis, S. (2009) Antiquity and long-term morphological stasis in a group of rove beetles (Coleoptera: Staphylinidae): description of the oldest *Octavius* species from Cretaceous Burmese amber and a review of the 'Euaesthetine subgroup' fossil record. *Cretaceous Research*, **30**, 1426–1434. <https://doi.org/10.1016/j.cretres.2009.09.002>.
- Clarke, D.J. & Grebennikov, V.V. (2009) Monophyly of Euaesthetinae (Coleoptera: Staphylinidae): phylogenetic evidence from adults and larvae, review of austral genera, and new larval descriptions. *Systematic Entomology*, **34**, 246–397.
- Coiffait, H. (1972) Coléoptères Staphylinidae de la Région Paléarctique Occidentale - Sous-familles: Xantholininae et Leptotyphlinae. *Publications de La Nouvelle Revue d'Entomologie*, **2**, 115–626.
- Donoghue, M.J., Doyle, J.A., Gauthier, J., Kluge, A.G. & Rowe, T. (1989) The importance of fossils in phylogeny reconstruction. *Annual Review of Ecology and Systematics*, **20**, 431–460. <https://doi.org/10.1146/annurev.es.20.110189.002243>.
- Goloboff, P.A. & Catalano, S.A. (2016) TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics*, **32**, 221–238. <https://doi.org/10.1111/cla.12160>.
- Grebennikov, V.V. & Newton, A.F. (2009) Good-bye Scydmaenidae, or why the ant-like stone beetles should become megadiverse Staphylinidae sensu latissimo (Coleoptera). *European Journal of Entomology*, **106**, 275–301. <https://doi.org/10.14411/eje.2009.035>.
- Grimaldi, D.A. & Engel, M.S. (2005) *Evolution of the Insects*. Cambridge University Press, New York, New York. [WWW document]. URL <http://www.cambridge.org/us/academic/subjects/life-sciences/entomology/evolution-insects?format=HB&isbn=9780521821490>.
- Heer, O. (1856) Über die fossilen Insekten von Aix in der Provence. *Vierteljahrsschrift Der Naturforschenden Gesellschaft in Zürich*, **1**, 1–40.
- Herman, L.H. (2001) Catalog of the Staphylinidae (Insecta: Coleoptera). 1758 to the end of the second millennium. II. Tachyporine group. *Bulletin of the American Museum of Natural History*, **265**, 651–1066 [WWW document]. URL [papers2://publication/uuid/2E101BC2-6783-4792-9AAF-6FD1F39B01CB](https://publication/uuid/2E101BC2-6783-4792-9AAF-6FD1F39B01CB).
- Jacquelin du Val, P.M.C. (1856) Famille des Staphylinides. *Manuel Entomologique. Genera des Coléoptères d'Europe*, 2nd edn, pp. 1–41. Deyrolle, Paris.
- Jałoszyński, P. (2015) A new Eocene genus of ant-like stone beetles sheds new light on the evolution of Mastigini. *Journal of Paleontology*, **89**, 1056–1067. <https://doi.org/10.1017/jpa.2015.75>.
- Jałoszyński, P., Yamamoto, S. & Takahashi, Y. (2016) *Scydmbosetia* gen. nov., the first definite Glandulariini from Upper Cretaceous Burmese amber (Coleoptera: Staphylinidae: Scydmaeninae). *Cretaceous Research*, **65**, 59–67. <https://doi.org/10.1016/j.cretres.2016.04.011>.
- Jałoszyński, P., Brunke, A., Metscher, B., Zhang, W.-W. & Bai, M. (2017a) *Clidicostigus* gen. nov., the first Mesozoic genus of Mastigini (Coleoptera: Staphylinidae: Scydmaeninae) from Cenomanian Burmese amber. *Cretaceous Research*, **72**, 110–116. <https://doi.org/10.1016/j.cretres.2016.12.022>.
- Jałoszyński, P., Yamamoto, S. & Takahashi, Y. (2017b) A new extinct genus of *Glandulariini* with two species from Upper Cretaceous Burmese amber (Coleoptera: Staphylinidae: Scydmaeninae). *Cretaceous Research*, **72**, 142–150. <https://doi.org/10.1016/j.cretres.2016.12.021>.
- Jeannel, R. (1922) Deux Staphylinides endogés aveugles des monts Bihor. *Buletinul Societății de Științe Din Cluj*, **1**, 337–347.
- Klimaszewski, J., Brunke, A., Assing, V., Langor, D., Newton, A., Bourdon, C. *et al.* (2013) Part 2: Staphylinidae. *Synopsis of Adventive Species of Coleoptera (Insecta) Recorded from Canada*, pp. 148–149. Pensoft Publishers, Moscow.
- LeConte, J.L. (1861) Fam. X. - Staphylinidae. *Classification of the Coleoptera of North America - Part I. Prepared for the Smithsonian Institution*, 2nd edn, Vol. **3**, pp. 58–71. Smithsonian Miscellaneous Collections, Washington, DC.
- Leschen, R.A.B. & Newton, A.F. (2003) Larval description, adult feeding behaviour, and phylogenetic placement of *Megalopinus* (Coleoptera: Staphylinidae). *The Coleopterists Bulletin*, **57**, 469–493.
- Lü, L., Cai, C.-Y. & Huang, D.-Y. (2017) The earliest oxyteline rove beetle in amber and its systematic implications (Coleoptera: Staphylinidae: Oxytelinae). *Cretaceous Research*, **69**, 169–177. <https://doi.org/10.1016/j.cretres.2016.09.008>.
- Maddison, W.P. and Maddison, D.R. (2017) *Mesquite: A Modular System for Evolutionary Analysis*. [WWW document]. URL <http://mesquiteproject.org>
- McKenna, D.D., Farrell, B.D., Caterino, M.S., Farnum, C.W., Hawks, D.C., Maddison, D.R. *et al.* (2014) Phylogeny and evolution of Staphyliniformia and Scarabaeiformia: forest litter as a stepping stone for diversification of nonphytophagous beetles. *Systematic Entomology*, **40**, 35–60. <https://doi.org/10.1111/syen.12093>.
- Moore, I. (1964) A new key to the subfamilies of the Nearctic Staphylinidae and notes on their classification. *The Coleopterist's Bulletin*, **18**, 83–91.
- Moore, I. & Legner, E.F. (1973) Keys to the genera of Staphylinidae of America north of Mexico exclusive of Aleocharinae (Coleoptera: Staphylinidae). *Hilgardia*, **42**, 548–563.
- Moore, I. & Legner, E.F. (1975) *A Catalogue of the Staphylinidae of America North of Mexico (Coleoptera)*. Division of Agricultural Sciences, University of California, Riverside, California.
- Moore, I. & Legner, E.F. (1979) *An Illustrated Guide to the Genera of the Staphylinidae of America North of Mexico: Exclusive of the Aleocharinae (Coleoptera)*. Division of Agricultural Sciences, University of California, Riverside, California. [WWW document]. URL https://books.google.dk/books/about/An_illustrated_guide_to_the_genera_of_th.html?id=R9RMAAAAYAAJ&redir_esc=y.
- Naomi, S.-I. (1989) Comparative morphology of the Staphylinidae and the allied groups (Coleoptera, Staphylinidea). VII Metendosternite and wings. *Japanese Journal of Entomology*, **57**, 82–90.
- Newton, A.F. & Thayer, M.K. (1992) Current classification and family-group names in Staphyliniformia (Coleoptera) / Alfred F. Newton, Jr., Margaret K. Thayer. *Fieldiana: Zoology*, **67**, 1–92. <https://doi.org/10.5962/bhl.title.3544>.
- Nixon, K.C. (2002) *WinClada*. Published by the author, Ithaca, New York.
- Parker, J. (2016) Emergence of a superradiation: pselaphine rove beetles in mid-Cretaceous amber from Myanmar and their evolutionary implications. *Systematic Entomology*, **41**, 541–566. <https://doi.org/10.1111/syen.12173>.
- Patterson, C. (1981) Significance of fossils in determining evolutionary relationships. *Annual Review of Ecology and Systematics*, **12**, 195–223. [WWW document]. URL <http://www.annualreviews.org/doi/pdf/10.1146/annurev.es.12.110181.001211> accessed on 24 August 2017.
- Pyron, R.A. (2015) Post-molecular systematics and the future of phylogenetics. *Trends in Ecology & Evolution*, **30**, 384–389. <https://doi.org/10.1016/j.tree.2015.04.016>.

