2019-06

Phylogenomics of the superfamily Dytiscoidea (Coleoptera: Adephaga) with an evaluation of phylogenetic conflict and systematic error.

Vasilikopoulos, A

10.1016/j.ympev.2019.02.022
Molecular Phylogenetics and Evolution
Elsevier

Phylogenomics of the superfamily Dytiscoidea (Coleoptera: Adephaga) with an evaluation of phylogenetic conflict and systematic error

Alexandros Vasilikopoulos[a*], Michael Balke[b,c], Rolf G. Beutel[d], Alexander Donath[a], Lars Podsiadlowski[a], James M. Pflug[e], Robert M. Waterhouse[f], Karen Meusemann[a,g,h], Ralph S. Peters[i], Hermes E. Escalona[g], Christoph Mayer[a], Shanlin Liu[j], Lars Hendrich[b], Yves Alarie[k], David T. Bilton[l,m], Fenglong Jia[n], Xin Zhou[o], David R. Maddison[e], Oliver Niehuis[g], Bernhard Misof[a*]

[a]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany

[b]Department of Entomology, SNSB-Bavarian State Collections of Zoology, Münchhausenstr. 21, 81247 Munich, Germany

[c]GeoBio-Center, Ludwig-Maximilians-Universität Munich, Richard-Wagner-Str. 10, 80333 Munich, Germany

[d]Institut für Zoologie und Evolutionsforschung, Friedrich-Schiller-Universität Jena, Ebertstr. 1, 07743 Jena, Germany

[e]Department of Integrative Biology, Oregon State University, 3029 Cordley Hall, Corvallis, 97331 Oregon, USA

[f]Department of Ecology and Evolution, University of Lausanne and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[g]Department of Evolutionary Biology and Ecology, Institute of Biology I (Zoology), Albert-Ludwigs-Universität Freiburg, Hauptstr. 1, 79104 Freiburg, Germany

[h]Australian National Insect Collection, National Research Collections Australia, CSIRO, ACT 2601, Canberra, Australia

[i]Center of Taxonomy and Evolutionary Research, Department of Arthropoda, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany

[j]China National GeneBank, BGI-Shenzhen, Shenzhen, Guangdong Province 518083, P.R. China

[k]Department of Biology, Laurentian University, P3E2C6 Sudbury, Ontario, Canada

[l]Marine Biology and Ecology Research Center, School of Biological and Marine Sciences, University of Plymouth, Drake Circus, PL48AA, UK

[m]Department of Zoology, University of Johannesburg, PO Box 524, Auckland Park, 2006 Johannesburg, South Africa

[n]Institute of Entomology, School of Life Sciences, Sun Yat-sen University, 510275 Guangzhou, P.R. China

[0]Department of Entomology, China Agricultural University, 100193 Beijing, P.R. China

[*]**Corresponding authors at**: Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, Adenauerallee 160, 53113, Bonn, Germany (A.Vasilikopoulos and B. Misof). E-mail addresses: a.vasilikopoulos@leibniz-zfmk.de (A.Vasilikopoulos), b.misof@leibniz-zfmk.de (B. Misof).

Abstract

The beetle superfamily Dytiscoidea, placed within the suborder Adephaga, comprises six families. The phylogenetic relationships of these families, whose species are aquatic, remain highly contentious. In particular the monophyly of the geographically disjunct Aspidytidae (China and South Africa) remains unclear. Here we use a phylogenomic approach to demonstrate that Aspidytidae are indeed monophyletic, as we inferred this phylogenetic relationship from analyzing nucleotide sequence data filtered for compositional heterogeneity and from analyzing amino-acid sequence data. Our analyses suggest that Aspidytidae are the sister group of Amphizoidae, although the support for this relationship is not unequivocal. A sister group relationship of Hygrobiidae to a clade comprising Amphizoidae, Aspidytidae, and Dytiscidae is supported by analyses in which model assumptions are violated the least. In general, we find that both concatenation and the applied coalescent method are sensitive to the effect of among-species compositional heterogeneity. Four-cluster likelihood-mapping suggests that despite the substantial size of the dataset and the use of advanced analytical methods, statistical support is weak for the inferred phylogenetic placement of Hygrobiidae. These results indicate that other kinds of data (e.g. genomic meta-characters) are possibly required to resolve the above-specified persisting phylogenetic uncertainties. Our study illustrates various data-driven confounding effects in phylogenetic reconstructions and highlights the need for careful monitoring of model violations prior to phylogenomic analysis.

Keywords: Hydradephaga, Aspidytidae, transcriptomics, RNA-seq, compositional bias.

1. Introduction

Almost half of the ca. 13,000 beetle species with an aquatic lifestyle (Jäch and Balke, 2008) belong to the suborder Adephaga, which also contains more than 38,000 species of the terrestrial Carabidae and Trachypachidae. The aquatic (or semi-aquatic) adephagan families Amphizoidae, Dytiscidae, Gyrinidae, Haliplidae, Hygrobiidae, and Noteridae have traditionally been considered as monophyletic and collectively referred to as "Hydradephaga" (Crowson, 1960). The monophyly of "Hydradephaga" has not been corroborated in extensive phylogenetic analyses of morphological data or in recent phylogenomic investigations (e.g. Baca et al., 2017; Beutel, 1993; Beutel et al., 2008, 2006; Beutel and Haas, 1996; Beutel and Roughley, 1988; Dressler et al., 2011; Dressler and Beutel, 2010; S. Zhang et al., 2018; but see López-López and Vogler, 2017). On the other hand, the monophyly of the superfamily Dytiscoidea (Amphizoidae, Aspidytidae, Dytiscidae, Hygrobiidae, Meruidae, and Noteridae) is well established (e.g. Baca et al., 2017; Beutel et al., 2013; Dressler et al., 2011; but see López-López and Vogler, 2017). Species of this superfamily can be encountered in virtually every kind of freshwater habitat, including springs, rivers, acidic swamps, lakes, and even in hypersaline or hygropetric habitats. Their widespread occurrence is primarily due to the astounding ecological versatility of species in the family Dytiscidae (Miller and Bergsten, 2016). Interestingly, the phylogenetic relationships within Dytiscoidea are still obscure, especially concerning the hypothesized monophyly of Aspidytidae and the phylogenetic affinities of its species to those of the families Amphizoidae and Hygrobiidae. In the present phylogenomic study, we investigate the above-outlined phylogenetic questions with the largest molecular dataset compiled to date for studying phylogenetic relationships in this group of beetles.

Most species of Dytiscoidea are strictly aquatic, but two families with species inhabiting hygropetric habitats have recently been described. The species of these families occur in geographically disjunct regions. Meruidae, with the single species *Meru phyllisae* Spangler and Steiner, 2005, is known only from the Guiana Shield region of Venezuela (Spangler and Steiner,

2005). Aspidytidae contain two species, *Sinaspidytes wrasei* (Balke, Ribera, Beutel, 2003) from China (Balke et al., 2003; Toussaint et al., 2015) and *Aspidytes niobe* Ribera, Beutel, Balke, Vogler, 2002 from the Cape region of South Africa (Beutel et al., 2010; Ribera et al., 2002a). Phylogenetic analyses have placed these two families in the superfamily Dytiscoidea (Beutel et al., 2006; Ribera et al., 2002a), along with the Dytiscidae (diving beetles, 4,489 species; Nilsson and Hájek, 2019), Noteridae (burrowing water beetles, 258 species; Nilsson, 2011), Hygrobiidae (squeak beetles, six species) and Amphizoidae (trout stream beetles, five species). The taxonomy of Dytiscoidea has been extensively studied, as have been its morphological and ecological adaptations (Balke and Hendrich, 2016; Miller and Bergsten, 2016) and the anatomy of adults and larvae (Belkaceme, 1991; Beutel, 1993, 1988, 1986a, 1986b; Dressler and Beutel, 2010). Moreover, species of the group are well documented in the fossil record and can be traced back to the Triassic (e.g. Beutel et al., 2013; Ponomarenko, 1993).

The phylogenetic relationships of dytiscoid beetles have been addressed in numerous studies investigating morphology, chemical gland compounds, fossil data, and DNA sequences (Alarie et al., 2011, 2004; Alarie and Bilton, 2005; Baca et al., 2017; Balke et al., 2008, 2005; Beutel et al., 2006; Beutel, 1993; Beutel et al., 2013, 2008; Beutel and Haas, 1996; Burmeister, 1976; Dettner, 1985; Kavanaugh, 1986; López-López and Vogler, 2017; McKenna et al., 2015; Ribera et al., 2002b; Toussaint et al., 2015). Analyses of these different data have not yielded congruent topologies (see Fig. 1 for selected hypotheses). The currently accepted view is that Meruidae + Noteridae represent the sister clade of the remaining four families of the superfamily Dytiscoidea (Fig. 1). However, the affinities of Amphizoidae, Aspidytidae, Dytiscidae, and Hygrobiidae remain unresolved. A clade consisting of Dytiscidae and Hygrobiidae is supported by some morphological features (Balke et al., 2005; Beutel et al., 2006; Dressler and Beutel, 2010), such as the presence of prothoracic glands (Beutel, 1988, 1986b; Forsyth, 1970) but molecular and total evidence analyses

have yielded incongruent topologies (e.g. Baca et al., 2017; Balke et al., 2005; Ribera et al., 2002a; Toussaint et al., 2015).

A sister group relationship between Amphizoidae and Aspidytidae has been suggested in previous studies analyzing molecular data (Balke et al., 2008, 2005; Hawlitschek et al., 2012; Toussaint et al., 2015), but Toussaint et al. (2015) recovered paraphyletic Aspidytidae (in relation to Amphizoidae). Specifically, in a multigene analysis of nucleotide sequence data, and after excluding the highly saturated third codon positions, *A. niobe* was placed as a sister taxon of Amphizoidae (Fig. 1f). This new hypothesis contributed to the existing confusion on character evolution within Dytiscoidea (Balke et al., 2005; Beutel et al., 2006; Ribera et al., 2002a), because morphological characters of the adult beetles (antenna: configuration of scape and pedicel) suggest a monophyletic Aspidytidae, while morphological characters of the larvae of *S. wrasei* show considerable structural affinities with those of Amphizoidae (Toussaint et al., 2015).

Given the above outlined uncertainties in the phylogenetic relationships of the families currently included in Dytiscoidea we 1) investigated whether Aspidytidae are monophyletic and 2) inferred the phylogenetic relationships among the families Amphizoidae, Aspidytidae, Dytiscidae, Hygrobiidae, and Noteridae based on an extensive transcriptomic dataset. In order to achieve these goals, we analyzed whole body transcriptomes of species of all major lineages of Dytiscoidea except Meruidae. We also investigated the effects of different potential sources of conflicting phylogenetic signal and phylogenomic incongruence when estimating phylogenetic relationships within Dytiscoidea, and evaluated the degree of confidence for alternative topologies using branch support tests and a data permutation approach.

2. Materials and methods

2.1 Taxon sampling

We compiled a dataset consisting of *de novo*-sequenced transcriptomes and of previously published transcriptomes of Dytiscoidea (Table 1). The sampled species represent all extant families of Dytiscoidea except Meruidae (for which transcriptomic data were not available). As there is high confidence in the hypothesized sister group relationship between Meruidae and Noteridae (Baca et al., 2017; Balke et al., 2008; Beutel et al., 2006; Dressler et al., 2011; Toussaint et al., 2015), we do not deem the lack of the species *M. phyllisae* from our dataset as problematic for investigating the major relationships of Dytiscoidea (see Fig. 1). Representatives of Gyrinidae and Haliplidae were included as outgroups (Baca et al., 2017; Beutel et al., 2013, 2006; Beutel and Haas, 1996; Beutel and Roughley, 1988; Dressler et al., 2011; Dressler and Beutel, 2010).

The *de novo*-sequenced and assembled transcriptomes were screened for putative adaptor, vector and cross-contaminated sequences (see Suppl. Text 1), and clean assemblies were subsequently submitted to the NCBI-TSA database (Table 1). For a detailed description of the procedures for specimen collection and preservation, RNA isolation, RNA library preparation, transcriptome sequencing, transcriptome assembly, cross-contamination screening and sequence submissions see the Supplementary Text 1. We used custom made Perl and Python scripts to calculate descriptive statistics for each transcriptome in our study (Table 1).

2.2 Orthology assignment and alignment refinement

We identified 3,085 clusters of single-copy genes (COGs) that are non-homologous or out-paralogous among each other at the hierarchical level Endopterygota, based on a customized profile query in OrthoDB v.9.1 (Zdobnov et al., 2017) (see Suppl. Text 1). Our query was based on six endopterygote species (subsequently referred to as reference species) with well sequenced and annotated genomes (Suppl. Table 1). Each transcriptome was searched for transcripts orthologous

to the sequences of a given COG (see Peters et al., 2017; Petersen et al., 2017). This search was performed with Orthograph v.0.6.1 (Petersen et al., 2017). Orthologous sequences for each COG (including those of the reference species) were combined in two FASTA files: one containing sequences at the transcriptional level (i.e. nucleotides, nCOGs), the other containing sequences at the translational level (i.e. amino acids, aaCOGs). The resulting nCOGs and aaCOGs are deposited at MENDELEY DATA (XXXXX).

