Data curation and modeling of compositional heterogeneity in insect phylogenomics: a case study of the phylogeny of Dytiscoidea (Coleoptera: Adephaga)

Chenyang Cai[a,b,*], Erik Tihelka[c], Davide Pisani[b], Philip C. J. Donoghue[b,*]

[a] *State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, and Centre for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Nanjing 210008, China*
[b] *School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK*
[c] *Department of Animal Science, Hartpury College, Hartpury, GL19 3BE, UK*

**Corresponding authors:*
E-mail addresses: cycai@nigpas.ac.cn (C.C.), phil.donoghue@bristol.ac.uk (P.C.J.D)

Keywords: Hydradephaga, Hygrobiidae, Transcriptomics, compositional heterogeneity, site-heterogeneous model

Abstract

Diving beetles and their allies are a virtually ubiquitous group of freshwater predators. Knowledge of the phylogeny of the adephagan superfamily Dytiscoidea has significantly improved since the advent of molecular phylogenetics. However, despite recent comprehensive phylogenomic studies, some phylogenetic relationships among the constituent families remain elusive. In particular, the position of the family Hygrobiidae remains uncertain. We address these issues by re-analyzing recently published phylogenomic datasets for Dytiscoidea, using approaches to reduce compositional heterogeneity and adopting site-heterogeneous mixture models. We obtained a consistent, well-resolved, and strongly supported tree, robust to analyses of various sizes of datasets. Consistent with previous studies, the monophyly of the geographically disjunct Aspidytidae is strongly supported. Our analyses support that Aspidytidae are the sister group of Amphizoidae, and more importantly, Hygrobiidae are sister to the diverse Dytiscidae, as convincingly demonstrated by morphology-based phylogenies. Our new results are congruent with recent morphology-based phylogenies. The phylogeny of Dytiscoidea can be resolved by reducing the effect of among-site compositional heterogeneity and adopting a better-fitting model accommodating site-specific amino acid preferences. Our analyses provide a backbone phylogeny of Dytiscoidea, which lays the foundation for better understanding the evolution of morphological characters, life habits, and feeding behaviors of dytiscoid beetles.

1. Introduction

The adephagan superfamily Dytiscoidea (Amphizoidae, Aspidytidae, Dytiscidae, Hygrobiidae, Meruidae, and Noteridae) is a well-established group of beetles (e.g. Baca et al., 2017; Beutel et al., 2013; Dressler et al., 2011; but see López-López and Vogler, 2017). Dytiscoid species occur in various freshwater habitats, including springs, rivers, acidic swamps, lakes, and even in hypersaline and hygropetric habitats. Bell (1966) suggested a clade, Dytiscoidea, comprising aquatic (or semi-aquatic) families such as Noteridae, Amphizoidae, Hygrobiidae, and Dytiscidae. The monophyly of Dytiscoidea has been confirmed in many phylogenetic analyses of morphological characters (Beutel and Haas, 1996; Beutel, 1998; Beutel and Haas, 2000) as well as analyses of molecular data (Ribera et al., 2002a,b; McKenna et al., 2015).

Although the phylogenetic relationships of dytiscoids have been extensively investigated based on morphology, gland chemical compounds, fossils, and molecular data (e.g. Alarie et al., 2011; Alarie and Bilton, 2005; Baca et al., 2017; Balke et al., 2008; Beutel et al., 2006, 2008, 2013; 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), these different datasets do not yield a congruent topology (Vasilikopoulos et al., 2019). Both morphology and molecular based phylogenies have indicated that Meruidae + Noteridae represent the sister clade of the remaining four dytiscoid families (summarized in Vasilikopoulos et al., 2019). The phylogenetic relationships among Amphizoidae, Aspidytidae, Dytiscidae and Hygrobiidae, however, remain unresolved. A recent phylogenomic study based on transcriptomes provided new insights into the backbone phylogeny of Dytiscoidea (Vasilikopoulos et al., 2019): Aspidytidae (cliff water beetles) was recovered as a monophyletic group, which is sister to the relictual family Amphizoidae. However, this phylogenomic study could not present conclusive evidence for some of the interfamilial relationships. After accounting for potential tree confounding factors, it has been considered that Hygrobiidae (squeak beetles) is most likely a sister group to a clade comprising Amphizoidae, Aspidytidae, and Dytiscidae (Vasilikopoulos et al., 2019). Such a relationship between Hygrobiidae and other dytiscoid families has also been supported by previously published Sanger sequence data and a combination of molecular and morphological data (Balke et al., 2005, 2008), but this particular relationship strongly contradicts the conventional hypothesis inferred from comparative morphological studies. For example, a clade consisting of Dytiscidae and Hygrobiidae is strongly supported by some critical morphological features (Beutel et al., 2006; Dressler and Beutel, 2010) such as the presence of prothoracic glands (Beutel, 1986, 1988). Despite extensive sampling of genes and some rare species, the phylogenomic study of Dytiscoidea with an evaluation of phylogenetic conflict and systematic error recently published by Vasilikopoulos et al. (2019) failed to resolve the phylogenetic position of the peculiar family Hygrobiidae. Other recent phylogenomic-scale studies have arrived at yet different results. The largest phylogeny of beetles published to date, based on 4,818 genes (McKenna et al., 2019), and an analysis of Adephaga based on ultraconserved elements (Gustafson et al., 2019) have both recovered Hygrobiidae as a sister to Amphizoidae + Aspidytidae.