- Rambaut, A., Suchard, M. and Drummond, A. (2013) *Tracer: A Program for Analysing Results from Bayesian MCMC Programs such as BEAST & MrBayes*. [WWW document]. URL <http://tree.bio.ed.ac.uk/software/tracer/> [accessed on 24 August 2017].
- Reitter, E. (1908) Staphyliniden-Gruppen der Othiini und Xantholinini aus Europa und den angrenzenden Ländern. *Bestimmungs-Tabellen Der Europäischen Coleopteren*, **64**, 1–27.
- Ronquist, F., Teslenko, M., van der Mark, P. et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539–542. <https://doi.org/10.1093/sysbio/sys029>.
- Rood, R.J., Brunke, A. & Solodovnikov, A. (2015) *Othius punctulatus* (Goeze, 1777) (Coleoptera: Staphylinidae) detected in the Pacific Northwest: a Palearctic genus new to the USA. *The Coleopterists Bulletin*, **69**, 412–414. <https://doi.org/10.1649/0010-065X-69.3.412>.
- Schaufuss, L.W. (1888) Einige Käfer aus dem baltischen Bernsteine. *Berliner Entomologische Zeitschrift*, **32**, 266–270.
- Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. (2012) NIH image to ImageJ: 25 years of image analysis. *Nature Methods*, **9**, 671–675. <https://doi.org/10.1038/nmeth.2089>.
- Schomann, A. & Solodovnikov, A. (2012) A new genus of Staphylinidae (Coleoptera) from the Lower Cretaceous: the first fossil rove beetles from the southern hemisphere. *Systematic Entomology*, **37**, 379–386.
- Schomann, A. & Solodovnikov, A. (2017) Phylogenetic placement of the austral rove beetle genus *Hyperomma* triggers changes in classification of Paederinae (Coleoptera: Staphylinidae). *Zoologica Scripta*, **46**, 336–347. <https://doi.org/10.1111/zsc.12209>.
- Scudder, S.H. (1900) *Adephagous and Clavicorn Coleoptera from the Tertiary Deposits at Florissant, Colorado*. Monographs of the United States Geological Survey, Vol. **40**. Government Printing Office, Washington, DC.
- Shi, G., Grimaldi, D.A., Harlow, G.E. et al. (2012) Age constraint on Burmese amber based on U–Pb dating of zircons. *Cretaceous Research*, **37**, 155–163. <https://doi.org/10.1016/j.cretres.2012.03.014>.
- Smetana, A. (1982) Revision of the subfamily Xantholininae of America north of Mexico (Coleoptera: Staphylinidae). *Memoirs of the Entomological Society of Canada*, **114**, 1–389. <https://doi.org/10.4039/entm114120fv>.
- Smith, A.B. (2009) Parsimony, phylogenetic analysis, and fossils. *Systematics and the Fossil Record*, pp. 31–72. Oxford, Blackwell Science Ltd.
- Solodovnikov, A. & Newton, A.F. (2005) Phylogenetic placement of Arrowinini trib.n. within the subfamily Staphylininae (Coleoptera: Staphylinidae), with revision of the relict South African genus *Arrowinus* and description of its larva. *Systematic Entomology*, **30**, 398–441. <https://doi.org/10.1111/j.1365-3113.2004.00283.x>.
- Solodovnikov, A., Yue, Y., Tarasov, S. & Ren, D. (2013) Extinct and extant rove beetles meet in the matrix: Early Cretaceous fossils shed light on the evolution of a hyperdiverse insect lineage (Coleoptera: Staphylinidae: Staphylininae). *Cladistics*, **29**, 360–403. <https://doi.org/10.1111/j.1096-0031.2012.00433.x>.
- Swisher, C.C., Wang, Y., Wang, X., Xu, X. & Wang, Y. (1999) Cretaceous age for the feathered dinosaurs of Liaoning, China. *Nature*, **400**, 58–61. <https://doi.org/10.1038/21872>.
- Thayer, M.K. (2016) Staphylinidae Latreille, 1802. *Handbook of Zoology. Coleoptera, Beetles - Volume 1: Morphology and Systematics (Archostemata, Adephaga, Myxophaga, Polyphaga partim)*, 2nd edn (ed. by R.G. Beutel and R.A.B. Leschen), pp. 394–442. Walter de Gruyter GmbH, Berlin/Boston, Massachusetts.
- Thayer, M.K., Newton, A.F. & Chatzimanolis, S. (2012) *Prosolierius*, a new mid-Cretaceous genus of Solieriinae (Coleoptera: Staphylinidae) with three new species from Burmese amber. *Cretaceous Research*, **34**, 124–134. <https://doi.org/10.1016/j.cretres.2011.10.010>.
- Thomson, C.G. (1859) *Skandinaviens Coleoptera, synoptiskt bearbetade*, 1st edn. Berlingska Boktryckeriet, Lund.
- Wang, S., Wang, Y., Hu, H. & Li, H. (2001) The existing time of Sihetun vertebrate in western Liao – evidence from U–Pb dating of zircon. *Chinese Science Bulletin*, **46**, 779–782.
- Wiens, J.J. & Morrill, M.C. (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Systematic Biology*, **60**, 719–731. <https://doi.org/10.1093/sysbio/syr025>.
- Yamamoto, S. (2016a) The first fossil of dasycerine rove beetle (Coleoptera: Staphylinidae) from Upper Cretaceous Burmese amber: phylogenetic implications for the omaline group subfamilies. *Cretaceous Research*, **58**, 63–68. <https://doi.org/10.1016/j.cretres.2015.09.022>.
- Yamamoto, S. (2016b) The oldest tachyporine rove beetle in amber (Coleoptera, Staphylinidae): a new genus and species from Upper Cretaceous Burmese amber. *Cretaceous Research*, **65**, 163–171. <https://doi.org/10.1016/j.cretres.2016.05.001>.
- Yamamoto, S. (2017) Discovery of the oxyporine rove beetle in the Mesozoic amber and its evolutionary implications for mycophagy (Coleoptera: Staphylinidae). *Cretaceous Research*, **74**, 198–204. <https://doi.org/10.1016/j.cretres.2017.02.018>.
- Yamamoto, S. & Maruyama, M. (2017) Phylogeny of the rove beetle tribe Gymnusini sensu n. (Coleoptera: Staphylinidae: Aleocharinae): implications for the early branching events of the subfamily. *Systematic Entomology*, **43**, 183–199. <https://doi.org/10.1111/syen.12267>.
- Yamamoto, S. & Solodovnikov, A. (2016) The first fossil Megalopsidiinae (Coleoptera: Staphylinidae) from Upper Cretaceous Burmese amber and its potential for understanding basal relationships of rove beetles. *Cretaceous Research*, **59**, 140–146. <https://doi.org/10.1016/j.cretres.2015.10.024>.
- Yamamoto, S., Maruyama, M. & Parker, J. (2016) Evidence for social parasitism of early insect societies by Cretaceous rove beetles. *Nature Communications*, **7**, 13658. <https://doi.org/10.1038/ncomms13658>.
- Yin, Z., Cai, C., Huang, D. & Li, L. (2017a) A second species of the genus *Cretoleptochromus* Cai & Huang (Coleoptera: Staphylinidae: Scydmaeninae) from mid-Cretaceous Burmese amber. *Cretaceous Research*, **75**, 115–119. <https://doi.org/10.1016/j.cretres.2017.03.018>.
- Yin, Z.-W., Parker, J., Cai, C.-Y., Huang, D.-Y. & Li, L.-Z. (2017b) A new stem bythinine in Cretaceous Burmese amber and early evolution of specialized predatory behaviour in pselaphine rove beetles (Coleoptera: Staphylinidae). *Journal of Systematic Palaeontology*, **16**, 531–541. <https://doi.org/10.1080/14772019.2017.1313790>.
- Yue, Y., Ren, D. & Solodovnikov, A. (2010) *Megolisthaerus chinensis* gen. et sp.n. (Coleoptera: Staphylinidae incertae sedis): an enigmatic rove beetle lineage from the Early Cretaceous. *Insect Systematics & Evolution*, **41**, 317–327. <https://doi.org/10.1163/187631210X527034>.
- Żyła, D. & Solodovnikov, A. (2017) First extinct representative of the rove beetle subtribe Acylophorina from Baltic amber and its phylogenetic placement. *Journal of Systematic Palaeontology*, 1–11. <https://doi.org/10.1080/14772019.2017.1399171>.
- Żyła, D., Yamamoto, S., Wolf-Schwenninger, K. & Solodovnikov, A. (2017) Cretaceous origin of the unique prey-capture apparatus in mega-diverse genus: stem lineage of Steninae rove beetles discovered in Burmese amber. *Scientific Reports*, **7**, 45904. <https://doi.org/10.1038/srep45904>.

Accepted 16 April 2018

Chapter 3


Every cloud has a silver lining: X-ray micro-CT reveals Orsunius rove beetle in Rovno amber from a specimen inaccessible to light microscopy

Janina L. Kypke and Alexey Solodovnikov (2018) *Historical Biology*, DOI: 10.1080/08912963.2018.1558222.

ARTICLE



Every cloud has a silver lining: X-ray micro-CT reveals *Orsunius* rove beetle in Rovno amber from a specimen inaccessible to light microscopy

Janina L. Kypke  and Alexey Solodovnikov 

Biosystematics Section, Zoological Museum, Natural History Museum of Denmark, Copenhagen, Denmark

ABSTRACT

The exceptionally well-preserved and diverse insect fossil record of the Baltic amber group is a unique source for science. However, bubbles and thick layers of white gaseous froth cover numerous inclusions in this type of amber, rendering their examination with traditional light microscopy impossible. Here, we show that X-ray micro-computed tomography can be an efficient tool in such cases. We scanned a completely covered rove beetle (Staphylinidae) fossil in a piece of the relatively recently discovered Rovno amber. The resulting reconstruction was detailed enough to identify the fossil as the first extinct species of *Orsunius* Assing, an extant genus from the systematically very challenging subtribe Medonina (Lathrobiini, Paederinae). We summarize all previous studies on the origin of gaseous inclusions around fossil specimens in Baltic amber and discuss the technical challenges of using μ -CT where a mostly hollow specimen is surrounded by bubbles of the same density. Since recent *Orsunius* exclusively occur in the Oriental biogeographic region, we discuss the possible climate change driven speciation scenarios between *Orsunius electronefelus* sp. nov. and the former.

urn:lsid:zoobank.org:pub:8ECD641E-629F-46DC-BC81-217C529ACA84

ARTICLE HISTORY

Received 17 November 2018
Accepted 7 December 2018

KEYWORDS

X-ray micro-CT; Rovno amber; *Orsunius*; Medonina; Paederinae; Staphylinidae

Introduction

Contemporary quantitative evolutionary biology strives to include fossil species in seamless datasets with their extant taxa for phylogeny reconstruction, biogeographic inference or palaeoenvironmental research. All these fields of science rely on the detailed morphological examination of fossil species (Smith 2009; John and Birks 2012; Laflamme et al. 2014; Bolton and Beaudoin 2017; Kuzmina 2017). This is often challenging, because even the best-preserved fossils retain only fragments of their phenotype, what makes observations of their important characters difficult to sometimes impossible. Amber inclusions of smaller organisms, especially insects, are a unique type of fossils where skeletal structures of extinct species are preserved so well that they can be studied with the same level of detail as in extant species (Grimaldi and Engel 2005; McCoy et al. 2018). Amber is also special for palaeontology since the resin often entraps soft-bodied animals underrepresented in the fossil record (Grimaldi 2009). Overall, amber provides an enormous source of fossils, especially the Baltic amber group, which is famous for being the most diverse record of fossil insects from the Eocene (Bogri et al. 2018). In particular, shifts in the diversity of insect species in reaction to the changing temperatures and carbon dioxide levels during the Eocene (Zanazzi et al. 2007; Liu et al. 2009; Hren et al. 2013) that can be traced via Baltic amber inclusions, is the best available analogue of how our current biota will be affected and react to the changing climate. Correct species identifications are crucial for such palaeoclimate studies as well, and the very life-like preservation of insects and other organisms in amber can be of

advantage in this process. However, even if the amber is light-coloured and transparent, unfortunate 'death postures' of the trapped animals, pieces of dirt, plant remains, and especially widespread and bothersome fluid or gaseous inclusions might still hinder a detailed study of important structures.