Alignment of the amino-acid sequences in each aaCOG, was performed with MAFFT v.7.309 (Katoh and Standley, 2013) using the algorithm L-INS-i. We screened the amino-acid multiple sequence alignments (MSAs) for potentially misaligned sequences and erroneously identified orthologs using the procedure outlined by Misof et al. (2014). We also adapted the alignment refinement procedure proposed by Misof et al. (2014). Amino-acid and nucleotide sequences that were still identified as outliers after the alignment refinement procedure were removed from the MSAs.

Following the alignment refinement procedure, we removed all sequences of the reference species from the aligned aaCOGs and also discarded their corresponding nucleotide sequences. This resulted in FASTA files that comprised exclusively (aligned) amino-acid or (unaligned) nucleotide sequences of Dytiscoidea and of the outgroup families Gyrinidae and Haliplidae. Next, we discarded all COGs from the ortholog set containing transcripts from fewer than three species. After removing gap-only and ambiguous-only positions from the remaining 2,991 aaCOGs we generated codon-based nucleotide sequence alignments, with a modified version of the script Pal2nal.pl (Suyama et al., 2006) as described by Misof et al. (2014). The 2,991 aligned aaCOGs and the corresponding codon-based alignments are deposited at MENDELEY DATA (XXXXX).

2.3 Concatenation-based and gene-tree-based analysis of amino-acid sequence data

We generated eleven amino-acid supermatrices (Table 2, Suppl. Fig. 1) and assessed the effects of different putative sources of topological incongruence on our concatenation-based phylogenetic inference, namely: 1) alignment masking (i.e. alignment column-filtering) of individual gene partitions when analyzed in a supermatrix context 2) effects of data coverage and phylogenetic information content on the dytiscoid phylogenetic relationships 3) taxonomic decisiveness of gene partitions with respect to a specific phylogenetic question, and 4) effects of compositionally heterogeneous genes in a supermatrix context. We modified the initial supermatrix (supermatrix A, Table 2) by masking the effects of each of the above-mentioned factors one by one (e.g. by removing the randomly similar sections in each gene or removing partitions with low information content). This hierarchical masking strategy progressively resulted in supermatrices to be analyzed with fewer genes and fewer amino-acid alignment sites. We used each generated dataset (Table 2, Suppl. Fig. 1) to infer the phylogeny of Dytiscoidea. The purpose of these analyses was to assess whether or not gradual masking of the initial supermatrix for any of the above factors affected the results of the phylogenetic inference. Amino-acid supermatrices A–K are deposited at MENDELEY DATA (XXXXX).

2.3.1 Masking of the individual amino-acid MSAs

It has been suggested that current methods of alignment masking may lead to biased phylogenetic inferences because alignment columns are filtered too aggressively (Tan et al., 2015). To assess the effect of alignment masking on our results, we first concatenated the original MSAs of aaCOGs without applying alignment masking (supermatrix A). We then applied ALISCORE v.1.2 (Kück et al., 2010; Misof and Misof, 2009) on each aaCOG separately with the options: -r 10^{27} (for the maximum number of pairwise sequence comparisons) and -e. The masked genes (aaCOGs) were then concatenated in a new masked supermatrix (supermatrix B). Concatenation of

both masked and unmasked amino-acid MSAs was conducted with FASconCAT-G v.1.02 (Kück and Longo, 2014).

2.3.2 Increasing data coverage and phylogenetic information content

We evaluated whether or not increasing the saturation (SV, the overall degree of data coverage with respect to gene presence or absence) and the phylogenetic information content (IC) of the supermatrix, as a function of data coverage and phylogenetic signal, had an effect on our tree reconstructions. IC and SV values were calculated with MARE v.0.1.2-rc (MAtrix REduction) (Misof et al., 2013). We generated and assessed the following amino-acid supermatrices:

1) supermatrix C: selected optimal subset (SOS, default output supermatrix) of the software MARE when using supermatrix B as input;

2) supermatrix D: inferred from supermatrix B after removing those genes with IC = 0;

3) supermatrix E: selected optimal subset (SOS, default output supermatrix) of the software MARE when using supermatrix D as input.

We also calculated the SV and the IC of every other amino-acid supermatrix (Table 2). In addition, we calculated the overall alignment completeness scores (C_a) for all supermatrices (Tables 2 and 3) with AliStat v.1.6 (https://github.com/thomaskf/AliStat, see Misof et al., 2014). The overall completeness score provides a direct measure of the overall degree of missing data in each analyzed supermatrix. Moreover, we generated heatmaps of pairwise completeness scores for every amino-acid and nucleotide sequence supermatrix that we analyzed (Suppl. Fig. 3–23).

2.3.3 Controlling for data decisiveness

We constructed two amino-acid sequence supermatrices to control for data decisiveness following the approach outlined by Dell'Ampio et al. (2014). Data decisiveness refers to the property of a partition to include data of every group of species that is relevant to address a specific

phylogenetic question (e.g. the monophyly of Aspidytidae). We generated a subset of supermatrix E by including only those aaCOGs in which all 14 species were present (supermatrix F). An additional decisive dataset (supermatrix G) was constructed by including only those aaCOGs that included at least one representative of Amphizoidae, Dytiscidae, Gyrinidae, Haliplidae, Hygrobiidae, Noteridae, and both representatives of Aspidytidae (*A. niobe* + *S. wrasei*). These two amino-acid sequence datasets were considered decisive for addressing the inter-familiar relationships of Dytiscoidea and the monophyly of Aspidytidae.

2.3.4 Controlling for among-species compositional heterogeneity

Compositional heterogeneity among species in a dataset is often neglected as a source of systematic error in molecular phylogenetic studies (Jermiin et al., 2004; Nesnidal et al., 2010; Philippe and Roure, 2011; Romiguier et al., 2016; Whitfield and Kjer, 2008). We explicitly explored whether among-species compositional heterogeneity biased tree reconstructions. Compositionally heterogeneous aaCOGs were excluded from the decisive amino-acid dataset (supermatrix F) to generate a decisive and more compositionally homogeneous matrix (supermatrix H, Suppl. Fig. 1). Among-species compositional heterogeneity was assessed for each partition separately, based on the partition-specific relative composition frequency variation value (RCFV) (Zhong et al., 2011) calculated by BaCoCa v.1.105 (Kück and Struck, 2014). We followed Fernandez et al. (2016) by considering compositional heterogeneity among species in a given aaCOG to be high when the overall RCFV value was greater than or equal to 0.1. We also filtered supermatrix A and supermatrix E using the same threshold (Table 3, supermatrices J and K) and compared results of tree reconstructions. Complementary to the RCFV approach, we used the software SymTest v.2.0.47 (https://github.com/ottmi/symtest) to calculate the overall deviation from stationarity, reversibility, and homogeneity (SRH) (Jermiin et al., 2008) between the amino-acid (or nucleotide) sequences of the species in each generated supermatrix (see Misof et al., 2014

and Suppl. Text 1). We generated heatmaps to visualize the pairwise deviations from SRH conditions in each generated supermatrix in our study (Suppl. Text 1, Suppl. Fig. 24–44).

2.3.5 Maximum likelihood phylogenetic analyses of amino-acid sequence data

For each of the amino-acid sequence supermatrices (A–K) ten independent partitioned tree searches were performed using IQ-TREE v.1.5.5 (or later) (Nguyen et al., 2015) by specifying the aligned aaCOG boundaries. Model selection for each aaCOG was performed with ModelFinder (Kalyaanamoorthy et al., 2017), implemented in IQ-TREE. We considered the following amino-acid substitution models: DAYHOFF (Dayhoff et al., 1978), DCMUT (Kosiol and Goldman, 2005), JTT (Jones et al., 1992), JTTDCMUT (Kosiol and Goldman, 2005), LG (Le and Gascuel, 2008), LG4X (Le et al., 2012), and WAG (Whelan and Goldman, 2001) allowing all possible combinations of modeling rate heterogeneity among sites (options: -mrate E,I,G,I+G,R -gmedian -merit AICc). We used the edge-linked partitioned model for tree reconstruction (option: -spp) allowing each gene to have its own rate but assuming a common topology and proportional branch lengths among all gene partitions (Chernomor et al., 2016). For each supermatrix the most appropriate model for each gene partition was selected during the first tree search (option -m MFP). The resulting NEXUS files of the first run were used as input for all remaining tree searches.

A common practice in phylogenomic analyses is to optimize the partitioning schemes and corresponding substitution models for the data within an algorithmic framework (Lanfear et al., 2014, 2012). Such optimizations of the partitioning schemes are time-consuming and could result in combining different genes in different meta-partition analyses due to the heuristic optimization procedures implemented in the existing software (Lanfear et al., 2014). This can lead to very different model assignments for different genes and therefore would add an additional uncontrollable effect when comparing different supermatrices. By defining the original masked gene boundaries for all supermatrices and by not optimizing the partitioning schemes we excluded

301 the effects of differential model fit (due to the different composition of the inferred meta-partitions
302 in each matrix) on the results of tree reconstructions. However, in order to avoid missing a unique
303 topology of Dytiscoidea due to suboptimal model fit we optimized the partitioning scheme for a
304 selection of amino-acid supermatrices. We selected the supermatrices H and E for this purpose,
305 because they gave rise to different topologies when analyzing amino-acid sequence data. We used
306 the relaxed clustering algorithm (rcluster) (Lanfear et al., 2014) and RaxML v.8.2 (options: -raxml -
307 rcluster-max 5000) (Stamatakis, 2014) in PartitionFinder v.2.1.1 (Lanfear et al., 2017) to merge
308 partitions according to the default weights under the AICc information criterion. We restricted the
309 model search in PartitionFinder to the following amino-acid substitution models: DAYHOFF+G,
310 DAYHOFF+G+F, DCMUT+G, DCMUT+G+F, JTT+G, JTT+G+F, LG+G, LG+G+F, LG4X,
311 WAG+G, and WAG+G+F. The inferred schemes and models for the corresponding meta-partitions
312 were defined as input for the IQ-TREE tree searches (v.1.5.5) again with the edge-linked model.
313 Ten independent tree searches were performed with the optimized partitioning schemes of
314 supermatrix E and H. The resulting NEXUS files with the optimized schemes of supermatrix E and
315 of supermatrix H are deposited at MENDELEY DATA (XXXXX). Statistical support of our
316 inferred relationships was assessed based on the non-parametric bootstrap measure (Felsenstein,
317 1985) and the bootstrap by transfer (TBE) support measure (Lemoine et al., 2018). We calculated
318 100 non-parametric bootstrap replicates and TBE support using the unoptimized partitioning
319 schemes of all the analyzed amino-acid datasets (Table 2). In addition, we calculated 100 non-
320 parametric bootstrap replicates and TBE support for the optimized partitioning schemes of
321 supermatrices E and H. Subsequently, we mapped the bootstrap support values on the maximum
322 likelihood trees (i.e. trees with the best log-likelihood among all ten tree searches).
323 For the optimized partitioning schemes of the supermatrices E and supermatrix H we also
324 performed one additional tree search with the options -bb 1,000 -alrt 10,000 -abayes to estimate
325 different measures of branch support implemented in IQ-TREE v.1.5.5: Ultrafast Bootstrap 1

(UFBoot1), SH-like aLRT, and aBayes respectively (Anisimova et al., 2011; Guindon et al., 2010; Minh et al., 2013). We also separately calculated branch support based on the updated version of Ultrafast Bootstrap in IQ-TREE v.1.6.8 (UFBoot2, option: -bnni) with 1,000 replicates (Hoang et al., 2017). After verifying topological congruence to the maximum likelihood tree, we mapped the different branch support values on the maximum likelihood tree (Fig. 2).

For a selection of amino-acid supermatrices, we performed one additional tree search using IQ-TREE v.1.5.5 (or later) by implementing the posterior-mean-site-frequency (PMSF) model (Wang et al., 2017), as a rapid approximation of the site-heterogeneous CAT-like mixture model (Quang et al., 2008) with 60 amino-acid profile categories and the exchange rates of the LG substitution matrix (option: -m LG+C60+G+F). We used the tree with the best log-likelihood that resulted from the analysis based on the partition model as a guide tree. The idea of applying this mixture model was to increase the biological realism of the modeled substitution processes, as it should be able to describe site-specific amino-acid preferences in the supermatrices. Moreover, proponents of the site-heterogeneous mixture models have recommended their use to alleviate systematic errors due to model violations (Lartillot et al., 2007) We calculated the non-parametric bootstrap measure (BS PMSF. Fig. 2a, 2b) when applying the PMSF model (LG+C60+G+F) with 100 replicates (Table 2).