One of the key sources of uncertainty and error in inferring phylogenies is compositional and rate heterogeneity (Bleidorn, 2017). Some of the most popular inference methods used in phylogenomics operate under the assumption that the rate of evolutionary change is equal for every position of a sequence alignment (Sheffield et al., 2009). However, this assumption is unrealistic and does not reflect the high compositional and rate heterogeneity observed in metazoan genomes (Lartillot and Philippe, 2008); not only does mutation rate vary among bases (Hodgkinson and Eyre-Walker, 2011), but different parts of the genome are under selection pressures of different intensities (Xing and Lee, 2006), resulting into what typically is a highly unequal evolutionary rate across any given sequence. Models which assume compositional and rate homogeneity can consistently recover incorrect topologies, albeit often with high statistical support (Ho and Jermiin, 2004; Jermiin et al., 2004; Cox et al., 2008; Sheffield et al., 2009). To combat these problems, an arsenal of methods has been developed to reduce site compositional heterogeneity in datasets, such as various data filtering

and data recoding approaches (Bleidorn, 2017). Moreover, some recent complex site-heterogeneous models can account for both compositional and rate heterogeneity across sites. These models, such as CAT-GTR, have been shown to fit real data better than conventional site-homogeneous models and suppress common sources of phylogenetic error such as long branch attraction (Lartillot et al., 2007; Blanquart and Lartillot, 2008; Wang et al., 2008; Foster et al., 2009). In fact, when reanalyzed with these methods, some of the most controversial debates in evolutionary biology in the past decade such as the origin of eukaryotes and metazoans seem to boil down to problems caused by compositional and/or rate heterogeneity (Cox et al., 2008; Feuda et al., 2017; Williams et al., 2020).

To understand the systematic position of Hygrobiidae and the backbone phylogeny of Dytiscoidea, we re-analyzed the recently published phylogenomic data for Dytiscoidea, based on multiple datasets with significantly reduced compositional heterogeneity using site-heterogeneous mixture models (CAT-GTR in PhyloBayes and LG+C20 in IQ-TREE). We also investigated the effects of different approaches of reducing the compositional heterogeneity of large datasets by the data block mapping and gathering using entropy (BMGE) method and Dayhoff recoding.

2. Materials and methods

2.1. Dataset selection

We used the amino acid transcriptome alignments from Vasilikopoulos et al. (2019). The authors produced and analyzed different variants of nucleotide and amino acid alignments of their data. Among the eleven amino-acid supermatrices they generated, their focal analyses were principally based upon the full dataset (Supermatrix A: 14 taxa, 1,661,023 amino-acid sites), and two reduced datasets to increase data coverage and phylogenetic information (Supermatrix E: 14 taxa, 948,772 amino-acid sites), and to reduce the negative effects of among-species compositional heterogeneity (Supermatrix H: 14 taxa, 211,275 amino-acid sites) (Vasilikopoulos et al., 2019). Here we focused on exactly the same three supermatrices download from MENDELEY DATA (http://dx.doi.org/10.17632/j8xwxdtbyb.1) to understand the back bone phylogeny of Dytiscoidea.

To reduce among-site compositional heterogeneity and ease the convergence of runs under site-heterogeneous models (CAT-GTR and LG+C20), we compared the performance of two data transformation methods: data block mapping and gathering using entropy (BMGE) and Dayhoff 6-state recoding.