X-ray micro-computed tomography (μ -CT) appeared an accessible and powerful method for the examination of the sub-optimally preserved Baltic amber specimens. It has repeatedly been used for the examination of various amber-imbedded fossils (Grimaldi et al. 2000; Dierick et al. 2007; Lak et al. 2008; Labandeira 2014), but not extensively enough to be considered a common practice for the study of amber inclusions. Here, we chose to scan a rove beetle specimen entrapped in Rovno amber, which is part of the Baltic amber group (Perkovsky et al. 2003a, 2003b). The rove beetles are a taxonomically very diverse and hence challenging group of organisms that is common in the Baltic amber group (Bogri et al. 2018). This particular specimen was so strongly surrounded by a cloudy cover of gaseous froth and bubbles that it was only possible to guess that it was a member of the subfamily Paederinae based on the tip of antennae, tibiae, small parts of pronotum and vague contours of the habitus accessible from light microscopy (Figure 1). Such specimens in Baltic amber partially or totally inaccessible for the light microscopy are very common, but mostly neglected. This bias is highly unfortunate and undesirable in science, especially when it comes to the palaeoclimate research where relative abundance of specimens also matters. In this experiment, we tested to which extent this possibly hollow specimen, entirely covered by a thick

CONTACT Alexey Solodovnikov  asolodovnikov@snm.ku.dk  Biosystematics Section, Zoological Museum, Natural History Museum of Denmark, Universitetsparken 15, Copenhagen, DK-2100 Denmark

<http://zoobank.org/urn:lsid:zoobank.org>

 Supplementary data for this article can be accessed [here](#).

© 2018 Informa UK Limited, trading as Taylor & Francis Group

Published online 20 Dec 2018

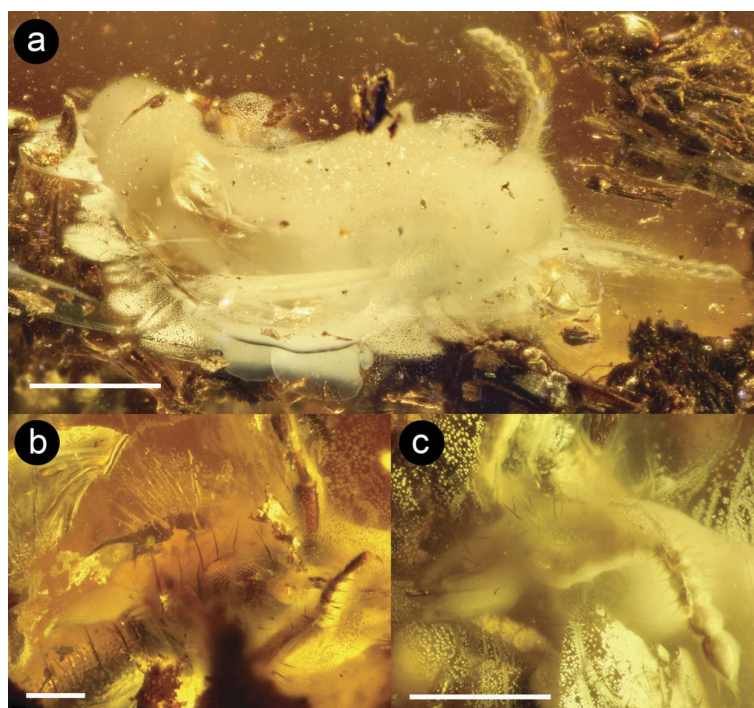


Figure 1. *Orsunius electronefelus* sp. nov., holotype K-7181, covered in milky froth and entrapped in Rovno amber from the Klesov deposit, Ukraine. (a) dorsal habitus. (b–c) lateral view of head and pronotum. Scale bar represents (a) 1 mm. (b) 0.1 mm. (c) 0.5 mm.

layer of air microbubbles, can be studied using μ -CT. Moreover, Rovno amber is a relatively recently discovered type of the Baltic amber group that remains understudied and our find is a new addition to the current fossil record.

Using X-ray μ -CT we were able to identify our specimen as an extinct species within the extant group of species of *Orsunius*, the recently described genus within the systematically very challenging subtribe Medonina of the tribe Lathrobiini. Here, we present the taxonomic and methodological results of this morphological reconstruction. Also, we summarize all previous studies on gaseous inclusions in amber that form the unfortunate clouds frequently blocking fossils in Baltic amber from observation. Finally, we discuss the challenges of the taxonomic study of amber fossils using advanced technologies like μ -CT where a mostly hollow specimen is surrounded by bubbles of the same density.

Geological setting and provenance

Baltic amber is a complex of ambers with different source areas spread throughout northern Europe (Bogri et al. 2018). All these ambers contain elevated concentrations of succinic acid and related succinates (Wolfe et al. 2009). While the exact species of the source tree that produced the resin has not been determined yet, micro-Fourier transform infrared spectroscopy (μ -FTIR) as well as palaeobotanical analyses lead to the proposition of it being a Palaeogene sciadopityaceous conifer (Wolfe et al. 2009). Despite extensive research on Baltic amber, determining its provenance time and source areas has been challenging and much debated (Perkovsky et al. 2007; Nadein et al. 2016; Bogri et al. 2018; Mänd et al. 2018). Propositions range from Eocene to Oligocene age

(56–23 Ma), however, with the majority proposing the Eocene (reviewed in Nadein et al. 2016; Bogri et al. 2018). Reasons for this imprecise age estimate are the large Palaeogene forests covering varying climatic zones, substantial climate change during the Eocene and into the Oligocene, as well as possible re-deposition of amber from the sources to today's deposits that excludes stratigraphic dating as a reliable option to determine the provenance time.

One of the more riddling amber types is from the Rovno region, a succinite from Ukraine and Belarus that has only been studied since the year 2000 (Szwedo and Sontag 2013). Both provenance and autochthony of Rovno amber compared to Baltic amber from deposits in Germany, Poland, Denmark, Lithuania or Russia have been questioned in the past (Perkovsky et al. 2007; Sontag and Szadziwski 2011; Szwedo and Sontag 2013; Perkovsky 2016). Although comparative studies of the palaeoentomofauna found differences in the abundance of common taxa and numerous taxa unique to the Rovno deposit (Perkovsky et al. 2007, 2010; Dlussky and Rasnitsyn 2009; Perkovsky 2016; Petrov and Perkovsky 2018), these differences have not been tested for their statistical significance and are hence being doubted (Szwedo and Sontag 2013). A recent analysis of the chemical-physical properties of Rovno and other types of Baltic amber found similar $\delta^{13}\text{C}$ values and μ -FTIR spectra, which point to their contemporaneous origin with similar source tree (Mänd et al. 2018). Stratigraphic analyses of the Mezhygorje Formation, the main source for Rovno amber, has been dated to be from the Early Oligocene (Rupelian age 33.9–28.4 Ma) (Jałoszyński and Perkovsky 2016). Since there is no precise date for other types of Baltic amber and there is even an assumption for a certain time range for the origin of various amber pieces, this could mean that Rovno

amber is of Early Oligocene, and hence younger than mostly expected. On the other hand, Mänd et al. (2018) found slight differences between $\delta^2\text{H}$ values of Rovno amber pieces collected in the same deposit, which suggest that the amber has been reworked from its source area. This means that the age of the Mezhygorje Formation should not be used as a reference for the age of Rovno amber deposited there, which is likely older. The latter hypothesis is further supported by substantial differences in $\delta^2\text{H}$ between Rovno and Baltic amber from the south-east coast of the Baltic Sea. These isotopic differences suggest a different water source that the resin-producing conifers accessed. Moreover, considering today's isoscape of Central Europe, these results place the source of Rovno amber south of the other tested Baltic amber sources. Apparently, the palaeoclimate of the Rovno amber source area was even warmer than at the area of origin of the other types of Baltic amber now found at the Baltic Sea coast. And potentially, the source of the Rovno amber, on the one hand, and other types of Baltic amber, on the other hand, were even separated by the Paratethys Sea (Mänd et al. 2018).

Material and methods

The amber piece, labelled K-7181, is from the Rovno region, mined in Klesov, Ukraine. It is coppery-yellow, transparent and, next to the specimen, includes wood fragments and gaseous or gas-liquid inclusions of varying size (Figure 1). The piece is deposited at IZAN, the I. I. Schmalhausen Institute of Zoology, National Academy of Sciences of Ukraine, Kiev.

Tomography and visualisation

X-ray μ -CT scans were acquired through the Cornell University Biotechnology Resource Center using a Zeiss-Xradia Versa 520 X-ray Microscope. Sample preparation was minimal. The amber piece was put into a small plastic container and held in place with soft polyurethane foam before being mounted on the scanner stage. It was scanned at 60 kV and 5 W, without a filter and reconstructed using the Zeiss reconstruction software from 1601 fluoroscopy projections taken over 360 degrees. The fluoroscopy image exposure times varied from scan to scan, but were sufficient to get good contrast between the amber, exoskeleton, and interior of the specimen (these times varied between 3 and 10 s).

The fossil was scanned several times to display different anatomical regions. A scan at 8.05 microns/pixel displayed the entire specimen. Higher resolution scans of the head and lower abdomen were taken at 1.86 microns/pixel.

For easier visualisation of the low-density space formed by the insect fossil surrounded by the amber of a higher density, the reconstructed image was inverted using Avizo Software (Avizo v. 9.4, Thermo Scientific™, <https://www.fei.com/software/amira-avizo/>), so that the empty space inside the amber would be more visible when performing volume renderings. For the final images, other inclusions surrounding the specimen like plant material, air or water inclusions were discarded. Where layers of large or small air bubbles covered the hollow specimen, their densities were equal to the specimen itself and hence they were initially automatically reconstructed as parts of the fossil

(Appendices, S1-2). In an additional step they were removed manually where possible, to reveal the exoskeleton of the fossil. The data have been deposited in Zenodo and have been assigned the following DOI: 10.5281/zenodo.2148279.

A short animation movie showing the 3D reconstruction in rotation was taken in Horos software v. 3.0 (Horosproject.org) and final adjustments were made in Adobe Premiere Pro CS6 v. 6.0.5 (Adobe Systems Inc. 1991–2012) (Appendices, S1).

Horos software was also used to take measurements using the whole-body scans (in mm). They are abbreviated as follows: HL, head length (from apex of clypeus to neck constriction); HW, maximal head width; PL, length of pronotum (along medial line); PW, maximal pronotum width; EL, elytral length (from apex of scutellum to the level of most distal extension of elytral apical margin); EW, combined width of both elytra; FL, forebody length (calculated as the sum of HL+PL+EL); TBL, total body length (sum of FL and length of abdomen).