2.3.6 Coalescent-based phylogenetic analysis

The supermatrix approach has been criticized for producing statistically inconsistent topologies as it fails to account for gene tree heterogeneity due to incomplete lineage sorting (ILS) (Kubatko and Degnan, 2007). However, research has shown that concatenation (even unpartitioned) can be more accurate than summary species tree methods under certain conditions (Bayzid and Warnow, 2013; Mirarab et al., 2016; Mirarab and Warnow, 2015; Xu and Yang, 2016) and that summary species tree methods can be sensitive to gene tree estimation errors or to low degree of variation in the analyzed sets of loci (Bayzid and Warnow, 2013; Meiklejohn et al., 2016). In an attempt to

explore the sensitivity of our phylogenetic results to the above mentioned potentially biasing factors, we conducted coalescent species tree analyses with ASTRAL III v.5.5.12 (Mirarab and Warnow, 2015; C. Zhang et al., 2018) as an alternative to the supermatrix approach. We expected that if both methods yield the same topologies for the datasets analyzed, any observed topological differences (between analyzed datasets) would unlikely be due to ILS, hybridization or due to biases resulting from gene tree estimation errors.

We performed the coalescent approach on 1) a selected subset of COGs from supermatrix E and 2) the full set of COGs from supermatrix H. When analyzing supermatrix E, we discarded all COGs with fewer than 13 species and more than 20 % ambiguous characters (X, -) to increase data coverage of the selected genes (Sayyari et al., 2017). When analyzing supermatrix H, we selected the full set of COGs to perform the species tree analysis, as this dataset had already a low proportion of missing data (Table 3, Suppl. Fig. 10). Individual gene trees were constructed under the maximum likelihood optimality criterion in IQ-TREE v.1.5.5. Model selection for each aaCOG was restricted to the amino-acid substitution matrices DCMUT, LG, JTT, and WAG under the AICc information criterion. We allowed a maximum of four free rate categories for modeling rate heterogeneity among sites in ModelFinder (option: -cmax 4). We calculated the branch lengths of the estimated species tree in coalescence units in ASTRAL with the option -q. We annotated the species tree with the option -t 2. This resulted in a tree labeled with quartet scores, total quartet support and local posterior probabilities (Sayyari and Mirarab, 2016). Quartet support values (q1, q2, q3) indicate the proportion of induced quartets in the gene trees that agree or disagree with a branch on the calculated species tree. Each alternative value corresponds to the three possible topologies around each branch of interest. The local posterior probabilities are calculated based on the quartet support values (Sayyari and Mirarab, 2016). The first quartet support and local posterior probability for each branch (q1 and pp1 respectively) correspond to the topology that is depicted in the tree that resulted from the coalescent based species tree analysis.

2.4 Maximum likelihood phylogenetic analyses of nucleotide sequence data

We generated the codon-based nucleotide alignment of supermatrix C, by excluding partitions with IC=0 (supermatrix nt.A, Suppl. Fig. 2, Table 3). With this nucleotide supermatrix, we evaluated whether or not 1) there is congruence between amino-acid and nucleotide sequence-based trees, 2) excluding first and third codon positions had a topological effect in the resulting phylogeny of Dytiscoidea, 3) RY-recoding of the nucleotide matrix and subsequent tree reconstruction indicated that heterogeneous base composition is a confounding factor, 4) phylogenetic analyses by including compositionally heterogeneous nCOGs biased tree reconstructions and 5) relative evolutionary rates of COGs affected tree reconstructions. All generated nucleotide sequence supermatrices (Table 3, Suppl. Fig. 2) are deposited at MENDELEY DATA (XXXXX).

Saturation of nucleotide substitutions at third codon positions is a well-known problem when addressing deep phylogenetic relationships (Philippe et al., 2011; Xia et al., 2003) and was also relevant in a recent multigene phylogenetic study of the dytiscoid relationships (Toussaint et al., 2015). Additionally, nucleotide sequences with highly heterogeneous GC content in the third codon positions may contribute to phylogenomic conflict (Romiguier et al., 2016). As a result, the authors of many studies have excluded saturated or compositionally heterogeneous sites prior to their phylogenetic analyses (e.g. Breinholt and Kawahara, 2013; Jarvis et al., 2014; Misof et al., 2014; Pauli et al., 2018; Peters et al., 2017). The second codon positions are arguably the most homogeneous sites among the codon triplets of a supermatrix (e.g. Misof et al., 2014; Timmermans et al., 2016) and should therefore deliver the least biased results. In order to dissect the influence of heterogeneous base composition or saturated substitutions on tree reconstructions, we compared the results of tree reconstructions when 1) including all codon positions of supermatrix nt.A for phylogenetic reconstruction, 2) including only the second codon positions and 3) recoding the nucleotide supermatrix nt.A into RY character states (R: Purines, Y: Pyrimidines). The expectation

is that a recoded matrix should alleviate problems related to compositional heterogeneity and substitution saturation, at the cost of partially eliminating phylogenetic signal.

We further explored the effect of masking (i.e. removing) the most compositionally heterogeneous genes (nCOGs) prior to the tree reconstructions (Table 3). In order to do so, we generated a decisive version of supermatrix nt.A by discarding those nCOGs with fewer than 14 taxa (Suppl. Fig. 2). We did not perform any tree searches for this intermediate decisive dataset. Subsequently, two reduced versions of this decisive supermatrix were generated by excluding genes with RCFV value greater than 0.08 (supermatrix nt.A.homogeneous1, Table 3) and by excluding genes with RCFV value greater than 0.06 (supermatrix nt.A.homogeneous2, Table 3). In addition, because the evolutionary rates of individual genes are often cited as an important predictor of their phylogenetic utility (Doyle et al., 2015; Klopfstein et al., 2017; Yang, 1998), we explored whether the relative evolutionary rates of the included sets of nCOGs biased tree reconstructions (Suppl. Text 1, Table 3). Lastly, we tested whether removal of the species *S. wrasei* from supermatices nt.A and nt.A.homogeneous2 affected the phylogenetic placement of Hygrobiidae (Table 3). We decided to remove *S. wrasei*, because it is the species that was associated with the longest tree branches among the two species of Aspidytidae when analyzing codon-based nucleotide sequence data (Fig. 3).

Ten independent tree searches were performed for each generated nucleotide dataset with IQ-TREE v.1.5.5 (or later). Tree searches and model selection in ModelFinder were based on an edge-linked partition model (options. -spp -gmedian -merit AICc), by considering the nCOG boundaries and the GTR substitution matrix (Tavaré, 1986), and by allowing all possible combinations for modeling among site rate variation. The RY recoded (in the form of binary data [0,1]) matrix was analyzed with an edge-linked partition model in IQ-TREE v.1.6.8 (options: -spp -st BIN -m MFP -gmedian -merit AICc). For a selection of nucleotide supermatrices, we optimized the partitioning scheme in PartitionFinder v.2.1.1 by restricting the model search to GTR and GTR+G with the

options -raxml and -rcluster-max 5000 using the AICc information criterion. For this purpose, we selected the datasets with the lowest levels of among-species compositional heterogeneity (Table 3). The resulting combinations of partitions and models were used as input for IQ-TREE v.1.5.5 for ten additional tree searches with the edge-linked model. Statistical branch support was estimated from 100 non-parametric bootstrap replicates, TBE support, 10,000 SH-like aLRT, aBayes, 1,000 UFBoot1 (IQ-TREE v.1.5.5), and 1,000 UFBoot2 (IQ-TREE v.1.6.8, -bnni) replicates on the datasets with the optimized partitioning schemes and on supermatrix nt.A. After verifying topological congruence to the maximum likelihood tree, we mapped these support values on the tree with the best log-likelihood among the trees that resulted from the ten maximum likelihood searches (Fig. 3, Suppl. Fig. 69). We additionally calculated 100 non-parametric bootstrap replicates and TBE support for every other nucleotide sequence dataset (Table 3). The NEXUS files with the optimized schemes of the supermatrices nt.B and nt.A.homogeneous2, calculated with PartitionFinder, are deposited at MENDELEY DATA (XXXXX).

2.5 Branch support tests with four-cluster likelihood-mapping and data permutations.

We tested the statistical robustness of phylogenomic estimates of four selected phylogenetic hypotheses (Suppl. Tables 2 and 3) by means of the four-cluster likelihood-mapping approach (FcLM) on supermatrix E (Strimmer and von Haeseler, 1997). This approach considers the proportion of taxon quartets in a supermatrix that support each of the three alternative topologies around a specific branch of interest (for details, see also the supplementary material provided by Misof et al., 2014). The formulation of each hypothesis was based on the best tree topology inferred from phylogenetically analyzing supermatrix E (Fig. 2b). We assumed taxa within each group definition to be monophyletic. For each FcLM test (Suppl. Tables 2 and 3) we additionally permuted the original matrix in three ways as described by Misof et al. (2014) to evaluate 1) whether or not the quartet support for a certain hypothesis results from genuine phylogenetic signal,

2) whether or not it is affected by confounding factors relating to compositional heterogeneity, 3) and whether or not the distribution of missing data affected the phylogenetic results (Suppl. Text 1). The FcLM approach and the permutations for testing hypotheses 1 and 3 were also applied on different amino-acid and nucleotide supermatrices (see also Suppl. Text 1 and Sann et al., 2018 for a description of FcLM tests applied at the nucleotide sequence level) with the same taxon group definitions in an attempt to investigate the source of topological incongruence. For each phylogenetic hypothesis tested, we discarded partitions or meta-partitions (if an optimized scheme was calculated for the respective matrix) that were uninformative with respect to a specific taxon-group definition. For the original dataset we used the same models selected during the IQ-TREE tree search for the respective dataset with the option -spp. For the permuted matrices we used the models LG (for amino-acid alignments) and GTR (for the nucleotide alignments) and the option -q for the partition file. All four-cluster likelihood-mapping analyses were conducted using IQ-TREE v.1.5.5.

3. Results

3.1 Orthology assignment and dataset assembly

On average, 2,689 transcripts per species (87 % of 3,085 COGs) passed the reciprocal best hit criterion (Min.= 2,133, Max.= 2,913) during the orthology assignment step. The dataset with the lowest number of assigned orthologs (2,133) was the transcriptome of the diving beetle *Thermonectus intermedius*, while the transcriptome of the species *S. wrasei* was the dataset with the highest number of assigned orthologous transcripts (2,913, Table 4). The average number of outlier sequences per species was 0.4 % (i.e. a mean of 12 outliers per species across 2,991 gene partitions). In total, 167 amino-acid (and corresponding nucleotide) sequences were removed after the alignment refinement step (Suppl. Table 4). The search for ambiguously aligned regions with ALISCORE resulted in the removal of a total number of 276,537 amino-acid sites from the original

amino-acid sequence alignments of supermatrix A (and 829,611 sites from their corresponding codon-based nucleotide sequence alignments).

3.2 Phylogenetic analyses of amino-acid sequence data

The different maximum likelihood searches for the same datasets resulted in congruent topologies (Fig. 2 and Suppl. Fig. 45–59) irrespective of whether or not we optimized the partitioning scheme (for supermatrices E and H respectively). The phylogenetic analyses with the site-heterogeneous mixture models yielded topologies identical to those obtained when using partition models for the amino-acid datasets analyzed (Suppl. Fig. 49, 51, 55, 57). All phylogenetic analyses inferred the monophyly Dytiscoidea as a whole and of each dytiscoid family, and supported a sister group relationship between Noteridae and all remaining families of Dytiscoidea. All the above relationships received high statistical support when analyzing amino-acid sequence data except for the monophyly of Aspidytidae when performing FcLM analysis on supermatrix E (see section 3.4.1). Moreover, a clade comprising the families Amphizoidae and Aspidytidae was suggested in all maximum likelihood analyses of amino-acid sequence data and is fully supported by all branch support measures (Fig. 2a and 2b). FcLM analysis on both the original and the permuted data of supermatrix E indicate high support for a clade consisting of Amphizoidae and Aspidytidae without detectable confounding signal (section 3.4.2, Hypothesis 2, Suppl. Table 2).

The phylogenetic analyses of the amino-acid supermatrices which were not corrected for among-species compositional heterogeneity, suggested Hygrobiidae as the sister clade to Aspidytidae + Amphizoidae with strong statistical branch support. Analyses of these datasets suggested that the three families collectively form a clade sister to the diving beetles (e.g. Fig. 2b). The analysis of supermatrix H (RCFV-corrected version of supermatrix F) yielded a different arrangement with Hygrobiidae being placed as a sister group to (Amphizoidae + Aspidytidae) + Dytiscidae (Fig. 2a). Furthermore, the phylogenetic analysis of the supermatrices J and K (RCFV-

corrected versions of supermatrices E and A respectively) also suggested the latter sister group relationship (Suppl. Fig. 58–59). Non-parametric bootstrap support for the clade (Amphizoidae + Aspidytidae) + Dytiscidae is not very high (supermatrix H: 79 %, Fig 2a, see also Suppl. Fig. 54, 58–59), but most measures such as BS PMSF, UFBoot1, aBayes, SH-aLRT and TBE strongly support this clade.