BMGE identifies phylogenetically informative sites by computing entropy-like scores weighted with BLOSUM similarity matrices in order to distinguish among biologically expected and unexpected variability for each aligned character (Criscuolo and Gribaldo, 2010). BMGE can select characters associated with a score value below a fixed threshold. The entropy score cut-off can be modified with the option '-h'. For example, the '-h 0.3' command used for Supermatrix A'' can select more conserved (or slower-evolving) sites in an amino acid sequence alignment (Criscuolo and Gribaldo, 2010). We prepared four stringently filtered datasets (Supermatrices A', A'', E' and H') by trimming the previously published supermatrices A, E and H using BMGE v.1.1 (Criscuolo and Gribaldo, 2010), which selects phylogenetically informative regions suitable for phylogenetic inference: BMGE -m BLOSUM95 -h 0.4 for supermatrices A', E' and H' and -m BLOSUM95 -h 0.3 for a more conserved supermatrix A''. (Criscuolo and Gribaldo, 2010). BLOSUM95 (Henikoff and Henikoff, 1992) was used as the studied taxa belonging to a single superfamily are represented by closely related amino acid sequences. To test the performance of different BMGE models we also reanalyzed supermatrix A with BLOSUM62 -h0.4 which uses an alignment of proteins with 62% identity.

We furthermore tested the effect of Dayhoff 6-state recoding. This method aims to buffer the effects of saturation and compositional bias by converting the 20 amino acids into 6 groups based on their shared chemical and physical properties (Dayhoff et al., 1978; Hrdý et al., 2004). As such, only changes between categories are considered as substitutions. Dayhoff 6-state recoding was implemented for datasets A', E', and H' in PhyloBayes. We reanalyzed the Dayhoff recoded data

using the CAT-GTR model.

2. Phylogenetic analyses of amino-acid sequence

We employed both site-heterogeneous (CAT-GTR and LG+C20) and site-homogenous (LG4X+R) models to evaluate competing hypotheses on the phylogenetic relationships among the main groups of Dytiscoidea. Two site-heterogeneous models were used: the CAT-GTR model as implemented in PhyloBayes for all trimmed datasets and LG+C20 implemented in IQ-TREE for supermatrix H'. CAT-GTR models compositional heterogeneity among sites incorporating the gamma distribution (Lartillot and Philippe, 2004; Lartillot et al., 2009), while LG+C20 represents a maximum likelihood (ML) variant of the CAT-GTR model (Si Quang et al., 2008). In addition, all trimmed alignments (supermatrices A', A'', E' and H') were used for maximum-likelihood phylogenetic reconstruction under the LG4X+R model (Le et al., 2012) as implemented in IQ-TREE.

For the CAT-GTR analyses, two independent Markov chain Monte Carlo (MCMC) chains were run until convergence (maxdiff < 0.3). For each PhyloBayes run, we used the bpcomp program to generate output of the largest (maxdiff) and mean (meandiff) discrepancy observed across all bipartitions. The ML models LG+C20 and LG4X+R were run using IQ-TREE v.1.6.10 with 1,000 ultra-fast bootstraps (Nguyen et al., 2015). All analyses were performed on the University of Bristol BlueCrystal Phase3 Cluster.

3. Results

Using the BMGE filtering method we obtained four new datasets, which represent subsets of the more conserved amino acid sites of the original supermatrices A, E, and H. The amino acid occupancy of all matrices was significantly improved, especially for larger datasets such as Supermatrices A and E: the data occupancy of Supermatrix A (1,661,023 sites) increased from 59.76% to 92.98% in Supermatrix A' (542,493 sites) and to 95.48% in Supermatrix A'' (399,769 sites), Supermatrix E (948,772 sites) increased from 66.54% to 91.97% in Supermatrix E' (334,457 sites), and Supermatrix H (211,275 sites) increased from 85.92% to 95.22% in Supermatrix H' (156,395 sites) (Fig. 1).