Microscopy and illustration

The fossil specimen was examined using a Leica M205 C stereoscope. Pictures were taken with a Canon EOS 6D camera attached to the stereoscope using EOS Utility 3.4.30.0 software. Photographs were stacked in Zerene Stacker (Zerene Systems LLC, 2012) and edited in Adobe Photoshop CS6 Extended v. 13.0.6 (Adobe Systems Inc. 1990–2012). Based on the screenshots of the final reconstruction of the fossil, a habitus line drawing was created in Adobe Illustrator CS6 v. 16.0.4 (Adobe Systems Inc. 1987–2012).

The distribution map of *Orsunius* was made in SimpleMappr (Shorthouse 2010) and finalised in Adobe Illustrator CS6 v. 16.0.4 (Adobe Systems Inc. 1987–2012). Coordinates for recent species were taken from Assing (2011a, 2014, 2015) and the location of the Rovno source area was taken from Mänd et al. (2018).

Systematic palaeontology

This published work and the nomenclatural acts it contains have been registered with Zoobank: urn:lsid:zoobank.org:pub:8ECD641E-629F-46DC-BC81-217C529ACA84.

Class **Insecta** Linnaeus, 1758
Order **Coleoptera** Linnaeus, 1758
Family **Staphylinidae** Latreille, 1802
Subfamily **Paederinae** Fleming, 1821
Tribe **Medonina** Casey, 1905
Genus ***Orsunius*** Assing, 2011

†*Orsunius electronefelus* sp. nov.
(Figures 2, 3)

Type species. Holotype, ♂, K7181, a well-preserved specimen fully covered by milky gaseous froth inside a coppery-orange, transparent piece of amber (c. 2.5 × 0.8 × 0.9 mm) (IZAN).

Diagnosis. Based on the characters provided in the section ‘Taxonomic placement’ (below) this new *Orsunius* species is

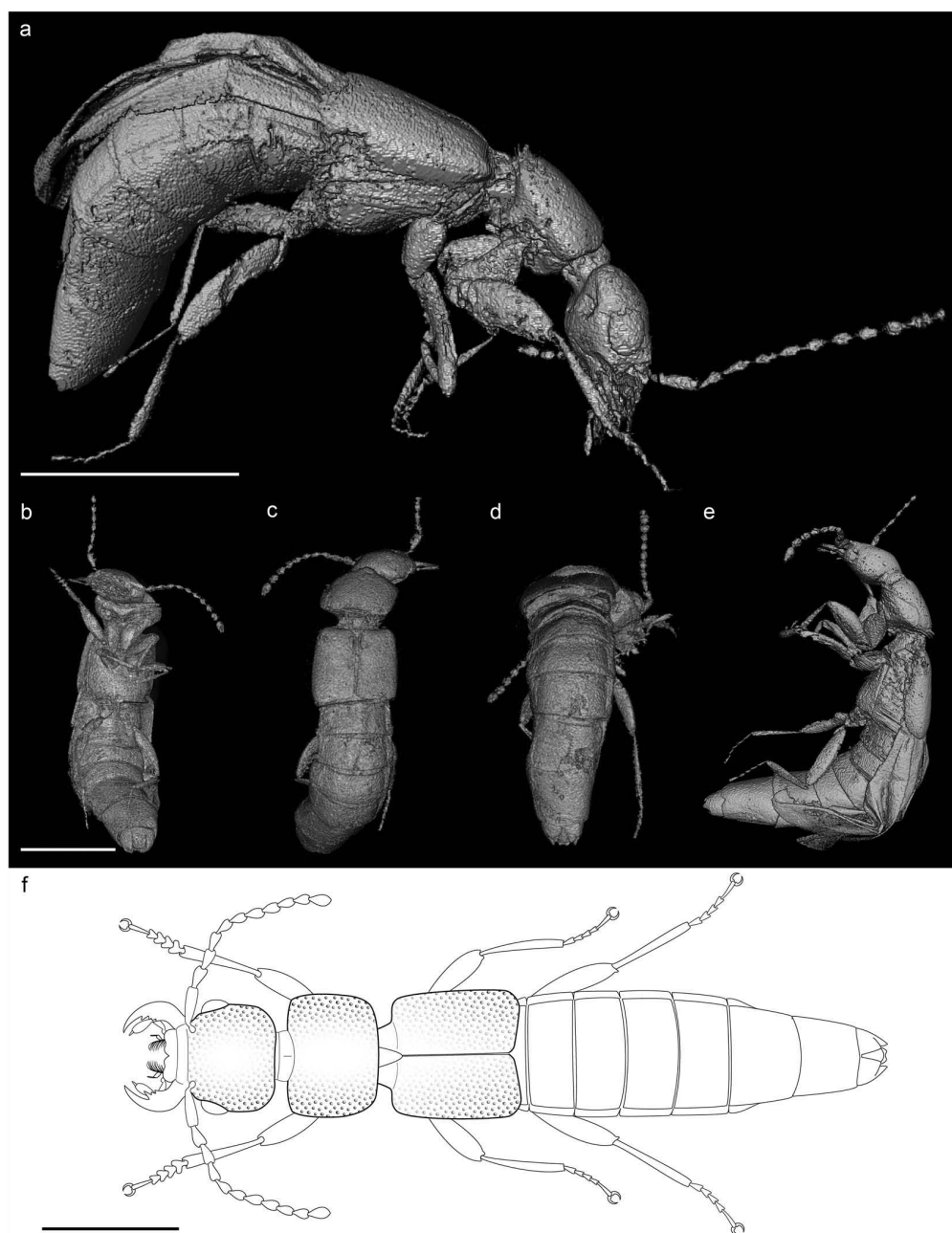


Figure 2. 3D reconstruction from μ -CT scans at varying angles and artistically reconstructed habitus of *Orsunius electronefelus* sp. nov., holotype K-7181. (a & e) lateral habitus. (b) ventral habitus. (c & d) dorsal habitus. (e), line drawing of dorsal habitus based on details from 3D reconstruction. Scale bars represent 1 mm.

placed in Assing's (2014) monophylum containing (*O. granulosus* Assing 2014, *O. cuneatus*, 2014 and *O. heissi*, 2014) (*granulosus*-group *sensu* Assing 2015). From all three extant species of the *granulosus*-group the fossil *O. electronefelus* sp. nov. differs in larger body and quadratic rather than transversely rectangular head.

Derivation of name. The species name is a compound derived from two ancient Greek words: 'electron' (ἤλεκτρον, pronounced 'i.lek.tron) is the amber and 'nefeli' (νεφέλη) is the nebula. For the conjugation with the masculine Latin genus *Orsunius*, we have Latinized the species epithet with the suffix 'us'.

Material. The preservation status of the fossil was hard to judge before seeing the results of the scan since it is almost entirely within a thick white cloud of (microscopically) small bubbles. However, parts of the head, pronotum and legs are visible and suggest a good preservation status. On these body parts, small details like punctation and setation can be observed. The abdomen seems to have been slightly squished at places, making it difficult to judge the correct shape.

Occurrence. Rovno amber, Klesov deposit, Ukraine. Upper Eocene.

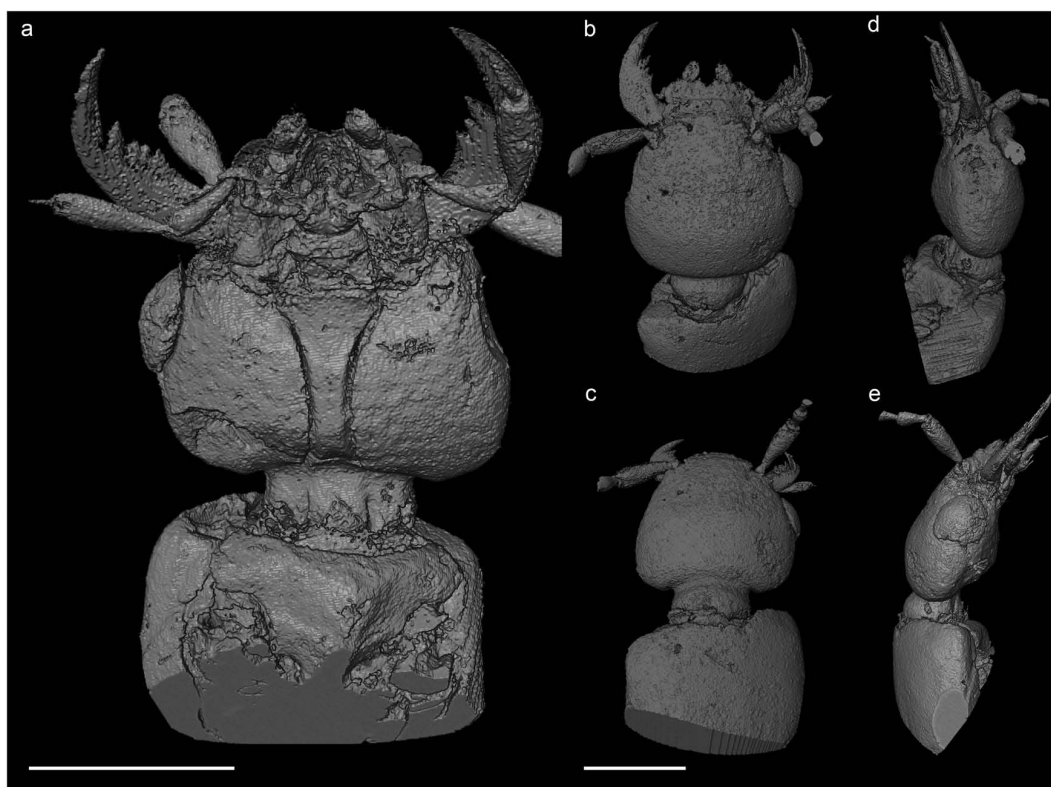


Figure 3. 3D reconstruction of head and partial pronotum of *Orsunius electronefelus* sp. nov. from μ -CT scans, holotype K-7181 at higher resolution. (a) ventral view. (b & c) dorsal view. (d & e) lateral view. Scale bars represent 0.5 mm.

Description

Habitus. With a TBL: 4.89 mm a rather large species. FL: 2.09 mm with coarse, granulose and dense punctation (where visible under the light microscope).