The coalescent-based species tree analyses with ASTRAL yielded topologies identical to those obtained from concatenation when analyzing supermatrices E and H (Suppl. Fig. 71–72). Overall, the local posterior probabilities in favor of the monophyly of the dytiscoid lineages except Noteridae (i.e. Aspidytidae + Amphizoidae + Dytiscidae + Hygrobiidae), the monophyly of Aspidytidae, and the monophyly of Amphizoidae + Aspidytidae are high in both coalescent phylogenetic analyses. On the one hand, quartet support shows conflict among the selected gene trees of supermatrix E concerning the monophyly of Aspidytidae (q1=0.44; q2=0.32; q3=0.22) and the placement of Hygrobiidae as a sister group to Aspidytidae and Amphizoidae (q1=0.37; q2=0.26; q3=0.36). On the other hand, the local posterior probabilities for the above relationships are high (0.99 and 0.90 respectively). A low quartet support for the monophyly of Aspidytidae is again observed when analyzing the gene trees of supermatrix H (q1=0.45; q2=0.32; q3=0.21), indicating conflict among the gene trees of this dataset for this relationship. A clade comprising Amphizoidae, Aspidytidae, and Dytiscidae (which resulted from the coalescent analysis of the genes in supermatrix H) received low quartet support (q1=0.37; q2=0.36; q3=0.26). This clade also received low support based on the local posterior probability value (0.73).

3.3 Phylogenetic analyses of nucleotide sequence data

In contrast to the analysis of the amino-acid sequence data, phylogenetic analysis of the codon-based nucleotide sequence data (supermatrix nt.A) yielded paraphyletic Aspidytidae, with *S. wrasei* placed as a sister taxon of Amphizoidae (Fig. 3b). However, after removal of the most

compositionally heterogeneous genes, the phylogenetic analyses provided strong statistical branch support for the monophyly of Aspidytidae (Fig. 3a, Suppl. Fig. 65–67). Analyzing exclusively second codon positions also provided strong support for the hypothesis of Aspidytidae representing a natural group (Suppl. Fig. 60 and 69). The best tree from the analysis of the RY-recoded supermatrix supported the monophyly of Aspidytidae as well (Suppl. Fig. 70). Some of the interfamiliar relationships recovered by the analysis of the recoded nucleotide sequence matrix are different than the relationships recovered from most of our analyses. The branch support values for those relationships are high but the internal branches of the tree are very short (Suppl. Fig. 70). As expected, including only the fastest evolving genes in the dataset delivered phylogenetic relationships (including paraphyletic Dytiscoidea) not seen in any of the other phylogenetic analyses. In contrast, removing the ca. 25 % or 75 % of the fastest evolving genes did not result in topological alterations compared with the original results of the analysis of supermatrix nt.A (Suppl. Fig. 61 and 63). Phylogenetic analyses of the concatenated codon-based nucleotide sequence dataset after removing outlier genes with respect to their relative evolutionary rate (Suppl. Fig. 64), yielded the same topology as the analysis of the supermatrix composed of exclusively slowly evolving genes (Suppl. Fig. 61).

Analysis of the nucleotide datasets did not corroborate the hypothesis of Hygrobiidae being the sister group to a clade comprising Aspidytidae, Dytiscidae and Amphizoidae, except when analyzing exclusively second codon positions. One additional difference between the trees derived from analyzing codon-based nucleotide sequence data and the tree based on the analysis of exclusively second codon positions is the placement of Amphizoidae as the sister group of Dytiscidae (Suppl. Fig. 60 and 69). However, this placement is in conflict with the phylogenies inferred when analyzing amino-acid data and which suggested a sister group relationship of Amphizoidae and Aspidytidae (Fig. 2) with high support. The results of the FcLM analysis on the amino-acid supermatrix E (Suppl. Table 3) are also in support of a clade Amphizoidae +

Aspidytidae without detectable confounding signal (see section 3.4.1). Removal of the species *S. wrasei* from the selected codon-based datasets (nt.A and nt.A.homogeneous2) did not affect the phylogenetic placement of Hygrobiidae (Suppl. Fig. 67–68). However, after removal of *S. wrasei* from the compositionally homogeneous matrix the monophyly of (Amphizoidae + Aspidytidae) + Hygrobiidae is only weakly supported (Suppl. Fig. 67).

3.4 Branch support tests with four-cluster likelihood-mapping and data permutations

3.4.1 Monophyly of Aspidytidae

All trees based on the MSAs of amino-acid sequences recovered a monophyletic Aspidytidae. The FcLM analysis of the amino-acid sequence data did not, however, strongly support the monophyly of Aspidytidae (Fig 2c: 55 % of quartets support a monophyletic Aspidytidae when analyzing the original data of supermatrix E). The FcLM results when analyzing supermatrix E show some weaker signal for the placement of *A. niobe* as sister group to Amphizoidae (40 % of quartets). Additionally, after eliminating phylogenetic signal in supermatrix E (permutation scheme I) putative confounding signal emerges supporting the monophyly of Aspidytidae (75 % of quartets). This signal is reduced after having applied permutation scheme II on supermatrix E (40 % of quartets), suggesting that it stems from non-stationary processes among species in supermatrix E (Suppl. Table 2). When the effect of among-species compositional heterogeneity is reduced in the original data (supermatrices H and K), the putative confounding signal supporting the monophyly of Aspidytidae decreases (25 % and 20 % of quartets, permutation scheme I, supermatrix H and K respectively) and the support for the monophyly of Aspidytidae when analyzing the original data increases (60 % of quartets are in favor of the monophyly of Aspidytidae when analyzing the original data of supermatrices H and K).

Maximum likelihood phylogenetic analysis of the supermatrix nt.A strongly supports the sister group relationship between *S. wrasei* and Amphizoidae, as indicated by all applied branch support

measures (Fig. 3b). This arrangement also received relatively high quartet support from the FcLM analysis on the original data of supermatrix nt.A (70 % of quartets, Suppl. Table 3). There is however strong putatively confounding phylogenetic signal in favor of this hypothesis after applying permutation scheme I on supermatrix nt.A (70 % of quartets). This signal is greatly reduced in permutation number II of the same matrix (20 % of quartets), suggesting that it stems from non-stationary processes among species in the supermatrix nt.A. The total number of different quartets that are informative with respect to the monophyly of Aspidytidae is low (20 quartets, Suppl. Table 2) due to the low number of species in our dataset.

3.4.2 Phylogenetic relationships of the dytiscoid families

In all our tree reconstructions, Noteridae were inferred as the sister taxon of all remaining Dytiscoidea (e.g. Fig. 2a, 2b, 3a, 3b). This phylogenetic placement received strong support from most applied statistics, and is also supported by the FcLM and data permutation tests on supermatrix E (100 % of quartets support a clade of Dytiscidae + Hygrobiidae + Amphizoidae + Aspidytidae as the sister group of Noteridae, Suppl. Table 2, Hypothesis 4). In addition, a clade of Aspidytidae + Amphizoidae is fully supported by all analyses based on the amino-acid and nucleotide sequences, except for the analyses of the second codon positions (Suppl. Fig. 60 and 69). We observed a strong signal in favor of Amphizoidae + Aspidytidae when analyzing the original data of supermatrix E (95.3 % of quartets support Amphizoidae + Aspidytidae, Suppl. Table 2), and no detectable confounding signal for this arrangement after applying permutation scheme I on the same amino-acid dataset (39.1 % of quartets support Amphizoidae + Aspidytidae when eliminating phylogenetic signal in supermatrix E).

The position of Hygrobiidae with respect to Amphizoidae, Aspidytidae and Dytiscidae differs between the trees that were inferred at the amino-acid sequence level when allowing for different degrees of compositional heterogeneity among species in the dataset (e.g. Fig. 2). The two

prevailing phylogenetic hypotheses that were inferred from analyzing amino-acid sequence data (Fig. 2a and 2b) received almost equally high support in the FcLM analyses of the different amino-acid and nucleotide data matrices with no detectable confounding factors (Fig. 2d, Suppl. Tables 2 and 3). This result indicates the substantial phylogenetic conflict among the analyzed quartets for this particular phylogenetic question. Again, the total number of quartets for investigating the phylogenetic hypothesis number 3 was not very high (128 quartets) due to taxon sampling limitations in our dataset.

4. Discussion

4.1 The phylogeny of the dytiscoid families and the monophyly of Aspidytidae

Previous analyses based on either morphological or molecular data were unable to deliver congruent reconstructions of dytiscoid phylogenetic relationships (e.g. Baca et al., 2017; Balke et al., 2008, 2005, Beutel et al., 2013, 2008; Toussaint et al., 2015). We addressed these phylogenetic problems with an unprecedented amount of phylogenomic data representing all dytiscoid families except Meruidae. Results of our phylogenomic analyses are consistent with the hypothesis of Noteridae (plus most likely Meruidae) being the sister group of a clade comprising the families Amphizoidae, Aspidytidae, Dytiscidae, and Hygrobiidae (Baca et al., 2017; Beutel et al., 2008; Dressler et al., 2011; McKenna et al., 2015). The monophyly of the latter clade received strong statistical support in all of our analyses. The phylogenetic relationships within this clade, however, are not robustly resolved and resolution depends on the phylogenetic approach and dataset. Nevertheless, our analyses demonstrate that selecting the datasets that violate model assumptions the least support a sister group relationship between Hygrobiidae and a clade comprising Amphizoidae, Aspidytidae, and Dytiscidae. The monophyly of the latter three families is also suggested by an unusual morphological apomorphy, a pair of large and sclerotized epipharyngeal sensilla (Dressler and Beutel, 2010). A clade comprising the squeak beetles and the diving beetles

626 (Hygrobiidae + Dytiscidae), as suggested by some studies based on the analysis of morphological
627 characters (e.g. Alarie and Bilton, 2005; Beutel et al., 2013; Beutel and Roughley, 1988; Dressler et
628 al., 2011) was not recovered in any of our analyses. This suggests that prothoracic glands (Forsyth,
629 1970) have evolved independently in the two families.

630 All analyses of amino-acid sequence data and nucleotide sequence data with reduced levels of
631 among-species compositional heterogeneity suggest monophyletic Aspidytidae. This result is
632 congruent with the analysis of the morphological characters of the adults of Aspidytidae (Balke et
633 al., 2003). Moreover, we received high branch support and high FcLM support for a clade
634 consisting of Amphizoidae and Aspidytidae in all analyses of amino-acid sequence data, and this
635 phylogenetic relationship is also supported by the analysis of codon-based nucleotide sequence
636 data. On the other hand, the analysis of second codon positions suggested a sister group relationship
637 of Amphizoidae and Dytiscidae. The cause of this incongruent result is unclear, but may be due to
638 insufficient or conflicting signal for this relationship in the second codon positions. Overall, we
639 consider a sister group relationship of Amphizoidae and monophyletic Aspidytidae as the most
640 plausible scenario suggested by our data.

641 The disjunct geographical distribution of Amphizoidae, Aspidytidae and Hygrobiidae in
642 combination with the extensive molecular divergence among the three families, and between the
643 two aspidytid species in particular, suggests that these groups represent old and relictual lineages. In
644 this aspect, we corroborate the results put forth by Toussaint et al. (2015) and Hawlitschek et al.
645 (2012), who came to similar conclusions, but these conclusions were based on phylogenetic results
646 from only a few molecular loci. Thus, our results provide a base line for future phylogenomic
647 analyses of dytiscoid relationships and help to identify the most pressing open questions.
648 Additionally, we want to emphasize that the disjunct, relict and micro-endemic distribution of
649 Aspidytidae demands appropriate actions to conserve their habitats and future existence.

650 The instability of the phylogenetic placement of Hygrobiidae among the different datasets
651 analyzed deserves special attention. The lack of resolution in phylogenetics is often attributed to
652 biological phenomena of ancient rapid cladogenesis (Whitfield and Kjer, 2008). Signatures of such
653 processes when analyzing genome-scale data are illustrated by either low levels of phylogenetic
654 signal or highly conflicting phylogenetic signal (Suh, 2016; Whitfield and Kjer, 2008). Our FcLM
655 results as well as the coalescent analyses showed substantial levels of phylogenomic conflict for the
656 interrelationships of the dytiscoid families Amphizoidae, Aspidytidae and Hygrobiidae. The large
657 molecular divergence observed between these families and within Aspidytidae, together with their
658 disjunct geographical distributions and the high levels of gene tree conflict for the interfamiliar
659 relationships observed here, are indications that these lineages may have originated via rapid
660 cladogenesis. On the other hand, such ancient rapid speciation events can be difficult to distinguish
661 from other causes related to data quality and conflict in the analyzed datasets (Whitfield and Kjer,
662 2008) and this hypothesis should be further tested using molecular dating and diversification
663 analyses.
664 The lack of phylogenetic resolution can be the result of deficient taxon sampling (Nabhan and
665 Sarkar, 2012). We acknowledge the sensitivity of phylogenetic reconstructions to taxon sampling,
666 yet we consider our dataset as the most comprehensive genome-scale dataset to date in terms of the
667 number of included species within the small families Amphizoidae, Aspidytidae and Hygrobiidae.
668 Furthermore, we acknowledge that the statistical power of the FcLM approach is highly dependent
669 on the number of sampled species. Increasing the available genomic data, especially within the
670 species-rich Dytiscidae and Noteridae, will inevitably boost the statistical power of the FcLM
671 analyses and further facilitate addressing the persisting phylogenetic uncertainties. Lastly, the
672 analysis of other kind of data such as whole genome sequences, and genomic meta-characters can
673 provide additional or complementary evidence to decipher the evolutionary history of Dytiscoidea
674 (Niehuis et al., 2012).