The largest discrepancies (maxdiff) in all PhyloBayes runs equal to 0 (maxdiff < 0.1), indicating they all represent 'good' runs (Lartillot et al., 2013). Like the analyses of amino acid sequence data in Vasilikopoulos et al. (2019), all analyses in the present study supported the monophyly of Dytiscoidea and of each dytiscoid family, and indicated a sister group relationship between Noteridae and the other families of Dytiscoidea, including Amphizoidae, Aspidytidae, Dytiscidae, and Hygrobiidae. All the above relationships received maximal statistical support (Bayesian Posterior Probabilities [BPP]=1) in all analyses (Fig. 2). Our PhyloBayes analysis of the original amino-acid supermatrix H, which were not trimmed using BMGE to reduce the compositional heterogeneity of amino acids, suggested Hygrobiidae as the sister group to Dytiscidae + (Aspidytidae + Amphizoidae) with maximal support (BPP=1), a topology identical to the one based on the same dataset (Supermatrix H) but under a site-homologous model (Fig. 2a in Vasilikopoulos et al., 2019). In addition to this analysis based on the original supermatrix (Supermatrix H), the PhyloBayes analyses based on our new filtered datasets (Supermatrices A', A'', E' and H') all resulted in an identical and fully supported topology: Noteridae + ((Amphizoidae + Aspidytidae) + (Dytiscidae + Hygrobiidae)) (Fig. 2). Trimming supermatrix A with BLOSUM62 -h 0.4 and subsequently analyzing this dataset with the CAT-GTR model yielded the same topology as the CAT-GTR analysis of BLOSUM95 data in Fig. S1. Analyzing the trimmed dataset with the simplistic ML model LG4X+R yielded the same topology as the LG+C20, again with a poorly resolved position of Hygrobiidae (Fig. S2).

In all tree reconstructions based on filtered datasets under a site-heterogeneous model, Noteridae was supported as the sister group to all remaining Dytiscoidea. Both clades of Aspidytidae + Amphizoidae and Dytiscidae + Hygrobiidae were strongly supported by all analyses based on the amino-acid datasets. We observed a confounding signal in the original amino-acid dataset

(Supermatrix H), which is probably negatively affected by the compositional heterogeneity. The position of Hygrobiidae within Dytiscoidea (as a sister group to Dytiscidae) was stable and consistent in all analyses of filtered amino acid datasets.

The analysis of supermatrix H' using the site-heterogeneous LG+C20 recovered Dytiscidae as a sister group to a clade comprising Amphizoidae, Aspidytidae, and Hygrobiidae, albeit this clade received low support. Aside from the position of Dytiscidae and Hygrobiidae, the latter of which was not supported (Maximum Likelihood Bootstrap [MLB] = 52, Fig. S3), other relationships were identical to those recovered by the CAT-GTR analysis.

Our maximum likelihood (IQ-TREE) LG4X+R analyses of the amino-acid supermatrices E' and H' resulted in identical topologies (Fig. 3) to those based on the original supermatrices E and H under optimized schemes, respectively (Fig. 2a,b in Vasilikopoulos et al., 2019). Moreover, the support values are interestingly correlated to those yielded in the original analyses. For instance, for the supermatrices A', A'' and H', the nodes uniting Amphizoidae + Aspidytidae and Dytiscidae were weakly supported (MLB = 73 for supermatrix H'). Similarly, within the family Dytiscidae the node between *Liopterus haemorrhoidalis* and *Cybister lateralimarginalis* + *Thermonectus intermedius* was moderately supported (MLB = 90 for supermatrix H'). Unlike the 10-partitioned ML tree of the original supermatrix A (Supplementary Fig. 45 in Vasilikopoulos et al., 2019), our maximum likelihood analyses of the filtered supermatrices A' and A'' both yielded a topology identical to the one under supermatrix H' or supermatrix H, in which Hygrobiidae is the sister group to the weakly supported (MLB = 54 in supermatrix A' and 58 in supermatrix A'') clade (Aspidytidae + Amphizoidae) + Dytiscidae (Fig. 3). Based on the maximum likelihood analyses of supermatrices A' and A'', we found that a more conserved dataset with slower-evolving sites can produce an identical but better supported topology under the same model (Fig. 3).

Dayhoff recoding of datasets A', E', and H' that were subsequently analyzed with CAT-GTR recovered Hygrobiidae as a sister group to a clade comprising Amphizoidae, Aspidytidae, and Dytiscidae (Fig. S4–S6).