Head. Almost quadratic, slightly wider than long (HW: 0.7 mm; HL: 0.69 mm) with rounded but pronounced hind angles and concave posterior margin in the middle; at least laterally with large macrosetae that can be seen through the microscope but not in the scan. The exact position of macrosetae remains uncertain. Eyes large, 0.74 times as long as temples, and protruding over head-contour. Labrum bilobed, with shallow but rather broad V-shaped median incision at anterior margin. Antennae setose and long (1.51 mm), longer than head and pronotum combined; antennomere 1 almost twice as long as the following antennomere; antennomere 3 slightly longer than 2; antennomeres 4–9, each, slightly shorter than 3, ellipsoid and of similar size; antennomeres 10–11 increasing in diameter, appearing rounder. Mandibles slightly curved and broad; left with three teeth, right with four teeth on inner margin; mandible tips probably crossed-over in resting position. Maxillae appear as two large lobes, an artefact of the reconstruction. Last (fourth) maxillary palpomere acicular and small, much shorter than preceding palpomere; the latter three times as long as broad and dilating apicad. Last labial palpomere needle-shaped and about half as long as preceding palpomere. Ligula bilobed. Gular sutures concave but do not meet. Neck *ca.* half as wide as head with well-developed dorsal and ventral constriction.

Pronotum. Almost as long as head but wider (PW: 0.77 mm; PL: 0.62 mm) with dense, distinct punctation and large macrosetae at lateral margins.

Elytra. Wider and longer than pronotum (EW: 0.96 mm; EL: 0.83 mm).

Hind wings. Fully developed.

Legs. Protarsomeres 1–4 dilated; meso- and metatarsomeres 1 slightly shorter than tarsomeres 2 and 3 combined.

Abdomen. Narrower than elytra and slightly tapering apicad, with two pairs of paratergites visible on segment VII in 3D reconstruction. Posterior margin of sternite VII straight; sternite VIII as wide as VII and only slightly tapering with weakly developed median incision on posterior margin; sternite IX comparatively wide, 1.6 times as broad as long with slightly concave posterior margin. Tergite IX incised in the middle along all its visible length; tergite X with slightly concave posterior margin.

Systematic placement. Based on the habitus resemblance and morphological features such as the strongly transverse labrum, the large head with pronounced posterior angles, dilated protarsi 1–4, superior marginal line of pronotum going downwards in the anterior half of pronotal length, the shallow incision at the posterior margin of male sternite VIII and the deeply incised tergite IX, the fossil is placed in the extant medonine genus *Orsunius* confined to the Oriental region in

distribution. Furthermore, within that genus *O. electronefelus* sp. nov. can be assigned to Assing's (2014) monophylum containing *O. granulosus*, *O. cuneatus* and *O. heissi* which share the presence of the fourth tooth on the right mandible, the rather broad V-shaped median incision of the labrum and the granulose punctation on the pronotum. Considering very poor taxonomic and phylogenetic knowledge of Medonina, a very speciose group, especially in the Oriental region, that comprises many convergent morphotypes, and still unavailable structures in our fossil, for example the aedeagus, the systematic assignment of *Orsunius electronefelus* should be tested through future investigations of rich assemblages of Baltic amber paederines along with a stronger phylogenetic approach to the Medonina in general.

Results

The specimen entrapped in Rovno amber was successfully scanned using X-ray μ -CT and the resulting 3D reconstruction was detailed enough to describe and systematically place the fossil (see systematic section for details). As expected, it turned out to be mostly hollow. No inner soft tissues have been preserved but a few strongly sclerotized endoskeletal structures, for example the apodemes of the tentorium visible in the semi-transparent image of the high-resolution scan of the head (Figure 4(a)). Unfortunately, the few remaining internal structures at the tip of the abdomen, that are situated where the aedeagus can normally be found, could not be reconstructed – neither from the whole body scan nor the high-resolution scan of only the tip of the abdomen (Figure 4(b)). It seems that softer parts of the aedeagus have been decomposed and lost.

Discussion

X-ray micro-computed tomography as a tool to study invertebrates

Three-dimensional (3D) images can be obtained using computed tomography (CT) where an X-ray beam passes through a sample at various viewing angles and a 3D image is reconstructed based on measurements of varying absorptions and attenuations of the beam (Herman 1985). Originally, this technique was developed for clinical diagnostic studies of human anatomy and physiology but the technology evolved quickly (Kalender 2006). Beginning with a resolution of 1 mm, today μ -CT scanners can image objects with a resolution down to a fraction of a micrometre depending on the X-ray source and the object that is scanned. A synchrotron X-ray (SR) source has the highest brilliance and is leading to more detailed images with an effective pixel size as small as 0.25 μ m, but even conventional μ -CT scanners can create images with resolutions <1 μ m (Chappard et al. 2006; Betz et al. 2007).

Since the first reported μ -CT scan of an insect, which in fact was a robber fly (Diptera: Asilidae) entrapped in Baltic amber and covered in milky froth (Grimaldi et al. 2000), entomologists have used μ -CT scanning to better understand the physiology, anatomy, palaeontology and even behaviour

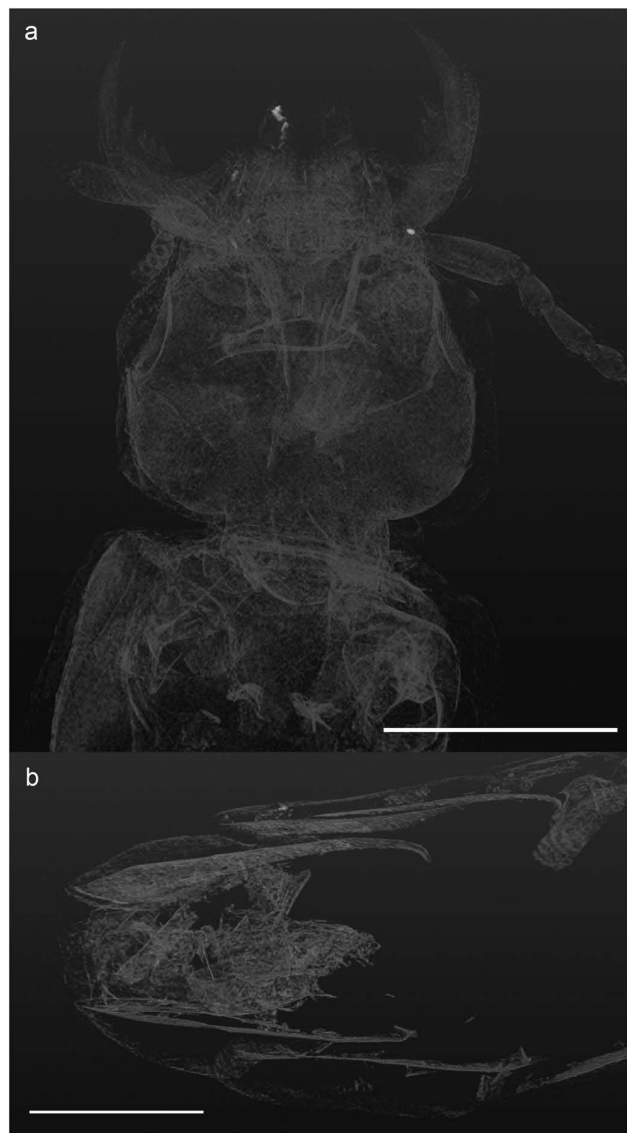


Figure 4. Semi-transparent images of head and lower abdomen in high-resolution μ -CT scans (1.86 microns/pixel). (a) head. (b) tip of abdomen, cropped. Scale bars represent (a) 0.5 mm. (b) 0.25 mm.

or locomotion of insects (e.g. Hörnschemeyer et al. 2002; Tafforeau et al. 2006; Lak et al. 2008; Kirejtshuk et al. 2009; Verdú et al. 2012; Simonsen and Kitching 2014; Smith et al. 2016). Additionally, μ -CT scans have found applications in pest control, e.g. tracing the larvae of wood-boring insects (Jennings and Austin 2011) and recently, they permitted to successfully scan insects alive (Poinapen et al. 2017) which opens up even more new applications.

In rove beetles, μ -CT has been used as a non-destructive alternative to classical microtomy in order to study the morphology of recent taxa and to understand the connections of internal and external body structures (Betz et al. 2007; Weide and Betz 2009; Zhang et al. 2010; Li et al. 2011). So far, only very few fossil rove beetles entombed in Burmese and Baltic amber from Poland and Russia have been scanned using μ -CT; one of them is an aleocharine (Yamamoto and Maruyama 2018) and the remaining four are scydmaenines

(Jałoszyński et al. 2017, 2018). None of them, however, was fully covered by gaseous froth and thus as inaccessible for the light microscopy as the rove beetle specimen we studied here.

Challenges using X-ray micro-computed tomography to scan amber inclusions hidden under gaseous froth

Gaseous froth and bubbles of varying size often surround fossil inclusions in amber. They render the study of an embedded fossil using traditional and most widely used light microscopy impossible and hence highly diminish the number of specimens available for examination. It might in fact create an inadvertent bias in the study of certain taxonomic groups over others if they are more affected by covering froth than others. This is cause for concern for comparative palaeoecological studies where the abundance of certain taxa is as important as their diversity. Such areas of 'white clouds' that cover inclusions to different extent are very common in Baltic amber but nonetheless have not been studied extensively.

Resembling the white cottony mycelium of a fungus these white impurities were first attributed to a Phycomycete (Berendt 1845), later assigned to be results of bacterial activities decomposing the trapped invertebrates (Schlüter and Kühne 1975; Mierzejewski 1978) and are now thought to be the visible results of a dehydration process that took place *post-mortem* (Weitschat & Wichard 1998/2002 cited by Judson 2003). In this case, irregular cloudy covers can be explained if the animal was not trapped entirely at first so that any liquid of its exposed body parts could evaporate before being covered by another layer of resin. As a result, only body parts entrapped in the resin first become covered by white impurities (Judson 2003). Since this phenomenon is predominantly known to occur in ambers of the Baltic amber group, it seems reasonable to assume that the chemical composition of the resin might also have influenced the formation of such white clouds. Statistical tests show that differences in the preservation of fossils trapped in amber belonging to different chemical groups (e.g. Baltic amber vs. Dominican amber) are significant (McCoy et al. 2018). One way how Baltic amber differs from others is that it contains comparatively high amounts of succinic acid and other succinates (Wolfe et al. 2009; Nadein et al. 2016; Mänd et al. 2018), but whether or how they might have interacted with entrapped animals and if they are responsible for the formation of white clouds around them has not yet been studied.