675

676 **4.2 Model violations bias the reconstruction of the phylogeny of Dytiscoidea**

677 We pointed out that model violations are one very likely source of the observed phylogenetic
678 discrepancies among the different datasets that we analyzed. This is not an unknown phenomenon,
679 as violations of model assumptions, uneven distribution of data coverage, data-type effects, or
680 unnoticed cross-contamination are some of the factors that can strongly bias the results of tree
681 reconstructions (Borowiec et al., 2019; Feuda et al., 2017; Jeffroy et al., 2006; Jermiin et al., 2004;
682 Nesnidal et al., 2013; Philippe et al., 2011; Reddy et al., 2017; Whitfield and Kjer, 2008). In the
683 presented analysis of the dytiscoid relationships we are able to show that masking the genes with
684 the highest levels of among-species compositional heterogeneity altered the topologies of the
685 inferred phylogenetic trees. This was the case irrespective of whether or not we analyzed amino-
686 acid sequence data or nucleotide sequence data. We deduce from this that scientists should seek to
687 take measures against violations of model assumptions in order to more accurately infer the real
688 evolutionary history of the taxa of interest.

689 At the amino-acid sequence level, we reconstructed phylogenetic relationships of Dytiscoidea
690 based on three supermatrices for which the most compositionally heterogeneous genes had been
691 removed (supermatrices H, J, and K). All of these reconstructions yielded congruent topologies,
692 with respect to the interrelationships of the dytiscoid families, which differed from the topologies
693 that resulted from the analyses of the compositionally heterogeneous amino-acid sequence datasets.
694 The effects of among-species compositional heterogeneity at the amino-acid sequence level is
695 further corroborated by our FcLM tests. Although Aspidytidae are recovered as a monophylum
696 when analyzing amino-acid sequence data, there is detectable confounding signal supporting this
697 monophyly in the compositionally heterogeneous supermatrix E. This putatively confounding
698 signal most likely stems from compositional heterogeneity among species in the alignment because
699 it is reduced when analyzing the datasets with reduced levels of among-species compositional

heterogeneity. Furthermore, despite the fact that phylogenetic analysis of both the compositionally homogeneous and the compositionally heterogeneous amino-acid datasets yielded monophyletic Aspidytidae, the compositionally homogeneous supermatrices showed slightly increased phylogenetic signal supporting the monophyly of Aspidytidae. We conclude from these observations that gene partitions with high degrees of among-species compositional heterogeneity biased some of our phylogenetic analyses and are one very likely source of incongruence between tree topologies inferred from analyzing amino-acid sequence data.

Summary coalescent phylogenetic analyses (Mirarab and Warnow, 2015) suggested topologies identical to those obtained when applying a concatenation approach. The observation that both approaches resulted in the same topology irrespective of what dataset we analyzed makes us confident that the incongruence between topologies of different datasets are not due to high levels of incomplete lineage sorting or ancient introgression. This observation further suggests that the applied summary species tree method is sensitive to the same compositional bias as the supermatrix approach.

Our results showed that reducing the degree of missing data and indecisive gene partitions in the amino-acid supermatrices did not affect the topology of the reconstructed dytiscoid phylogeny. The analysis of the amino-acid sequence supermatrix with 100 % data coverage across all species delivered the same topology as the analyses of the non-homogeneous datasets, further supporting the idea that non-random distribution of missing data unlikely accounts for the observed topological differences. Additionally the use of site-heterogeneous amino-acid mixture models in a maximum likelihood framework yielded identical topologies compared with the analysis based on site-homogeneous partition models. The overall information content of the supermatrices (Misof et al., 2013) could not be related to the topological incongruence.

It has been argued that alignment masking might be detrimental to reliable phylogenetic reconstructions (Tan et al., 2015). Tan and colleagues (2015) argue that alignment masking

725 eliminates too much phylogenetic signal and therefore reduces the resolution of single-gene
726 phylogenetic inferences. We found no evidence that alignment masking affected the topology of the
727 dytiscoid phylogeny in the analyses of concatenated and masked aaCOGs.

728 The analysis of the nucleotide sequence data revealed that first and third codon positions are
729 heterogeneous in their base composition, because their inclusion results in a major deviation from
730 SRH conditions. Congruently, the Bowker's pairwise symmetry tests corroborate previous
731 hypotheses that the smallest deviations from SRH conditions are consistently observed in datasets
732 composed solely of second codon positions. Reducing among-species compositional heterogeneity,
733 by recoding the nucleotide sequence data or by removing compositionally heterogeneous genes,
734 restored the monophyly of the cliff water beetles, congruent with tree reconstructions based on the
735 amino-acid sequence datasets. These results indicate that the paraphyly of Aspidytidae as it was
736 found by Toussaint et al. (2015) could also be an artifact resulting from compositional biases in the
737 underlying dataset. Additional evidence for the effect of compositional bias on the analysis of the
738 nucleotide sequence data comes from the results of the FcLM. The FcLM results on supermatrix
739 nt.A suggest that the paraphyletic Aspidytidae stems from non-stationary processes among species
740 in the analyzed dataset, as the signal in favor of this relationship is greatly reduced when applying
741 permutation scheme II. The FcLM results of the nucleotide matrix after reducing among-species
742 compositional heterogeneity shows that there is weak signal supporting the original results (40 %)
743 but there are no detectable confounding effects observed for this arrangement. Taken together these
744 results suggest that the observed paraphyly Aspidytidae obtained when analyzing supermatrix nt.A
745 probably stems from systematic bias owing to among-species compositional heterogeneity in first
746 and third codon positions.

747 We compared the resolution of three distinct sets of genes relative to their evolutionary rate and
748 found that except for the set of genes with the highest relative evolutionary rates, the selection of
749 gene sets did not influence the results. In the extreme case of analyzing a set of the ca. 25 % of the

fastest evolving genes in our supermatrix, we recovered many unexpected relationships, which in turn suggests that including only fast evolving genes results in erroneous phylogenetic estimates of the dytiscoid relationships. Analyses based on the 25 % of the most slowly evolving genes yielded results congruent with those obtained when analyzing all genes (i.e., those of supermatrix nt.A). We also find that after extending the phylogenetic analysis to the 75 % of the slowest evolving genes (i.e. by removing only the 25 % of the fastest evolving genes), the relationships recovered are the same as when analyzing supermatrix nt.A, including the paraphyly of Aspidytidae. Hence, we hypothesize that the paraphyly of Aspidytidae, obtained when analyzing the nucleotide sequence data of supermatrix nt.A, is very likely not driven by the confounding effects of genes with very high evolutionary rates.

5. Conclusions

Our extensive phylogenomic analyses resolve some outstanding issues in adephagan beetle phylogeny, as well as pointing to some problems which apply to phylogenomic approaches more generally. We present evidence that the cliff water beetles (Aspidytidae) constitute a monophylum despite their highly disjunct geographical distribution and large molecular divergence. In addition, our analyses suggest that Aspidytidae are the closest relatives of Amphizoidae. The close affinity of Amphizoidae and Aspidytidae is supported by most of our phylogenetic analyses and by FcLM tests of amino-acid sequence data. Our study could not provide conclusive evidence for some of the interfamiliar relationships of Dytiscoidea, yet we show that excluding genomic regions with high among-species compositional heterogeneity yields different topologies for our transcriptomic dataset. After accounting for most potential tree confounding factors, we consider a sister group relationship between Hygrobiidae and a clade comprising Amphizoidae, Aspidytidae, and Dytiscidae to most likely represent the evolutionary relationships. Overall, we demonstrated in our study how confounding parameters can lead to misleading results. Our study also highlights the

775 importance of interpreting, integrating and summarizing across different datasets and tree-inference
776 approaches for drawing major phylogenetic conclusions. It is obvious that incongruence due to
777 model violations, uneven distribution of missing data, unequal evolutionary rates, as well as
778 conflicting phylogenetic signal among gene trees will prevail in primarily sequence-based
779 phylogenomic analyses, and measures need to be taken against violations of model assumptions. An
780 alternative or complementary route would be the comparative analyses of genomic meta-characters
781 such as the position of introns, the evolution of gene families, or the structure of genes. The
782 tremendous advances in sequencing technologies are currently opening a window into these fields
783 of research (Niehuis et al., 2012).
784
785 **Acknowledgments**
786 This study has partially been enabled by the 1KITE consortium. (www.1kite.org). AV, BM, MB,
787 ON and RGB thank the German Research Foundation (BM649/15-1, NI 1387/7-1, BA2152/24-1)
788 which partially provided funding for this project. MB thanks the Mohammed bin Zayed Species
789 Conservation Fund (MBZ) for a grant to study the Chinese Aspidytidae. RMW is grateful to the
790 Swiss National Science Foundation (PP00P3_170664). Funding for transcriptome sequencing and
791 assembly was in part supported by BGI-Shenzhen. We are grateful to the 1KITE beetle subproject
792 (http://www.1kite.org/subprojects.html) for granting access to a few transcriptomes prior to their
793 planned release. We thank Alexandra Cieslak and Ignacio Ribera for providing transcriptome
794 sequence data for our preliminary analyses. We also thank Bui Quang Minh (Australian National
795 University) and Ondrej Hlinka (CSIRO) for helpful comments and help with the HPC cluster during
796 the analyses of our data.
797
798
799

Authors' contributions

AV, BM, MB, ON, and RGB conceived the study. BM, DRM, MB, ON, RSP, and XZ contributed to coordination of taxon sampling and transcriptome sequencing. BM, DRM, MB, ON, RGB, and XZ, contributed to funding acquisition. DRM, DTB, FJ, HEE, KM, LH, MB, RSP, YA, and XZ collected samples and/or contributed to the data processing of the sequenced transcriptomes. AD, AV, JMP, LP, and SL performed the *de novo* transcriptome assembly and cross-contamination checks. AD, AV, and JMP performed the NCBI sequence submissions. AV, ON, and RMW performed the orthology inference and orthology assignment analyses. AV performed the phylogenetic analyses with contributions, suggestions and comments from BM, KM, and CM. AV, BM, ON, MB and RGB wrote the first draft of the manuscript, with AV taking the lead. All authors contributed with comments and suggestions in the later versions of the manuscript.

Declarations of interest: none

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at (doi link upon acceptance). The filtered and unfiltered COGs as well as all inferred matrices and their partition files are available at the MENDELEY DATA repository (XXXXX).

References

Alarie, Y., Beutel, R.G., Watts, H.S., 2004. Larval morphology of three species of Hygrobiidae (Coleoptera: Adephaga: Dytiscoidea) with phylogenetic considerations. Eur. J. Entomol. 101, 293–311.

Alarie, Y., Bilton, D.T., 2005. Larval morphology of Aspidytidae (Coleoptera: Adephaga) and its phylogenetic implications. Ann. Entomol. Soc. Am. 98, 417–430.

Alarie, Y., Short, A.E.Z., Garcia, M., Joly, L., 2011. Larval morphology of Meruidae (Coleoptera: Adephaga) and its phylogenetic implications. Ann. Entomol. Soc. Am. 104, 25–36. https://doi.org/10.1603/AN10054

828 Anisimova, M., Gil, M., Dufayard, J.F., Dessimoz, C., Gascuel, O., 2011. Survey of branch support
829 methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation
830 schemes. Syst. Biol. 60, 685–699. https://doi.org/10.1093/sysbio/syr041

831 Baca, S.M., Alexander, A., Gustafson, G.T., Short, A.E.Z., 2017. Ultraconserved elements show
832 utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of
833 'Hydradephega'. Syst. Entomol. 42, 1–10. https://doi.org/10.1111/syen.12244

834 Balke, M., Hendrich, L., 2016. 7.6 Dytiscidae Leach, in: Beutel, R.G., Leschen, R.A.B. (Eds.),
835 Handbook of Zoology. Volume IV Arthropoda: Insecta Part 38. Coleoptera, Beetles. Volume
836 1. Morphology and Systematics (Archostemata, Adephaga, Myxophaga, Polyphaga Partim).
837 W. DeGruyter, Berlin, pp. 116–147.

838 Balke, M., Ribera, I., Beutel, R., Viloria, A., Garcia, M., Vogler, A.P., 2008. Systematic placement
839 of the recently discovered beetle family Meruidae (Coleoptera: Dytiscoidea) based on
840 molecular data. Zool. Scr. 37, 647–650. https://doi.org/10.1111/j.1463-6409.2008.00345.x

841 Balke, M., Ribera, I., Beutel, R.G., 2005. The systematic position of Aspidytidae, the
842 diversification of Dytiscoidea (Coleoptera, Adephaga) and the phylogenetic signal of third
843 codon positions. J. Zool. Syst. Evol. Res. 43, 223–242. https://doi.org/10.1111/j.1439-
844 0469.2005.00318.x

845 Balke, M., Ribera, I., Beutel, R.G., 2003. ASPIDYTIDAE: On the discovery of a new beetle
846 family: detailed morphological analysis, description of a second species, and key to fossil and
847 extant adephagan families (Coleoptera), in: Jäch, M.A., Ji, L. (Eds.), Water Beetles of China.
848 Zoologisch-Botanische Gesellschaft & Wiener Coleopterologenverein, Wien, pp. 53–66.