4. Discussion

Despite extensive analyses of both morphological and molecular data, it has proven challenging to achieve a congruent reconstruction of dytiscoid phylogeny (e.g. Baca et al., 2017; Balke et al., 2005, 2008 Beutel et al., 2008, 2013; Toussaint et al., 2015; Vasilikopoulos et al., 2019). To tackle this phylogenetic problem, we used a large published phylogenomic dataset representing all dytiscoid families except Meruidae. Unlike the inconsistent and equivocal results under various datasets in Vasilikopoulos et al. (2019), our analyses based on a complex and better-fitting model and multiple datasets with reduced compositional heterogeneity yielded a consistent and fully supported tree of Dytiscoidea. We suggest that Noteridae (plus most likely Meruidae, Vasilikopoulos et al., 2019) is the basal-most lineage within Dytiscoidea, sister to a clade comprising Amphizoidae, Aspidytidae, Dytiscidae, and Hygrobiidae (McKenna et al., 2015; Vasilikopoulos et al., 2019). As confirmed in the recent phylogenomic study of Vasilikopoulos et al. (2019) and other morphological and/or molecular phylogenies (e.g. Balke et al., 2005, 2008), Aspidytidae is monophyletic and sister to Amphizoidae with strong support in all Bayesian analyses of the amino-acid sequence data.

The phylogenetic position of Hygrobiidae is well resolved by our re-analyses, unlike the results in Vasilikopoulos et al. (2019), in which the phylogenetic position is affected by a highly conflicting phylogenetic signal. A clade encompassing Hygrobiidae and Dytiscidae, as suggested by some studies based on the analysis of morphological characters (e.g. Beutel et al., 2013; Beutel and Roughley, 1988; Dressler et al., 2011), is strongly supported in all analyses of filtered datasets. Despite several obvious anatomical differences between Hygrobiidae and Dytiscidae (Alarie et al., 2004; Dettner, 2016), many studies including an analysis of molecular data (Shulsl et al., 2001) suggest that these families are sister groups. A close relationship between Hygrobiidae and Dytiscidae is also supported by a combined phylogenetic analysis (Ribera et al., 2002a), larval

morphology (Alarie and Bilton, 2005), and traces of antimicrobic pygidial gland compounds such as benzoic acid and *p*-hydroxybenzaldehyde (Dettner, 1987). More importantly, they share a similar prothoracic defensive gland (Forsyth, 1970), which is another potential synapomorphy of the two families (Dettner, 2016).

Previous simulation studies showed that site trimming using BMGE produces datasets leading to accurate trees, and this method has been widely applied to inferring deep phylogenies (e.g. Zaremba-Niedzwiedzka et al., 2017; Martijn et al., 2018; Lahr et al., 2019; Philippe et al., 2019; Strassert et al., 2019). Our filtered datasets, with a significantly improved signal/noise ratio, are suitable for phylogenetic analyses, and the phylogenetic trees are less affected by phylogeny reconstruction artefacts due to compositional heterogeneity (e.g. Feuda et al., 2017; Lozano-Fernandez et al., 2019a). Regardless of the BLOSUM method used for trimming, the topologies were identical further demonstrating the robustness of our analyses. Unlike the tree reconstructing methods used in Vasilikopoulos et al. (2019), we employed the more complex site-heterogeneous CAT-GTR model implemented in PhyloBayes, which can account for potential site-specific amino acid preferences (or compositional heterogeneity) (e.g. Lozano-Fernandez et al., 2019a; Schwentner et al., 2017; Wolfe et al., 2019). The CAT-GTR model is mostly regarded to be best suited to suppress artefacts in phylogenetic estimation such as long-branch attraction, especially for large-scale analyses (Feuda et al., 2017; Lartillot et al., 2007; Lozano-Fernandez et al., 2019b). In addition, based on the comparative analyses of both amino acid and nucleotide sequence data by Vasilikopoulos et al. (2019), amino acids should be preferred to nucleotides in phylogenomic analyses of ancient relationships (e.g. Inagaki and Roger, 2006; Rota-Stabelli et al., 2013; Schwentner et al., 2017).