Gaseous froth is not the only potential setback when working with amber. The preservation of fossils can vary from being completely intact, not only outside but also inside, to being completely empty with nothing but a thin carbon layer left over that reflects what used to be tissue (Dierick et al. 2007). McCoy et al. (2018) report that in Baltic amber (not including Rovno amber) internal structures were preserved in 55% of the known fossils that have been scanned. Since all published studies of fossils in Rovno amber are limited to their external morphology and the only existing synchrotron scan of a chrysomelid beetle in Rovno amber (Nadein et al. 2016) only reports external structures, chances for finding fossilised internal tissues in our target rove beetle were

unknown. However, considering that Rovno amber most likely stems from resin produced by the same source tree as other types of Baltic amber, a similar preservation to the latter might be expected. Finding a mostly empty specimen with almost no internally preserved structures is therefore not very surprising. In general, however, we know very little about the degree of internal preservation of Baltic amber specimens because the great majority of the specimens were examined using a light microscope only.

Albeit a satisfying final product, a 3D model of the fossil that could successfully be used to describe a species, one major problem in the process of reconstructing the 3D image of the fossil was to separate any froth and larger bubbles that were directly attached to the body of the beetle (Figure 5(a,b)).

In the process of image segmentation, a variety of software is available to help to partition the image into regions that are alike with respect to, e.g. material, structure or function. This way, they can subsequently be studied in isolation, independently coloured, measured or completely removed from the final image. An inversion of the image's raw data helps in the segmentation of the 3D volume because a hollow specimen of low density is 'filled' while the amber of high density is 'emptied'. That way the fossil image can be reconstructed just like any other object surrounded by air instead of an amber matrix. Segmentation can be performed relatively easy if these regions differ in their densities which any software would visualize in varying shades of grey, respectively, with higher contrast between structures of more different densities (Figure 5(c,d)). Using a variety of selection tools that the software offers, regions of the same intensity can easily be selected. Depending on the tool used, regions will be selected in 2D or 3D. Selected regions can be saved in separate layers for further analysis. In our case, where bubbles were directly attached to the hollow beetle, no tool was able to distinguish between them as they had the same density and hence appeared to be the same. This meant that more time needed to be invested to hand-select any bubbles attached to the beetle so that they could be removed layer by layer. This is very difficult in a 3D image where one can only work on a 2D plane, so that the surface of some body parts could not be reconstructed perfectly and appears slightly uneven.

***Orsunius electronefelus* sp. nov. in the context of *Medonina* systematics and the palaeoenvironment**

As shown in detail in the Systematic Palaeontology section, μ -CT revealed the morphology of our fossil in such detail that it could be placed in the *granulosus*-group *sensu* Assing (2015) of *Orsunius*, an extant genus from the paederine subtribe *Medonina* (Casey 1905). *Medonina* is a species-rich assemblage of relatively few poorly defined genera, of which only some have recently been revised or newly described and hence have clear synapomorphies or diagnostic character-combinations separating them from the others (Assing 2011a, 2011b, 2011c, 2013, 2018). On the other hand, the majority of medonines have not yet been revised and phylogenetic relationships between their species and genera remain uncertain. While the medonines of the West Palearctic are the



Figure 5. Example of manual segmentation process to remove bubbles attached to *O. electronefelus* sp. nov., holotype K-7181. (a) lateral habitus after automated reconstruction. (b) lateral habitus with bubbles selected in red. (c) cross section through 3D scan with beetle in light grey and bubbles selected in red. (d) as in (c) but longitudinal section. Scale bars represent 1 mm.

ones studied best, most of the members of this group, if we account for undescribed species as well, occur in the subtropical to tropical areas globally (A. Solodovnikov, unpub. data; A. Newton, pers. comm.). *Orsunius* is one of such thermophilic genera, which today is widely spread at lower to intermediate elevations from the very south of the East Palearctic through Oriental region south-eastwards to the Wallace line (Figure 6) (Assing 2011c, 2014). Currently, there are 22 extant species described in *Orsunius* (Assing 2015) and *O. electronefelus* sp. nov. is the first extinct species discovered in this genus.

The hitherto described arthropod fauna in Rovno amber is inconclusive with regards to its climatic preference compared to other succinites. There are species that are unique to the Rovno amber but there are also many species that occur in other types of Baltic amber (Perkovsky et al. 2007; Dlussky and Rasnitsyn 2009; Perkovsky 2016). Additionally, species described from Rovno amber belong to the genera whose recent representatives have diverse types of distribution ranges, from tropical to Holarctic (Alekseev and Alekseev 2016; Perkovsky 2016; Petrov and Perkovsky 2018). However, stable isotope analyses suggest that the source of Rovno amber was further south than the source of other types of the Baltic amber. It was generally warm and winters were mild with a predominantly subtropical palaeoflora during the time where and when Rovno amber was formed (Perkovsky 2016; Mänd et al. 2018; Sokoloff et al. 2018). This suggests that *O. electronefelus* sp. nov. was a thermophilic species, which is fully congruent with the thermopreference of its extant congeners. There are different possible scenarios that could explain the change in the genus distribution from then to now. All of them need to consider the cooling climate at the transition from the Eocene to the Oligocene (Zanazzi et al. 2007; Liu et al. 2009; Hren et al. 2013), which rendered the habitat unsuitable for *Orsunius*. One option is that *Orsunius* had a narrow distribution range in the palaeoarea, which now is part of Europe, and a cooling climate could have forced *Orsunius* to migrate towards the warmer refugium in South-East Asia (Figure 6). Alternatively, *Orsunius* could have had a very wide distribution range across the area now forming Eurasia. In this case, a changing climate might have led to an extinction of all European congeners, leaving behind only species in the Oriental region. Amongst all to date described extant species within the genus, the majority have three mandibular teeth on both mandibles and three species of the *granulosus*-group *sensu* Assing (2015), like the fossil, have an additional fourth tooth on the right mandible. Assuming that two groups with regard to the number of mandibular teeth may be monophyla, the discovery of a fossil with four teeth on the right mandible in the western part of the Eurasian palaeoarea suggests the origin of the four-toothed lineage there. While the rich extant fauna of the three-toothed lineage in South-East Asia is easiest to explain by their *in situ* origin in that area. These hypotheses suggest an ancestral wide distribution range for *Orsunius* and allopatric origin of its four- and three-toothed lineages, respectively, with the subsequent south-eastwards migration of the four-toothed species to the area now corresponding to

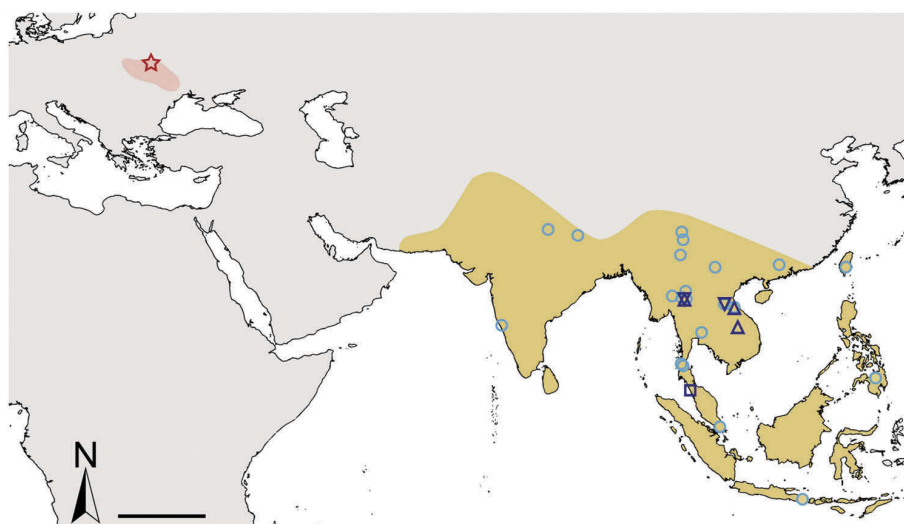


Figure 6. Distribution map of the genus *Orsunius*. Red star, the Rovno amber Klesov deposit, Ukraine, where *O. electronefelus* sp. nov. was found. Collecting localities for recent species are shown in blue. Circles, *Orsunius* spp. with even number of molar teeth (3 + 3). Inverse triangle, *O. cuneatus*. Triangle, *O. granulatus*. Square, *O. heissi*. Red background, suggested source area of Rovno amber (after Mänd et al. 2018). Yellow background, the oriental biogeographic region. Scale bar represents 1320 km.

South-East Asia which was triggered by the cooling climate. Eventually, distributions of both lineages overlapped to produce the recent distribution of the genus (Figure 6). Combined phylogenetic analyses of the extant and extinct species could shed more light on the value of the mandibular character, as well as the monophyly of *Orsunius* and evolution and systematics of this genus overall.

Conclusions

In a case like our amber fossil fully covered by milky gaseous froth, creating a 3D μ -CT scan is the only viable alternative to a light microscope, even though the specimen is hollow, i.e. has the same density as air bubbles. A 3D reconstruction can be manipulated and cropped in order to study the morphology of the concealed body parts. Major obstacles that we encountered in the process of image segmentation were those gaseous inclusions that were in direct contact with the insect body. They cannot yet be removed in an automated and fast way, which leaves room for improvement. Also, creating 3D images using μ -CT is often not an option that is freely available and the learning curve for any available software is steep. Still, the inversion of the scanned data facilitates the reconstruction of fossils trapped in amber and is recommended. Synchrotron μ -CT might generally be an even better option compared to conventional μ -CT since it leads to more detailed images in which setation or micro-sculpture might actually be visible, too. However, such facilities are even more limited and thus it takes time to get a slot to use a synchrotron and it is more expensive.

X-ray μ -CT enabled us to clearly identify the first fossil species in the recent genus *Orsunius* from the taxonomically challenging group of Paederinae rove beetles. *O. electronefelus* sp. nov. is by far not the only described species in Rovno and other ambers of the Baltic region whose extant relatives also

only occur in warmer areas of South-East Asia (summarized in Bogri et al. 2018). Considering a distinct source area for the species found in amber from Rovno, more work on the identification of source areas for Baltic ambers from more northern deposits will hopefully help to better understand distribution patterns of the palaeofauna of the Late Eocene. Additionally, the larger the taxonomic inventory and the more detail is provided for each species, the better the palaeofauna will be as a proxy for palaeoclimatic reconstructions, for evolutionary studies and as a basis to model outcomes of the current climate change.