849 Bayzid, M.S., Warnow, T., 2013. Naive binning improves phylogenomic analyses. Bioinformatics
850 29, 2277–2284. https://doi.org/10.1093/bioinformatics/btt394

851 Belkaceme, T., 1991. Skelet und Muskulatur des Kopfes und Thorax von *Noterus laevis* Sturm. Ein
852 Beitrag zur Morphologie und Phylogenie der Noteridae (Coleoptera: Adephaga). Stuttgarter
853 Beiträge zur Naturkunde, Ser. A 462, 1–94.

854 Beutel, R.G., 1993. Phylogenetic analysis of Adephaga (Coleoptera) based on characters of the
855 larval head. Syst. Entomol. 18, 127–147. https://doi.org/10.1111/j.1365-3113.1993.tb00658.x

856 Beutel, R.G., 1988. Studies of the metathorax of the trout-stream beetle, *Amphizoa lecontei*
857 Matthews (Coleoptera : Amphizoidae): Contribution towards clarification of the systematic
858 position of Amphizoidae. Int. J. Insect Morphol. Embryol. 17, 63–81.
859 https://doi.org/10.1016/0020-7322(88)90031-1

860 Beutel, R.G., 1986a. Skelet und Muskulatur des Kopfes der Larve von *Haliplus lineatocollis* Mrsh.
861 (Coleoptera). Stutt. Beitr. Naturkd. 390, 1–15.

862 Beutel, R.G., 1986b. Skelet und Muskulatur des Kopfes und Thorax von *Hygrobia tarda* (Herbst).
863 Ein Beitrag zur Klärung der phylogenetischen Beziehungen der Hydradephaga (Insecta:
864 Coleoptera). Stutt. Beitr. Naturkd. 388, 1–54.

865 Beutel, R.G., Balke, M., Ribera, I., 2010. 3.1. Aspidytidae Ribera, Beutel, Balke and Vogler, 2002.,
866 in: Leschen, R.A.B., Beutel, R.G., Lawrence, J.F. (Eds.), Handbook of Zoology, Arthropoda:
867 Insecta. Coleoptera, Beetles. Vol. 2: Morphology and Systematics (Elateroidea,
868 Bostrichiformia, Cucujiformia Partim). pp. 21–28.

869 Beutel, R.G., Balke, M., Steiner, W.E., 2006. The systematic position of Meruidae (Coleoptera,
870 Adephaga) and the phylogeny of the smaller aquatic adephagan beetle families. Cladistics 22,
871 102–131. https://doi.org/10.1111/j.1096-0031.2006.00092.x

872 Beutel, R.G., Haas, A., 1996. Phylogenetic analysis of larval and adult characters of Adephaga
873 (Coleoptera) using cladistic computer programs. Entomol. Scand. 27, 197–205.
874 https://doi.org/10.1163/187631296X00043

875 Beutel, R.G., Ribera, I., Bininda-Emonds, O.R.P., 2008. A genus-level supertree of Adephaga
876 (Coleoptera). Org. Divers. Evol. 7, 255–269. https://doi.org/10.1016/j.ode.2006.05.003

877 Beutel, R.G., Roughley, R.E., 1988. On the systematic position of the family Gyrinidae
878 (Coleoptera: Adephaga). J. Zool. Syst. Evol. Res. 26, 380–400. https://doi.org/10.1111/j.1439-
879 0469.1988.tb00324.x

880 Beutel, R.G., Wang, B., Tan, J.J., Ge, S.Q., Ren, D., Yang, X.K., 2013. On the phylogeny and
881 evolution of Mesozoic and extant lineages of Adephaga (Coleoptera, Insecta). Cladistics 29,
882 147–165. https://doi.org/10.1111/j.1096-0031.2012.00420.x

883 Borowiec, M.L., Rabeling, C., Brady, S.G., Fisher, B.L., Schultz, T.R., Ward, P.S., 2019.
884 Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants.
885 Mol. Phylogenet. Evol. 134, 111–121.
886 https://doi.org/https://doi.org/10.1016/j.ympev.2019.01.024

887 Boussau, B., Walton, Z., Delgado, J.A., Collantes, F., Beani, L., Stewart, I.J., Cameron, S.A.,
888 Whitfield, J.B., Johnston, J.S., Holland, P.W.H., Bachtrog, D., Kathirithamby, J., Huelsenbeck,
889 J.P., 2014. Strepsiptera, phylogenomics and the long branch attraction problem. PLoS One 9,
890 e107709. https://doi.org/10.1371/journal.pone.0107709

891 Breinholt, J.W., Kawahara, A.Y., 2013. Phylotranscriptomics: Saturated third codon positions
892 radically influence the estimation of trees based on next-gen data. Genome Biol. Evol. 5,
893 2082–2092. https://doi.org/10.1093/gbe/evt157

894 Burmeister, E.G., 1976. Der Ovipositor der Hydradephaga (Coleoptera) und seine phylogenetische
895 Bedeutung unter besonderer Berücksichtigung der Dytiscidae. Zoomorphologie 85, 165–257.

896 Chernomor, O., Von Haeseler, A., Minh, B.Q., 2016. Terrace aware data structure for
897 phylogenomic inference from supermatrices. Syst. Biol. 65, 997–1008.
898 https://doi.org/10.1093/sysbio/syw037

899 Crowson, R.A., 1960. The Phylogeny of Coleoptera. Annu. Rev. Entomol. 5, 111–134.
900 https://doi.org/10.1146/annurev.en.05.010160.000551

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. Atlas Protein Seq. Struct. 5, 345–351.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., Minh, B.Q., von Haeseler, A., Ebersberger, I., Pass, G., Misof, B., 2014. Decisive data sets in phylogenomics: Lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. 31, 239–49. https://doi.org/10.1093/molbev/mst196

Dettner, K., 1985. Ecological and phylogenetic significance of defensive compounds from pygidial glands of Hydradephaga (Coleoptera). Proc. Acad. Nat. Sci. Philadelphia 137, 156–171.

Doyle, V.P., Young, R.E., Naylor, G.J.P., Brown, J.M., 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64, 824–837. https://doi.org/10.1093/sysbio/syv041

Dressler, C., Beutel, R.G., 2010. The morphology and evolution of the adult head of Adephaga (Insecta: Coleoptera). Arthropod Syst. Phylogeny 68, 239–287.

Dressler, C., Ge, S.Q., Beutel, R.G., 2011. Is *Meru* a specialized noterid (Coleoptera, Adephaga)? Syst. Entomol. 36, 705–712. https://doi.org/10.1111/j.1365-3113.2011.00585.x

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39, 783–791.

Fernandez, R., Edgecombe, G.D., Giribet, G., 2016. Exploring phylogenetic relationships within myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. Syst. Biol. 65, 871–889. https://doi.org/10.1093/sysbio/syw041

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., Pisani, D., 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr. Biol. 27, 3864–3870.e4. https://doi.org/10.1016/j.cub.2017.11.008

Forsyth, D.J., 1970. The structure of the defence glands of the Cicindelidae, Amphizoidae, and Hygrobiidae (Insecta: Coleoptera). J. Zool. 160, 51–69. https://doi.org/10.1111/j.1469-7998.1970.tb02897.x

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Hawlitschek, O., Hendrich, L., Balke, M., 2012. Molecular phylogeny of the squeak beetles, a family with disjunct Palearctic-Australian range. Mol. Phylogenet. Evol. 62, 550–554. https://doi.org/10.1016/j.ympev.2011.09.015

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2017. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522. https://doi.org/10.1093/molbev/msx281

936 Jäch, M.A., Balke, M., 2008. Global diversity of water beetles (Coleoptera) in freshwater.
937 Hydrobiologia 595, 419–442. https://doi.org/10.1007/s10750-007-9117-y

938 Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C.,
939 Nabholz, B., Howard, J.T., Suh, A., Weber, C.C., da Fonseca, R.R., Li, J., Zhang, F., Li, H.,
940 Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M.S., Zavidovych, V.,
941 Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch,
942 K., Schierup, M., Lindow, B., Warren, W.C., Ray, D., Green, R.E., Bruford, M.W., Zhan, X.,
943 Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E.P., Bertelsen, M.F., Sheldon, F.H.,
944 Brumfield, R.T., Mello, C. V, Lovell, P. V, Wirthlin, M., Schneider, M.P.C., Prosdocimi, F.,
945 Samaniego, J.A., Velazquez, A.M.V., Alfaro-Núñez, A., Campos, P.F., Petersen, B., Sicheritz-
946 Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D.M., Zhou, Q., Perelman,
947 P., Driskell, A.C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J.,
948 Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F.E., Braun, M.,
949 Fjeldsa, J., Orlando, L., Barker, F.K., Jønsson, K.A., Johnson, W., Koepfli, K.-P., O'Brien, S.,
950 Haussler, D., Ryder, O.A., Rahbek, C., Willerslev, E., Graves, G.R., Glenn, T.C., McCormack,
951 J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V, Stamatakis, A., Mindell, D.P., Cracraft,
952 J., Braun, E.L., Warnow, T., Jun, W., Gilbert, M.T.P., Zhang, G., 2014. Whole-genome
953 analyses resolve early branches in the tree of life of modern birds. Science. 346, 1320-1331.
954 https://doi.org/10.1126/science.1253451

955 Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of
956 incongruence? Trends Genet. 22, 225–231. https://doi.org/10.1016/j.tig.2006.02.003

957 Jermiin, L.S., Ho, S.Y.W., Ababneh, F., Robinson, J., Larkum, A.W.D., 2004. The biasing effect of
958 compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53,
959 638–643. https://doi.org/10.1080/10635150490468648

960 Jermiin, L.S., Jayaswal, V., Ababneh, F., Robinson, J., 2008. Phylogenetic model evaluation. In:
961 Keith, J. (Ed.), Bioinformatics, Data, Sequences Analysis and Evolution, vol. I. Humana Press,
962 Totowa, pp. 331–363. https://doi.org/10.1007/978-1-60327-159-2_16

963 Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices
964 from protein sequences. Comput. Appl. Biosci. 8, 275–282.

965 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017.
966 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 287–
967 289. https://doi.org/10.1038/nmeth.4285

968 Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7:
969 improvements in performance and usability. Mol. Biol. Evol. 30, 772–80.
970 https://doi.org/10.1093/molbev/mst010

971 Kavanaugh, D.H., 1986. A systematic review of Amphizoid beetles (Amphizoidae: Coleoptera) and
972 their phylogenetic relationships to other Adephaga. Proc. Calif. Acad. Sci. 44, 67–109.

973 Klopfstein, S., Massingham, T., Goldman, N., 2017. More on the best evolutionary rate for
974 phylogenetic analysis. Syst. Biol. 66, 769–785. https://doi.org/10.1093/sysbio/syx051

975 Kosiol, C., Goldman, N., 2005. Different versions of the dayhoff rate matrix. Mol. Biol. Evol. 22,
976 193–199. https://doi.org/10.1093/molbev/msi005

977 Kück, P., Longo, G.C., 2014. FASconCAT-G: extensive functions for multiple sequence alignment
978 preparations concerning phylogenetic studies. Front. Zool. 11, 81.
979 https://doi.org/10.1186/s12983-014-0081-x

980 Kück, P., Meusemann, K., Dambach, J., Thormann, B., von Reumont, B.M., Wägele, J.W., Misof,
981 B., 2010. Parametric and non-parametric masking of randomness in sequence alignments can
982 be improved and leads to better resolved trees. Front. Zool. 7, 10.
983 https://doi.org/10.1186/1742-9994-7-10

984 Kück, P., Struck, T.H., 2014. BaCoCa - A heuristic software tool for the parallel assessment of
985 sequence biases in hundreds of gene and taxon partitions. Mol. Phylogenet. Evol. 70, 94–98.
986 https://doi.org/10.1016/j.ympev.2013.09.011

987 Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: Combined selection of
988 partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29,
989 1695–1701. https://doi.org/10.1093/molbev/mss020

990 Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A., 2014. Selecting optimal partitioning
991 schemes for phylogenomic datasets. BMC Evol. Biol. 14, 82. https://doi.org/10.1186/1471-
992 2148-14-82

993 Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2017. Partitionfinder 2: New
994 methods for selecting partitioned models of evolution for molecular and morphological
995 phylogenetic analyses. Mol. Biol. Evol. 34, 772–773. https://doi.org/10.1093/molbev/msw260

996 Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in
997 the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7, 1–14.
998 https://doi.org/10.1186/1471-2148-7-S1-S4

999 Le, S.Q., Dang, C.C., Gascuel, O., 2012. Modeling protein evolution with several amino acid
1000 replacement matrices depending on site rates. Mol. Biol. Evol. 29, 2921–2936.
1001 https://doi.org/10.1093/molbev/mss112

1002 Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol.
1003 25, 1307–1320. https://doi.org/10.1093/molbev/msn067

1004 Lemoine, F., Domelevo Entfellner, J.B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira,
1005 T., Gascuel, O., 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data.
1006 Nature 556, 452–456. https://doi.org/10.1038/s41586-018-0043-0

1007 López-López, A., Vogler, A.P., 2017. The mitogenome phylogeny of Adephaga (Coleoptera). Mol.
1008 Phylogenet. Evol. 114, 166–174. https://doi.org/10.1016/j.ympev.2017.06.009

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., Leschen, R.A.B., Marvaldi, A.E., Mchugh, J. V., Newton, A.F., Robertson, J.A., Thayer, M.K., Whiting, M.F., Lawrence, J.F., Ślipiński, A., Maddison, D.R., Farrell, B.D., 2015. The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. Syst. Entomol. 40, 835–880. https://doi.org/10.1111/syen.12132

Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Kimball, R.T., Braun, E.L., 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: Evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65, 612–627. https://doi.org/10.1093/sysbio/syw014

Miller, K.B., Bergsten, J., 2016. Diving beetles of the world. Systematics and Biology of the Dytiscidae. Johns Hopkins University Press, Baltimore.