When all datasets (even filtered using BMGE) are analyzed using maximum likelihood (ML) under the less fitting LG4X+R model, a tree is supported where Amphizoidae is the sister group to Aspidytidae, but the systematic position of Hygrobiidae is, as observed in the previous study (Vasilikopoulos et al., 2019), not stable. It is noteworthy that, in all ML trees of the filtered amino acid datasets the support values of the nodes between Hygrobiidae and other dytiscoid families are always not well supported (LG+C20: MLB = 52 in Supermatrix H'; LG4X+R : MLB = 54 in Supermatrix A' and 58 in Supermatrix A'', MLB = 82 in Supermatrix E', and MLB = 73 in Supermatrix H'). Similar weakly supported results, also obtained in Vasilikopoulos et al. (2019) under the simplistic site-homogeneous model, are probably artefactual. As indicated in Vasilikopoulos et al. (2019), the systematic position of Hygrobiidae cannot be resolved unambiguously under the ML analyses with the model they adopted. This difficulty is probably, in part, due to a lack of sufficient phylogenetic signal for the Hygrobiidae and Dytiscidae clade, since the internode between these two families is very short under the CAT-GTR model, perhaps reflecting early rapid diversification of these beetles. Such a problem is also found in other phylogenomic studies of other pancrustacean animals (e.g. Schwentner et al., 2017; Lozano-Fernandez et al., 2019b), where the sister group of Hexapoda, Remipedia, can only be recovered under a site-heterogeneous model (CAT-GTR) but not a homogeneous model. Recent studies that have recovered Hygrobiidae as a sister to a clade containing Amphizoidae and Aspidytidae (Gustafson et al., 2019; McKenna et al., 2019) have likewise both relayed on time-saving site-homogeneous models or their ML extensions which do not account for compositional heterogeneity and can lead to the recovery of misleading topologies, as demonstrated in our analyses.

Dayhoff recoding led to the recovery of Hygrobiidae as a sister group to a clade comprising Amphizoidae, Aspidytidae, and Dytiscidae. While the relationship received full support when the recoded datasets were analyzed with CAT-GTR (BPP = 1), we view this relationship as highly unlikely. It was suggested by Vasilikopoulos et al. (2019) with uncertainty over the placement of Hygrobiidae but was never recovered by any other formal phylogenetic analysis specifically addressing the phylogeny of Dytiscoidea (Ribera et al., 2002a; Balke et al., 2005, 2008; Beutel et al., 2006, 2013, 2019; Toussaint et al., 2015; Baca et al., 2017; López-López and Vogler, 2017;

Gustafson et al., 2019) and is incongruent with morphological evidence discussed below. While in theory Dayhoff-6 recoding should alleviate the effects of compositional heterogeneity, recoding also reduces genuine phylogenetic signal. Trees inferred from Dayhoff-6 recoded data often have low support values and oft-times recover surprising relationships (e.g. Rota-Stabelli et al., 2012; Lozano-Fernandez et al., 2019b). Indeed, the loss of phylogenetic signal in Dayhoff recoding may in some cases outweigh the benefits of suppressed compositional heterogeneity (Hernandez and Ryan, 2019), and so the decision whether to use 6-state recoding has to be made with this caveat in mind.

Overall, our results are consistent with morphology-based views of dytiscoid relationships. The sister-group relationship between Hygrobiidae and Dytiscidae was proposed by Burmeister (1976) based on morphology of the ovipositor and by Ruhnau (1986) based on larval morphology. Both adult Dytiscidae and Hygrobiidae also share the presence of prothoracic glands, among other characters (Forsyth, 1970; Beutel, 1986; Beutel, 1988). A clade comprising the two families was recently recovered by a maximum parsimony analysis of morphological data (Beutel et al., 2019). This same analysis also recovered Aspidytidae as a sister to Amphizoidae, in congruence with our CAT-GTR trees. It should be noted however that some deeper nodes in Beutel et al. (2019) did not receive high bootstrap support values, which is a common problem in morphological phylogenies (Fig. S7). With the relationships among Dytiscoidea strongly supported in our analyses (Fig. 2), our results confirm Beutel and colleague's morphology-based phylogeny of Dytiscoidea.

5. Concluding remarks

The phylogenetic relationships presented here provide an updated hypothesis about the evolution of Dytiscoidea and the systematic position of the relictual family Hygrobiidae. By careful filtering of the original supermatrices and employing a site-heterogeneous mixture model (CAT-GTR), the interrelationships of the five dytiscoid families can be resolved with confidence. Our phylogenomic result is congruent with the conventional morphology-based phylogenetic tree of Dytiscoidea. Tackling potential sources of systematic error strengthens support for a relationship between Hygrobiidae and Dytiscidae. Integrating various previous studies of the systematic position of the small family Meruidae (Balke et al., 2008; Baca et al., 2017; Beutel et al., 2013, 2019; Toussaint et al., 2015; McKenna et al., 2015), we propose an integrated phylogenetic framework for the six extant families of Dytiscoidea: (Meruidae + Noteridae) + ((Aspidytidae + Amphizoidae) + (Dytiscidae + Hygrobiidae)) (Fig. 4). Based on this tree of Dytiscoidea, it will now be possible to address and test a series of hypotheses regarding the evolution of many critical morphological innovations in Dytiscoidea.