Acknowledgments

We would like to thank A. Petrenko via E.E. Perkovsky for the loan of the material. We are also thankful to Teresa Porri, CT Manager at the Imaging Facility of the Cornell University Biotechnology Resource Center for providing the imaging data, acquired with NIH S10OD012287 funding for the ZEISS-Xradia Versa 520 X-ray Microscope, and more importantly, with hands-on help in reconstructing and analysing the scan. Furthermore, we thank Laura Tufano for giving JLK a quick introduction to film-making and the constructive feedback on the animated movie. Dagmara Żyła is greatly acknowledged for contributing specimens and identifications to our sample of Baltic amber Paederinae and multiple discussions about palaeontology and systematics that helped writing this paper as well. This study has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 642241 (PhD fellowship of JLK). This material reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains.

Author contribution

JK and AS designed the study. JK performed μ -CT scans and all subsequent work to reconstruct, illustrate and identify the fossil species. She drafted the paper which was finalized together with AS. The authors declare no conflicting financial interests.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the H2020 Marie Skłodowska-Curie Actions [642241]; NIH [S10OD012287].

ORCID

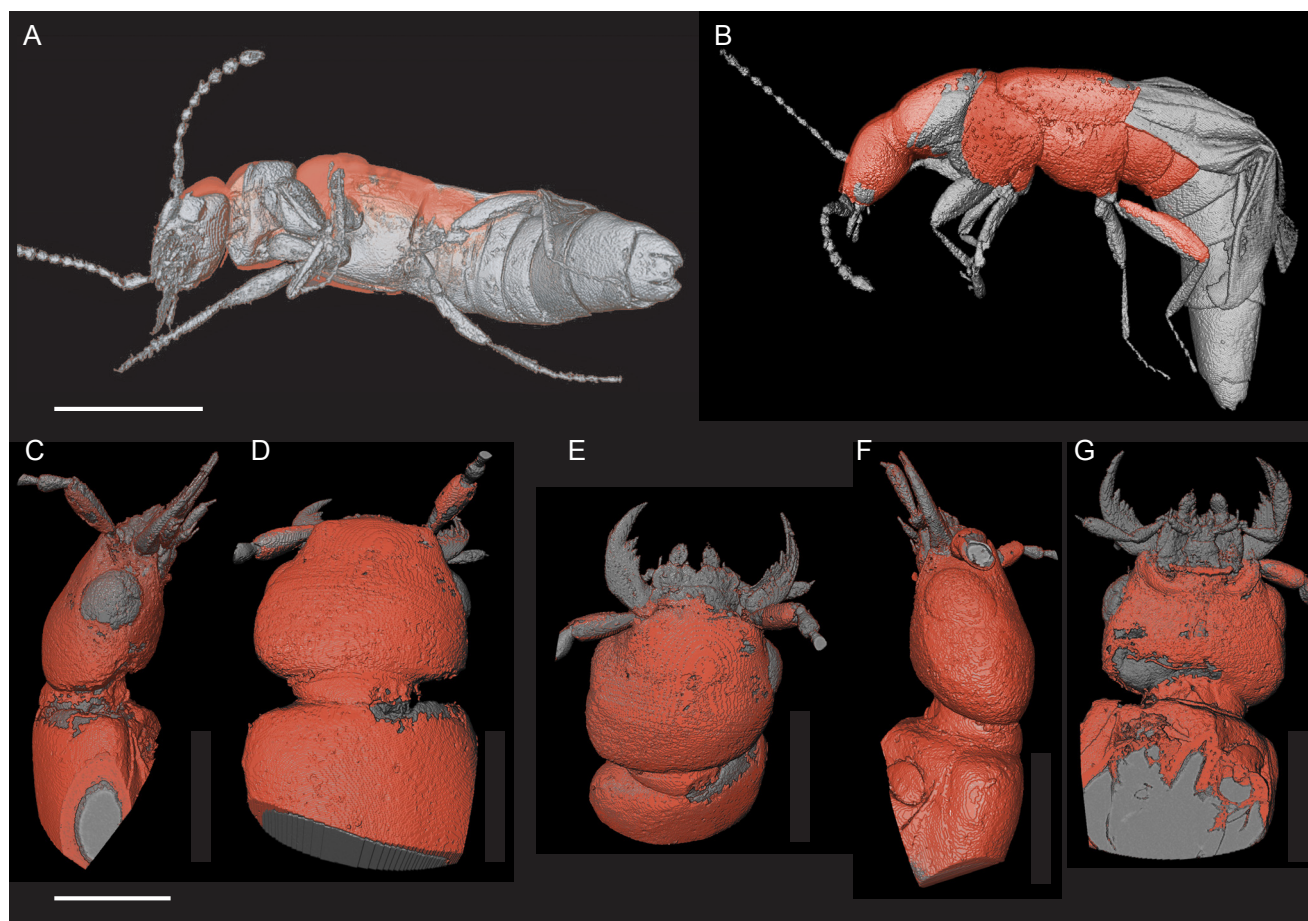
Janina L. Kypke  <http://orcid.org/0000-0003-2120-3133>
Alexey Solodovnikov  <http://orcid.org/0000-0003-2031-849X>

References

- Alekseev VI, Alekseev PI. 2016. New approaches for reconstruction of the ecosystem of an Eocene amber forest. *Biol Bull.* 43:75–86.
- Assing V. 2011a. A revision of *Eusclerus* Sharp, with a redescription of *Sciocharis bupthalma* Scheerpeltz (Coleoptera: Staphylinidae: Paederinae: Medonina). *Linzer biol Beitr.* 43:1169–1177.
- Assing V. 2011b. A revision of the genus *Neosclerus* cameron. *Beiträge zur Entomologie.* 61:89–148.
- Assing V. 2011c. *Orsunius* gen. nov. from the oriental region (Coleoptera: Staphylinidae: Paederinae: Medonina). *Linzer biol Beitr.* 43:221–244.
- Assing V. 2013. A revision of Palaearctic *Medon* IX. New species, new synonymies, a new combination, and additional records (Coleoptera: Staphylinidae: Paederinae). *Entomologische Blätter und Coleoptera.* 109:233–270.
- Assing V. 2014. On *Orsunius* II. Eight new species and additional records (Coleoptera: Staphylinidae: Paederinae: Medonina). *Linzer biol Beitr.* 46:461–479.
- Assing V. 2015. On *Orsunius* III. Four new species from China and Thailand, and additional records (Coleoptera: Staphylinidae: Paederinae: Medonina). *Linzer biol Beitr.* 47:83–96.
- Assing V. 2018. A revision of *Medon*. XI. Five new species, additional records, and the first confirmed records from the Oriental region (Coleoptera: Staphylinidae: Paederinae). *Contrib Entomol.* 68:69–81.
- Berendt GC. 1845. Die organischen Bernstein-Einschlüsse im Allgemeinen. In: Berendt GC, editor. *Die Im Bernstein Befindlichen Organischen Reste Der Vorwelt Gesammelt in Verbindung Mit Mehreren Bearbeitet Und Herausgegeben.* Volume 1. In Commission der Nicolaischen Buchhandlung. Berlin; p. 41–60. doi:10.5962/bhl.title.51864.
- Betz O, Wegst U, Weide D, Heethoff M, Helfen L, Lee W-K, Cloetens P. 2007. Imaging applications of synchrotron X-ray phase-contrast microtomography in biological morphology and biomaterials science. I. General aspects of the technique and its advantages in the analysis of millimetre-sized arthropod structure. *J Microsc.* 227:51–71.
- Bogri A, Solodovnikov A, Żyła D. 2018. Baltic amber impact on historical biogeography and palaeoclimate research: oriental rove beetle *Dysanabatium* found in the Eocene of Europe (Coleoptera, Staphylinidae, Paederinae). *Pap Palaeontol.* 4:433–452.
- Bolton MS, Beaudoin AB. 2017. Climate reconstructions based on post-glacial macrofossil assemblages from four river systems in southwestern Alberta. *Can Water Resour J/Revue Canadienne Des Ressources Hydriques.* 42:289–305.
- Casey TL. 1905. Transactions of the academy of science of St. Louis. *Acad Sci of St. Louis.* 15:17–248.
- Chappard C, Basillais A, Benhamou L, Bonassie A, Brunet-Imbault B, Bonnet N, Peyrin F. 2006. Comparison of synchrotron radiation and conventional x-ray microcomputed tomography for assessing trabecular bone microarchitecture of human femoral heads. *Med Phys.* 33:3568–3577.
- Dierick M, Cnudde V, Masschaele B, Vlassenbroeck J, Van Hoorebeke L, Jacobs P. 2007. Micro-CT of fossils preserved in amber. *Nucl Instrum Meth Phys Res A.* 580:641–643.
- Dlussky GM, Rasnitsyn AP. 2009. Ants (Insecta: Vespida: Formicidae) in the upper Eocene amber of central and Eastern Europe. *Paleontol J.* 43:1024–1042.
- Fleming J. 1821. Supplement to the fourth, fifth and sixth edition of the Encyclopaedia Britannica. With preliminary dissertations on the history of sciences. In: Napier M, editor. *Encyclopaedia Britannica.* Edinburgh & London: Archibald Constable and Company & Hurst, Robinson, and Company; p. 41–56.
- Grimaldi D. 2009. Fossil record. In: Resh VH, Cardé RT, editors. *Encyclopedia of insects.* Cambridge (Massachusetts): Elsevier Inc; p. 396–403. doi: 10.1016/B978-0-12-374144-8.00114-4.
- Grimaldi D, Engel M. 2005. *Evolution of the insects.* New York: Cambridge University Press; p. 755.
- Grimaldi D, Nguyen T, Ketcham R. 2000. Ultra-high-resolution x-ray computed Tomography (UHR CT) and the study of fossils in amber. In: Grimaldi D, editor. *Studies on fossils in amber; with particular reference to the cretaceous of New Jersey.* Leiden (The Netherlands): Backhuys Publishers; p. 77–91.
- Herman GT. 1985. Chapter 3: X-ray-computed tomography - basic principles. In: Robb RA, editor. *Three-dimensional biomedical imaging - volume II.* Boca Raton (Florida): CRC Press; p. 61–106.
- Hörschemeyer T, Beutel RG, Pasop F. 2002. Head structures of *Priacma serrata* LeConte (coleptera, archostemata) inferred from X-ray tomography. *J Morphol.* 252:298–314.
- Hren MT, Sheldon ND, Grimes ST, Collinson ME, Hooker JJ, Bugler M, Lohmann KC. 2013. Terrestrial cooling in Northern Europe during the Eocene-Oligocene transition. *Proc Nat Acad Sci.* 110:7562–7567. doi: 10.1073/pnas.1210930110.
- Jałoszyński P, Brunke A, Metscher B, Zhang -W-W, Bai M. 2017. *Clidicostigus* gen. nov., the first Mesozoic genus of Mastigini (Coleoptera: Staphylinidae: Scydmaeninae) from Cenomanian Burmese amber. *Cretac Res.* 72:110–116.
- Jałoszyński P, Brunke AJ, Yamamoto S, Takahashi Y. 2018. Evolution of Mastigitae: Mesozoic and Cenozoic fossils crucial for reclassification of extant tribes (Coleoptera: Staphylinidae: Scydmaeninae). *Zool J Linn Soc.* 184:623–652.
- Jałoszyński P, Perkovsky E. 2016. Diversity of Scydmaeninae (Coleoptera: staphylinidae) in upper Eocene Rovno amber. *Zootaxa.* 4157:1–85.
- Jennings JT, Austin AD. 2011. Novel use of a micro-computed tomography scanner to trace larvae of wood boring insects. *Aust J Entomol.* 50:160–163.
- John H, Birks B. 2012. Overview of numerical methods in Palaeolimnology. In: Birks HJB, Lotter AF, Juggins S, Smol J, editors. *Tracking environmental change using lake sediments: Data handling and numerical techniques.* London: Springer Science & Business Media; p. 19–92.
- Judson MLI. 2003. Baltic amber fossil of *Garypinus electri* Beier provides first evidence of phoresy in the pseudoscorpion family Garypnidae (Arachnida: chelonethi). In: Logunov DV, Penney D, editors. *Proceedings of the 21st European colloquium of arachnology.* St-Petersburg: Arthropoda Selecta; p. 127–131.
- Kalender WA. 2006. X-ray computed tomography. *Phys Med Biol.* 51: R29–R43.
- Kirejtshuk AG, Azar D, Tafforeau P, Boistel R, Fernandez V. 2009. New beetles of Polyphaga (Coleoptera, Polyphaga) from lower Cretaceous Lebanese amber. *Denisia.* 26:119–130.
- Kuzmina SA. 2017. Macroentomology analysis: methods, opportunities, and examples of reconstructions of paleoclimatic and paleoenvironmental conditions in the Quaternary of the northeastern Siberia. *Contemp Probl Ecol.* 10:336–349.
- Labandeira C. 2014. Amber. *Paleontol Soc PAP.* 20:163–216.
- Laflamme M, Schiffbauer JD, Darroch SAF. 2014. Reading and writing of the fossil record: preservational pathways to exceptional fossilization. *Paleontol Soc PAP.* 20:x–xii.
- Lak M, Néraudeau D, Nel A, Cloetens P, Perrichot V, Tafforeau P. 2008. Phase contrast X-ray synchrotron imaging: opening access to fossil inclusions in opaque amber. *Microsc Microanal.* 14:251–259.
- Latreille PA. 1802. *Histoire Naturelle Générale et Particulière Des Crustacés et Des Insectes: ouvrage Faisant Suite Aux Oeuvres de*