Minh, B.Q., Nguyen, M.A.T., Von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195. https://doi.org/10.1093/molbev/mst024

Mirarab, S., Bayzid, M.S., Warnow, T., 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65, 366–380. https://doi.org/10.1093/sysbio/syu063

Mirarab, S., Warnow, T., 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspock, U., Aspock, H., Bartel, D., Blanke, A., Berger, S., Bohm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B.M., Schutte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Xu, X., Yang, H., Wang, J., Kjer, K.M., Zhou, X., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767. https://doi.org/10.1126/science.1257570

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. https://doi.org/10.1186/1471-2105-14-348

1047 Misof, B., Misof, K., 2009. A Monte Carlo approach successfully identifies randomness in multiple
1048 sequence alignments: a more objective means of data exclusion. Syst. Biol. 58, 21–34.
1049 https://doi.org/10.1093/sysbio/syp006

1050 Nabhan, A.R., Sarkar, I.N., 2012. The impact of taxon sampling on phylogenetic inference: A
1051 review of two decades of controversy. Brief. Bioinform. 13, 122–134.
1052 https://doi.org/10.1093/bib/bbr014

1053 Nesnidal, M.P., Helmkampf, M., Bruchhaus, I., Hausdorf, B., 2010. Compositional heterogeneity
1054 and phylogenomic inference of metazoan relationships. Mol. Biol. Evol. 27, 2095–2104.
1055 https://doi.org/10.1093/molbev/msq097

1056 Nesnidal, M.P., Helmkampf, M., Meyer, A., Witek, A., Bruchhaus, I., Ebersberger, I., Hankeln, T.,
1057 Lieb, B., Struck, T.H., Hausdorf, B., 2013. New phylogenomic data support the monophyly of
1058 Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa
1059 are caused by systematic bias. BMC Evol. Biol. 13, 1–13. https://doi.org/10.1186/1471-2148-
1060 13-253

1061 Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective
1062 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32,
1063 268–274. https://doi.org/10.1093/molbev/msu300

1064 Niehuis, O., Hartig, G., Grath, S., Pohl, H., Lehmann, J., Tafer, H., Donath, A., Krauss, V.,
1065 Eisenhardt, C., Hertel, J., Petersen, M., Mayer, C., Meusemann, K., Peters, R.S., Stadler, P.F.,
1066 Beutel, R.G., Bornberg-Bauer, E., McKenna, D.D., Misof, B., 2012. Genomic and
1067 morphological evidence converge to resolve the enigma of Strepsiptera. Curr. Biol. 22, 1309–
1068 1313. https://doi.org/10.1016/j.cub.2012.05.018

1069 Nilsson, A.N., 2011. A world catalogue of the family Noteridae, or the burrowing water beetles
1070 (Coleoptera, Adephaga). Version 16.VIII.2011. Distributed as a PDF file via Internet.
1071 Available from: http://www.waterbeetles.eu (accessed 30 June 2018) [WWW Document].

1072 Nilsson, A.N., Hájek, J., 2019. A world catalogue of the family Dytiscidae, or the diving beetles
1073 (Coleoptera, Adephaga). Version 1.I.2019. Distributed as a PDF file via Internet. Available
1074 from: http://www.waterbeetles.eu (accessed 07 February 2019) [WWW Document].

1075 Pauli, T., Burt, T.O., Meusemann, K., Bayless, K., Donath, A., Podsiadlowski, L., Mayer, C.,
1076 Kozlov, A., Vasilikopoulos, A., Liu, S., Zhou, X., Yeates, D., Misof, B., Peters, R.S.,
1077 Mengual, X., 2018. New data, same story: Phylogenomics does not support Syrphoidea
1078 (Diptera: Syrphidae, Pipunculidae). Syst. Entomol. 43, 447–459.
1079 https://doi.org/10.1111/syen.12283

1080 Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A.,
1081 Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P.A., Heraty, J., Kjer, K.M., Klopfstein, S.,
1082 Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B.,
1083 Niehuis, O., 2017. Evolutionary history of the Hymenoptera. Curr. Biol. 27, 1013–1018.
1084 https://doi.org/10.1016/j.cub.2017.01.027

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

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9. https://doi.org/10.1371/journal.pbio.1000602

Philippe, H., Roure, B., 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. 9, 91. https://doi.org/10.1186/1741-7007-9-91

Ponomarenko, A.G., 1993. Two new species of Mesozoic dytiscoid beetles from Asia. Paleont. J. 27, 182–191.

Quang, L.S., Gascuel, O., Lartillot, N., 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24, 2317–2323. https://doi.org/10.1093/bioinformatics/btn445

Reddy, S., Kimball, R.T., Pandey, A., Hosner, P.A., Braun, M.J., Hackett, S.J., Han, K.-L., Harshman, J., Huddleston, C.J., Kingston, S., Marks, B.D., Miglia, K.J., Moore, W.S., Sheldon, F.H., Witt, C.C., Yuri, T., Braun, E.L., 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66, 857–879. https://doi.org/10.1093/sysbio/syx041

Ribera, I., Beutel, R.G., Balke, M., Vogler, A., 2002a. Discovery of Aspidytidae, a new family of aquatic Coleoptera. Proc. R. Soc. B Biol. Sci. 269, 2351–2356. https://doi.org/10.1098/rspb.2002.2157

Ribera, I., Hogan, J.R., Vogler, A.P., 2002b. Phylogeny of hydradephagan water beetles inferred from 18S rRNA sequences. Mol. Phylogenet. Evol. 23, 43–62. https://doi.org/10.1006/mpev.2001.1080

Romiguier, J., Cameron, S.A., Woodard, S.H., Fischman, B.J., Keller, L., Praz, C.J., 2016. Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. Mol. Biol. Evol. 33, 670–678. https://doi.org/10.1093/molbev/msv258

Sann, M., Niehuis, O., Peters, R.S., Mayer, C., Kozlov, A., Podsiadlowski, L., Bank, S., Meusemann, K., Misof, B., Bleidorn, C., Ohl, M., 2018. Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. BMC Evol. Biol. 18, 1–15. https://doi.org/10.1186/s12862-018-1155-8

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668. https://doi.org/10.1093/molbev/msw079

Sayyari, E., Whitfield, J.B., Mirarab, S., 2017. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. Mol. Biol. Evol. 34, 3279–3291. https://doi.org/10.1093/molbev/msx261

1121 Spangler, P.J., Steiner, W.E., 2005. A new aquatic beetle family, Meruidae, from Venezuela
1122 (Coleoptera: Adephaga). Syst. Entomol. 30, 339–357. https://doi.org/10.1111/j.1365-
1123 3113.2005.00288.x

1124 Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
1125 phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

1126 Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: A simple method to visualize
1127 phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U. S. A. 94, 6815–6819.
1128 https://doi.org/10.1073/pnas.94.13.6815

1129 Suh, A., 2016. The phylogenomic forest of bird trees contains a hard polytomy at the root of
1130 Neoaves. Zool. Scr. 45, 50–62. https://doi.org/10.1111/zsc.12213

1131 Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: Robust conversion of protein sequence
1132 alignments into the corresponding codon alignments. Nucleic Acids Res. 34, 609–612.
1133 https://doi.org/10.1093/nar/gkl315

1134 Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., Dessimoz, C., 2015.
1135 Current methods for automated filtering of multiple sequence alignments frequently worsen
1136 single-gene phylogenetic inference. Syst. Biol. 64, 778–791.
1137 https://doi.org/10.1093/sysbio/syv033

1138 Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences.
1139 Lect. Math. Life Sci. 17, 57–86.

1140 Timmermans, M.J.T.N., Barton, C., Haran, J., Ahrens, D., Culverwell, C.L., Ollikainen, A.,
1141 Dodsworth, S., Foster, P.G., Bocak, L., Vogler, A.P., 2016. Family-level sampling of
1142 mitochondrial genomes in Coleoptera: Compositional heterogeneity and phylogenetics.
1143 Genome Biol. Evol. 8, 161–175. https://doi.org/10.1093/gbe/evv241

1144 Toussaint, E.F.A., Beutel, R.G., Morinière, J., Jia, F., Xu, S., Michat, M.C., Zhou, X., Bilton, D.T.,
1145 Ribera, I., Hájek, J., Balke, M., 2015. Molecular phylogeny of the highly disjunct cliff water
1146 beetles from South Africa and China (Coleoptera: Aspidytidae). Zool. J. Linn. Soc. 176, 537–
1147 546. https://doi.org/10.1111/zoj.12332

1148 Wang, H.C., Minh, B.Q., Susko, E., Roger, A.J., 2017. Modeling site heterogeneity with posterior
1149 mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67,
1150 216–235. https://doi.org/10.1093/sysbio/syx068

1151 Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from
1152 multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–
1153 699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

1154 Whitfield, J.B., Kjer, K.M., 2008. Ancient rapid radiations of insects: challenges for phylogenetic
1155 analysis. Annu. Rev. Entomol. 53, 449–472.
1156 https://doi.org/10.1146/annurev.ento.53.103106.093304

1157 Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its
1158 application. Mol. Phylogenet. Evol. 26, 1–7. https://doi.org/10.1016/S1055-7903(02)00326-3

1159 Xu, B., Yang, Z., 2016. Challenges in species tree estimation under the multispecies coalescent
1160 model. Genetics 204, 1353–1368. https://doi.org/10.1534/genetics.116.190173

1161 Yang, Z., 1998. On the Best Evolutionary Rate for Phylogenetic Analysis. Syst. Biol. 47, 125–133.

1162 Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simao, F.A., Ioannidis, P.,
1163 Seppey, M., Loetscher, A., Kriventseva, E. V., 2017. OrthoDB v9.1: Cataloging evolutionary
1164 and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.
1165 Nucleic Acids Res. 45, D744–D749. https://doi.org/10.1093/nar/gkw1119

1166 Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: Polynomial time species tree
1167 reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 15–30.
1168 https://doi.org/10.1186/s12859-018-2129-y

1169 Zhang, S.Q., Che, L.-H., Li, Y., Dan Liang, Pang, H., Ślipiński, A., Zhang, P., 2018. Evolutionary
1170 history of Coleoptera revealed by extensive sampling of genes and species. Nat. Commun. 9,
1171 205. https://doi.org/10.1038/s41467-017-02644-4

1172 Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K.M., Struck, T.H., 2011.
1173 Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform
1174 annelids. BMC Evol. Biol. 11, 369. https://doi.org/10.1186/1471-2148-11-369

Table 1: An overview of the newly sequenced and previously published transcriptomes that were analyzed in the present study. NCBI accession numbers and descriptive statistics to each transcriptome are provided. Species whose transcriptomes were analyzed are given in alphabetic order.

Table 2: Detailed information and statistics of each generated amino-acid supermatrix analyzed in this study. The overall alignment completeness score of each matrix was calculated with the software AliStat. Matrix phylogenetic information content and saturation were calculated with the software MARE. The RCFV value was calculated with BaCoCa. Pairwise tests of symmetry for the Bowker's test were performed with SymTest. (C_a: overall alignment completeness score, SV: matrix saturation values, IC: matrix phylogenetic information content).

Table 3: Detailed information and statistics of each generated nucleotide supermatrix analyzed in this study. The overall alignment completeness score of each matrix was calculated with AliStat. Pairwise tests of symmetry for the Bowker's test were performed with SymTest. Median p-values 0.00E+00 for the Bowker's test indicate very small numbers. (C_a: Overall alignment completeness score).

Table 4: Summarized statistics of the results of the transcript orthology assignment at the amino-acid sequence level. Species whose transcriptomes were analyzed are given in alphabetic order. The summary statistics were calculated with the helper scripts provided with the Orthograph package.