Acknowledgements

Financial support was provided by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB26000000), the National Natural Science Foundation of China (41688103), the Second Tibetan Plateau Scientific Expedition and Research (2019QZKK0706), and the Newton International Fellowship from the Royal Society.

References

Alarie, Y., Beutel, R.G., Watts, C.H., 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.

Baca, S.M., Alexander, A., Gustafson, G.T., Short, A.E.Z., 2017. Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of

338 'Hydradephega'. Syst. Entomol. 42, 1–10.
339 Balke, M., Ribera, I., Beutel, R.G., 2005. The systematic position of Aspidytidae, the diversification
340 of Dytiscoidea (Coleoptera, Adephaga) and the phylogenetic signal of third codon positions. J.
341 Zool. Syst. Evol. Res. 43, 223–242.
342 Balke, M., Ribera, I., Beutel, R., Viloria, A., Garcia, M., Vogler, A.P., 2008. Systematic placement of
343 the recently discovered beetle family Meruidae (Coleoptera: Dytiscoidea) based on molecular
344 data. Zool. Scr. 37, 647–650.
345 Bell, R.T., 1966. Trachypachus and the origin of the Hydradephaga (Coleoptera). The Coleopts. Bull.
346 20, 107–112.
347 Beutel, R.G., 1986. Skelet und Muskulatur des Kopfes und Thorax von *Hygrobia tarda* (Herbst). Ein
348 Beitrag zur Klärung der phylogenetischen Beziehungen der Hydradephaga (Insecta:
349 Coleoptera). Stutt. Beitr. Naturkd. 388, 1–54.
350 Beutel, R.G., 1988. Studies of the metathorax of the trout-stream beetle, *Amphizoa lecontei*
351 Matthews (Coleoptera:Amphizoidae): Contribution towards clarification of the systematic
352 position of Amphizoidae. Int. J. Insect Morphol. Embryol. 17, 63–81.
353 Beutel, R.G., 1998. Trachypachidae and the phylogeny of Adephaga (Coleoptera). Proceedings of the
354 Carabid Symposium, XX. ICE, Firenze. Museo Regionale di Scienze Naturali (Torino) 1998,
355 81–106.
356 Beutel, R.G., Haas, A., 1996. Phylogenetic analysis of larval and adult characters of Adephaga
357 (Coleoptera) using cladistic computer programs. Entomol. Scand. 27, 197–205.
358 Beutel, R.G., Haas, F., 2000. Phylogenetic relationships of the suborders of Coleoptera (Insecta).
359 Cladistics 16, 1–39.
360 Beutel, R.G., Balke, M., Steiner, W.E., 2006. The systematic position of Meruidae (Coleoptera,
361 Adephaga) and the phylogeny of the smaller aquatic adephagan beetle families. Cladistics 22,
362 102–131.
363 Beutel, R.G., Ribera, I., Bininda-Emonds, O.R.P., 2008. A genus-level supertree of Adephaga
364 (Coleoptera). Org. Divers. Evol. 7, 255–269.
365 Beutel, R.G., Roughley, R.E., 1988. On the systematic position of the family Gyrinidae (Coleoptera:
366 Adephaga). J. Zool. Syst. Evol. Res. 26, 380–400.
367 Beutel, R.G., Wang, B., Tan, J.J., Ge, S.Q., Ren, D., Yang, X.K., 2013. On the phylogeny and
368 evolution of Mesozoic and extant lineages of Adephaga (Coleoptera, Insecta). Cladistics 29,
369 147–165.
370 Beutel, R.G., Ribera, I., Fikáček, M., Vasilikopoulos, A., Misof, B., Balke, M., 2019. The
371 morphological evolution of the Adephaga (Coleoptera). Syst. Entomol., in press. DOI:
372 10.1111/syen.12403
373 Blanquart, S., Lartillot, N., 2008. A site-and time-heterogeneous model of amino acid replacement.
374 Mol. Biol. Evol. 25(5), 842–858.
375 Bleidorn C., 2017. Phylogenomics: An Introduction, first ed. Springer, Berlin.
376 Burmeister, E.G., 1976. Der Ovipositor der Hydradephaga (Coleoptera) und seine phylogenetische
377 Bedeutung unter besonderer Berücksichtigung der Dytiscidae. Zoomorphologie 85, 165–257.
378 Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., Embley, T.M., 2008. The archaebacterial origin of
379 eukaryotes. Proc. Natl. Acad. Sci, 105(51), 20356–20361.
380 Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): selection of
381 phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10, 210.
382 Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. In
383 Atlas of Protein Sequence and Structure, M.O. Dayhoff, ed. (National Biomedical Research
384 Foundation), pp. 345–352.
385 Dettner, K., 1985. Ecological and phylogenetic significance of defensive compounds from pygidial
386 glands of Hydradephaga (Coleoptera). Proc. Acad. Nat. Sci. Philadelphia 137, 156–171.
387 Dettner, K., 2016. Hygrobiidae, Régimbart, 1879. In: Beutel, R.G. & Leschen, R.A.B. (Eds.),