- Leclerc de Buffon, et Partie Du Cours Complet d'histoire Naturelle Rédigé Par C.S. Sonnini, Membre de Plusieurs Sociétés Savantes. Paris: F. Dufart; p. 5–391.
- Li D, Zhang K, Zhu P, Wu Z, Zhou H. 2011. 3D configuration of mandibles and controlling muscles in rove beetles based on micro-CT technique. *Anal Bioanal Chem.* 401:817–825.
- Linnaeus C. 1758. *Systema Naturae per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis.* Holmiae: Impensis Direct. Laurentii Salvii; p.823
- Liu Z, Pagani M, Zinniker D, DeConto R, Huber M, Brinkhuis H, Shah SR, Leckie RM, Pearson A. 2009. Global cooling during the Eocene-Oligocene climate transition. *Science.* 323:1187–1190.
- Mänd K, Muehlenbachs K, McKellar RC, Wolfe AP, Konhauser KO. 2018. Distinct origins for Rovno and Baltic ambers: evidence from carbon and hydrogen stable isotopes. *Palaeogeogr Palaeoclimatol Palaeoecol.* 505:265–273.
- McCoy VE, Soriano C, Gabbott SE. 2018. A review of preservational variation of fossil inclusions in amber of different chemical groups. *Earth Env Sci T R So.* 107:203–211.
- Mierzejewski P. 1978. Electron microscopy study on the milky impurities covering arthropod inclusions in the Baltic amber. *Prace Muzeum Ziemi (Prace Geologiczne).* 28:79–84.
- Nadein KS, Perkovsky EE, Moseyko AG. 2016. New late Eocene Chrysomelidae (Insecta: Coleoptera) from Baltic, Rovno and Danish ambers. *Pap Palaeontol.* 2:117–137.
- Perkovsky EE. 2016. Tropical and Holarctic ants in Late Eocene ambers. *Vestnik Zoologii.* 50:111–122.
- Perkovsky EE, Rasnitsyn AP, Vlaskin AP, Taraschuk MV. 2007. A comparative analysis of the Baltic and Rovno amber arthropod faunas: representative samples. *Afr Inverteb.* 48:229–245.
- Perkovsky EE, Zosimovich VY, Vlaskin AP. 2010. Rovno Amber. In: Penney D, editor. *Biodiversity of fossils in amber from the major world deposits.* Manchester (UK): Siri Scientific Press; p. 116–136.
- Perkovsky EE, Zosimovich VY, Vlaskin AY. 2003a. Rovno amber fauna: a preliminary report. *Acta Zool Cracov.* 46:423–430.
- Perkovsky EE, Zosimovich VY, Vlaskin AY. 2003b. Rovno amber insects: first results of analysis. *Russian Entomol J.* 12:119–126.
- Petrov AV, Perkovsky EE. 2018. A new genus and species of Scolytinae (Coleoptera: Curculionidae) from the Rovno amber. *Paleontol J.* 52:164–167.
- Poinapen D, Konopka JK, Umoh JU, Norley CJD, McNeil JN, Holdsworth DW. 2017. Micro-CT imaging of live insects using carbon dioxide gas-induced hypoxia as anesthetic with minimal impact on certain subsequent life history traits. *BMC Zool.* 2(1). doi:10.1186/s40850-017-0018-x
- Schlüter T, Kühne WG. 1975. Die einseitige Trübung von Harzinkluden - ein Indiz gleicher Bildungsumstände. *Entomologica Germanica.* 2:308–3015.
- Shorthouse DP. 2010. SimpleMappr. SimpleMappr, an online tool to produce publication-quality point maps. World Wide Web Address: [accessed 2018 Oct 02]. <http://www.simplemappr.net/>.
- Simonsen T, Kitching IJ. 2014. Virtual dissections through micro-CT scanning: a method for non-destructive genitalia 'dissections' of valuable Lepidoptera material. *Syst Entomol.* 39:606–618.
- Smith AB. 2009. Parsimony, phylogenetic analysis, and fossils.. In: *Systematics and the fossil record.* Oxford (UK): Blackwell Science Ltd; p. 31–72.
- Smith DB, Bernhardt G, Raine NE, Abel RL, Sykes D, Ahmed F, Pedroso I, Gill RJ. 2016. Exploring miniature insect brains using micro-CT scanning techniques. *Sci Rep.* 6:21768.
- Sokoloff DD, Ignatov MS, Remizova MV, Nuraliev MS, Blagoderov V, Garbout A, Perkovsky EE. 2018. Staminate flower of *Prunus s. l.* (Rosaceae) from Eocene Rovno amber (Ukraine). *J Plant Res.* 1(3). doi:10.1007/s10265-018-1057-2
- Sontag E, Szadziński R. 2011. Biting midges (Diptera: ceratopogonidae) in Eocene baltic amber from the Rovno region (Ukraine). *Polish Journal of Entomology/Polskie Pismo Entomologiczne.* 80:779–800.
- Szwedo J, Sontag E. 2013. The flies (Diptera) say that amber from the gulf of gdańsk, bitterfeld and Rovno is the same Baltic amber. *Polish Journal of Entomology/Polskie Pismo Entomologiczne.* 82:379–388.
- Tafforeau P, Boistel R, Boller E, Bravin A, Brunet M, Chaimanee Y, Cloetens P, Feist M, Hoszowska J, Jaeger JJ, et al. 2006. Applications of X-ray synchrotron microtomography for non-destructive 3D studies of paleontological specimens. *Appl Phys A Mater Sci Process.* 83:195–202.
- Verdú JR, Alba-Tercedor J, Jiménez-Manrique M. 2012. Evidence of different thermoregulatory mechanisms between two sympatric *Scarabaeus* species using infrared thermography and micro-computer tomography. Seebacher F, ed. *PLoS One.* 7:e33914.
- Weide D, Betz O. 2009. Head morphology of selected Staphylinidae (Coleoptera: Staphyliniformia) with an evaluation of possible ground-plan features in Staphylinidae. *J Morphol.* 270:1503–1523.
- Wolfe AP, Tappert R, Muehlenbachs K, Boudreau M, McKellar RC, Basinger JF, Garrett A. 2009. A new proposal concerning the botanical origin of Baltic amber. *Proc Royal Soc B.* 276:3403–3412. doi: 10.1098/rspb.2009.0806.
- Yamamoto S, Maruyama M. 2018. Phylogeny of the rove beetle tribe Gymnusini *sensu n.* (Coleoptera: Staphylinidae: Aleocharinae): implications for the early branching events of the subfamily. *Syst Entomol.* 43:183–199.
- Zanazzi A, Kohn MJ, MacFadden BJ, Terry DO. 2007. Large temperature drop across the Eocene–Oligocene transition in central North America. *Nature.* 445:639–642.
- Zhang K, Li DE, Zhu P, Yuan Q, Huang W, Liu X, Hong Y, Gao G, Ge X, Zhou H, et al. 2010. 3D visualization of the microstructure of *Quedius beesoni* Cameron using micro-CT. *Anal Bioanal Chem.* 397:2143–2148.

Supplementary Material



S1. 3D reconstructions of *Orsunius electronefelus* sp. nov. from μ -CT scans, holotype K7181. Bubbles attached to the body marked in red. A, ventral habitus. B, lateral habitus. C – G, higher resolution scan of head and partial pronotum. C & F, lateral view. D & E, dorsal view. G, ventral view. Scale bars represent A & B, 1 mm; C – G, 0.5 mm.

A short animated movie has been added to the online version of this article. It shows the 3D reconstruction of *O. electronefelus* sp. nov. in rotation and illustrates the major findings (Supplementary Material, S2).