Species name/Transcriptome	Family	TSA accesssion	BioSample accesion	Bioproject accession	Reference/Source	No. contigs	After local VecScreen	After contam. check	Contigs published	Mean length	Median length	N50 length	Max. length
Amphizoa insolens LeConte, 1853	Amphizoidae	GFUZ01000000	SAMN07501457	PRJNA398088	NCBI-TSA	N/A	N/A	N/A	23,404	1,265	854	1,858	17,558
Amphizoa lecontei Matthews, 1872	Amphizoidae	GFUH01000000	SAMN07289768	PRJNA392306	this study	53,433	53,331	53,298	53,272	869	467	1,540	15,581
Aspidytes niobe Ribera, Beutel, Balke, Vogler, 2002	Aspidytidae	GFUO01000000	SAMN07279561	PRJNA391973	this study	22,688	22,683	22,269	22,272	1,173	716	1,996	9,941
Batrachomatus nannup (Watts, 1978)	Dytiscidae	GFUJ01000000	SAMN07280954	PRJNA392058	this study	43,890	43,601	43,554	43,521	741	446	1,151	15,127
Cybister lateralimarginalis (DeGeer, 1774)	Dytiscidae	GDLH01000000	SAMN03799556	PRJNA286512	1KITE, this study	31,471	31,470	31,403	31,402	981	577	1,586	47,239
Dineutus sp.	Gyrinidae	GDNB01000000	SAMN03799560	PRJNA286516	1KITE, this study	25,920	25,915	24,679	24,661	862	600	1,281	11,252
Gyrinus marinus Gyllenhal, 1808	Gyrinidae	GAUY02000000	SAMN02047132	PRJNA219564	1KITE, Misof et al. (2014)	23,637	23,637	23,510	23,491	866	535	1,426	13,197
Haliplus fluviatilis Aubé, 1836	Haliplidae	GDMW01000000	SAMN03799569	PRJNA286525	1KITE, this study	46,197	46,191	45,977	45,915	847	445	1,504	34,051
Hygrobia hermanni (Fabricius, 1775)	Hygrobiidae	GFUK01000000	SAMN07297121	PRJNA392382	this study	62,884	62,877	62,691	62,715	923	559	1,430	19,834
Hygrobia nigra (Clark, 1862)	Hygrobiidae	GFUN01000000	SAMN07287246	PRJNA392270	this study	28,837	28,835	28,561	28,569	918	567	1,492	10,964
Liopterus haemorrhoidalis (Fabricius, 1787)	Dytiscidae	GFUI01000000	SAMN07280875	PRJNA392045	this study	66,642	66,327	66,281	66,211	604	394	824	8,663
Noterus clavicornis (DeGeer, 1774)	Noteridae	GDNA01000000	SAMN03799605	PRJNA286561	1KITE, this study	21,719	21,716	21,606	21,601	1,046	639	1,695	37,302
Sinaspidytes wrasei (Balke, Ribera, Beutel, 2003)	Aspidytidae	GDNH01000000	SAMN03799537	PRJNA286492	1KITE, this study	41,855	41,748	37,769	37,371	874	400	1,725	25,916
Thermonectus intermedius Crotch, 1873	Dytiscidae	N/A	N/A	N/A	Boussau et al. (2014)	N/A	N/A	N/A	15,833	1,351	867	1,938	38,615

(Table 1)

Amino-acid matrix ID	No. of taxa	No. of amino-acid sites	No. of gene partitions	C_a	SV	IC	Percentage of pairwise p-values < 0.05 for the Bowker's test	Optimization of partitioning scheme	No. tree searches with unoptimized partitioning scheme	No. meta-partitions	No. tree searches with optimized partitioning scheme	No. bootstraps with unoptimized partitioning scheme	No. tree searches with the PMSF model	No. bootstraps with the PMSF CAT-like model	Information
A	14	1,661,023	2,991	0.5976280	0.893	0.521	100.00 %	NO	10	-	-	100	-	-	Unmasked matrix
B	14	1,384,486	2,991	0.6824300	0.891	0.523	100.00 %	NO	10	-	-	100	-	-	Masked genes of matrix A with ALISCORE
C	14	955,158	1,901	0.6668550	0.921	0.650	96.70 %	NO	10	-	-	100	-	-	Default MARE matrix (SOS) of matrix B
D	14	1,366,298	2,948	0.6888650	0.898	0.530	100.00 %	NO	10	-	-	100	1	100	Removed genes with IC=0 from matrix B.
E	14	948,772	1,884	0.6654340	0.921	0.639	95.60 %	YES	10	902	10	100	1	100	Default MARE matrix (SOS) of matrix D.
F	14	468,720	900	0.7548040	1.000	0.673	90.11 %	NO	10	-	-	100	-	-	Decisive 1: selected species with all genes from matrix E
G	14	806,143	1,634	0.7016170	0.951	0.661	93.41 %	NO	10	-	-	100	-	-	Decisive 2: Aspidytidae both present and at least one species for each of the remaining families (filtered matrix E)
H	14	211,275	416	0.8592440	1.000	0.660	73.63 %	YES	10	170	10	100	1	100	Removed genes with RCFV >= 0.1 from matrix F
I	14	218,940	1	1.0000000	N/A	N/A	94.51 %	N/A	10 (unpartitioned)	-	-	100	1	100	Selected sites with 100 % species coverage from matrix D
J	14	391,961	814	0.7751530	0.927	0.639	84.62 %	NO	10	-	-	100	-	-	Removed genes with RCFV >= 0.1 from matrix E
K	14	721,765	1,344	0.6862060	0.868	0.494	95.60 %	NO	10	-	-	100	-	-	Removed genes with RCFV >= 0.1 from matrix A

(Table 2)

1212
1213

Nucleotide dataset	No. of taxa	No. of nucleotide sites	No. of gene partitions	C_a	Percentage of pairwise p-values < 0.05 for the Bowker's test	Median paiwise p-value for the Bowker's test	No. tree searches with the unoptimized partitioning scheme	No. bootstraps with the unoptimized partitioningscheme	Optimization of the partitioning scheme	No. tree searches with the optimized partitioning scheme	No. bootstraps with the optimized partitioning scheme	Information
supermatrix.nt.A	14	4,098,894	2,948	0.6889	98.90 %	0.00E+00	10	100	NO	-	-	Codon-based nucleotide sequence alignment of supermatrix C
supermatrix nt.B	14	1,366,298	2,948	0.6889	97.80 %	3.20E-39	10	100	YES	10	100	Second codon positions of supermatrix nt.A
supermatrix nt.A.recoded	14	4,098,894	2,948	N/A	N/A	N/A	10	100	NO	-	-	RY recoded matrix of supermatrix nt.A
supermatrix nt.A.homogeneous1	14	617,355	498	0.8427	98.90 %	0.00E+00	10	100	NO	-	-	Removed genes with RCFV > 0.08 from the decisive version of supermatrix nt.A
supermatrix nt.A.homogeneous2	14	186,498	170	0.8849	98.90 %	8.40E-75	10	100	YES	10	100	Removed genes with RCFV > 0.06 from a decisive version of supermatrix nt.A
supermatrix nt.A.slow	14	920,700	737	0.6074	98.90 %	0.00E+00	10	100	NO	-	-	Removed genes with a relative rate > Q1 of sorted rates from supermatrix nt.A
supermatrix nt.A.fast	14	1,204,353	749	0.6623	100.00 %	0.00E+00	10	100	NO	-	-	Removed genes with a relative rate < Q3 of sorted rates from supermatrix nt.A
supermatrix nt.A.fast_removed	14	2,913,135	2,212	0.7002	100.00 %	0.00E+00	10	100	NO	-	-	Removed genes with a relative rate > Q3 of sorted rates from supermatrix nt.A
supermatrix nt.A.out_removed	14	3,811,368	2,804	0.7001	98.90 %	0.00E+00	10	100	NO	-	-	Removed genes with outlier values of relative rates from supermatrix nt.A
supermatrix.nt.A.sw	13	4,092,338	2,948	0.6805	98.72 %	0.00E+00	10	100	NO	-	-	Removed species *Sinaspidytes wrasei* from supermatrix nt.A
supermatrix nt.A.homogeneous2.sw	13	186,468	170	0.8810	98.72 %	1.06E-48	10	100	NO	-	-	Removed species *Sinaspidytes wrasei* from supermatrix nt.A.homogeneous2

1214
1215

(Table 3)

Species name/Transcriptome	No. of orthologous hits	Proportion of COGs (%)	Total no. of amino acids	No. of X residues	No. of stop codons	N50 of protein lengths	Mean protein length	Median protein length	Maximum protein length	Minimum protein length
Amphizoa insolens LeConte, 1853	2,820	91.41 %	1,109,394	0	13	491	393	325	3,633	30
Amphizoa lecontei Matthews, 1872	2,765	89.63 %	984,227	0	39	446	355	304	2,409	9
Aspidytes niobe Ribera, Beutel, Balke, Vogler, 2002	2,780	90.11 %	1,077,674	20	26	485	387	328	2,159	20
Batrachomatus nannup (Watts, 1978)	2,561	83.01 %	797,222	0	41	391	311	265	2,142	6
Cybister lateralimarginalis (DeGeer, 1774)	2,680	86.87 %	1,084,064	16	21	508	404	332	6,510	10
Dineutus sp.	2,642	85.64 %	781,715	72	11	362	295	259	2,168	15
Gyrinus marinus Gyllenhal*, 1808*	2,571	83.34 %	830,399	12	16	395	322	291	1,478	13
Haliplus fluviatilis Aubé, 1836	2,891	93.71 %	1,171,464	88	33	502	405	337	2,924	17
Hygrobia hermanni (Fabricius, 1775)	2,903	94.10 %	1,249,213	17	40	541	430	351	3,455	12
Hygrobia nigra (Clark, 1862)	2,662	86.29 %	950,213	13	32	444	356	309	1,977	9
Liopterus haemorrhoidalis (Fabricius, 1787)	2,450	79.42 %	698,178	0	48	351	284	246	2,249	13
Noterus clavicornis (DeGeer, 1774)	2,868	92.97 %	1,128,976	6	38	485	393	329	6,482	6
Sinaspidytes wrasei (Balke, Ribera, Beutel, 2003)	2,913	94.42 %	1,187,784	51	28	515	407	340	3,305	8
Thermonectus intermedius Crotch, 1873	2,133	69.14 %	897,627	0	6	524	420	340	6,828	6

(Table 4)

(Figures of the main text should be colored only in the online version of the article. The figures should be used in double-column format)

Figure 1: Overview of different phylogenetic hypotheses on family phylogenetic relationships among Dytiscoidea proposed in previous studies that had analyzed molecular and morphological data. (Note that Meruidae were not included in all studies. However, since their sister group relationship to Noteridae is generally considered undisputed, we consistently included them in the overview: "Meruidae + Noteridae"). a) Balke et al. (2005) based on morphological data, b) Baca et al. (2017) based on UCE data, c) Beutel et al. (2013, 2006) based on morphological data, d) Ribera et al. (2002a) based on morphological and molecular data, e) Balke et al. (2005, 2008) based on molecular data and Balke et al. (2005) based on morphological and molecular data, f) Toussaint et al. (2015) based on molecular data and McKenna et al. (2015) based on molecular data with only *Aspidytes* included.

Figure 2: Different phylogenetic hypotheses deduced from the analysis of amino-acid sequence data. a) Phylogram with the best log-likelihood score on the optimized scheme of supermatrix and b) phylogram with the best log-likelihood score on the optimized scheme of supermatrix E. Branch support is denoted based on 100 non-parametric bootstrap replicates (BS), 100 non-parametric bootstraps based on the PMSF model (BS PMSF), 10,000 SH-like aLRT replicates (SH-aLRT), aBayes support, 1,000 Ultrafast Bootstraps 1 (UFBoot1), 1,000 Ultrafast Bootstraps 2 (UFBoot2, -bnni), and 100 bootstraps by transfer (TBE). Both trees are rooted with Gyrinidae. Congruent and incongruent clades between the two trees (in terms of included terminal taxa) are illustrated in different colors. c) Results of the FcLM analysis on the original data of supermatrix E for the phylogenetic hypothesis 1 (i.e. monophyly of Aspidytidae). d) Results of the FcLM analysis on the original data of supermatrix E for the phylogenetic hypothesis 3 (i.e. Hygrobiidae are the sister

group of Amphizoidae + Aspidytidae). Beetle photos: 1) *Sinaspidytes wrasei*, 2) *Noterus crassicornis*, 3) *Hygrobia hermanni*, 4) *Amphizoa lecontei*, 5) *Cybister lateralimarginalis* (photos and copyright: M. Balke).

Figure 3: Comparison of phylogenetic hypotheses resulted from the analysis of the codon-based nucleotide sequence data. Congruent and incongruent clades between the two trees (in terms of included terminal taxa) are illustrated in different colors. a) Phylogram with the best log-likelihood score on the optimized scheme of supermatrix nt.A.homogeneous2. b) Phylogram with the best log-likelihood score on the unoptimized partitioning scheme of supermatrix nt.A. Branch support is denoted based on 100 non-parametric bootstrap replicates (BS), 10,000 SH-like aLRT replicates (SH-aLRT), aBayes support, 1,000 Ultrafast Bootstraps 1 (UFBoot1), 1,000 Ultrafast Bootstraps 2 (UFBoot2, -bnni), and 100 bootstraps by transfer (TBE). Both trees were rooted with Gyrinidae.