Handbook of Zoology. Vol. 4. Arthropoda: Insecta, Part 38, Coleoptera. Vol. 1. Morphology and Systematics (Archostemata, Adephaga, Myxophaga, Polyphaga partim) 2nd edition. Walter de Gruyter, Berlin, New York, pp. 112–118.
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.
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.
Forsyth, D.J., 1970. The structure of the defence glands of the Cicindelidae, Amphizoidae, and Hygrobiidae (Insecta: Coleoptera). J. Zool. 160, 51–69.
Foster, P.G., Cox, C.J., Embley, T.M., 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Phil. Trans. Roy. Soc. B 364(1527), 2197–2207.
Gustafson, G.T., Baca, S.M., Alexander, A.M., Short, A.E., 2019. Phylogenomic analysis of the beetle suborder Adephaga with comparison of tailored and generalized ultraconserved element probe performance. Syst. Entomol., in press. DOI: 10.1111/syen.12413
Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.
Hernandez, A.M., Ryan, J.F., 2019. Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses. BioRxiv Preprint. http://dx.doi.org/10.1101/729103
Ho, S.Y., Jermiin, L.S., 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53(4), 623–637.
Hrdý, I., Hirt, R.P., Doležal, P., Bardonová, L., Foster, P.G., Tachezy, J., Embley, T.M., 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature, 432(7017), 618–622.
Inagaki, Y., Roger, A.J., 2006. Phylogenetic estimation under codon; models can be biased by codon usage heterogeneity. Mol. Phylogenet. Evol. 40, 428–434.
Jermiin, L.S., Ho, S.Y., Ababneh, F., Robinson, J., Larkum, A.W., 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53(4), 638–643.
Kavanaugh, D.H., 1986. A systematic review of Amphizoid beetles (Amphizoidae: Coleoptera) and their phylogenetic relationships to other Adephaga. Proc. Calif. Acad. Sci. 44, 67–109.
Lahr, D.J., Kosakyan, A., Lara, E., Mitchell, E.A., Morais, L., Porfirio-Sousa, A.L., Ribeiro, G.M., Tice, A.K., Pánek, T., Kang, S., Brown, M.W., 2019. Phylogenomics and morphological reconstruction of Arcellinida testate amoebae highlight diversity of microbial eukaryotes in the Neoproterozoic. Curr. Biol. 29, 991–1001.
Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7, S4.
Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.
Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.
Lartillot, N., Philippe, H., 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Phil. Trans. Roy. Soc. B. 363(1496), 1463–1472.
Lartillot, N., Rodrigue, N., Stubbs, D., Richer, J., 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. 62, 611–615.

Le, S.Q., Dang, C.C., Gascuel, O., 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29, 2921–2936.
López-López, A., Vogler, A.P., 2017. The mitogenome phylogeny of Adephaga (Coleoptera). Mol. Phylogenet. Evol. 114, 166–174.
Lozano-Fernandez, J., Tanner, A.R., Giacomelli, M., Carton, R., Vinther, J., Edgecombe, G.D., Pisani, D., 2019a. Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. Nature Commun. 10, 2295.
Lozano-Fernandez, J., Giacomelli, M., Fleming, J., Chen, A., Vinther, J., Thomsen, P.F., Glenner, H., Palero, F., Legg, D.A., Iliffe, T.M., Pisani, D., Olesen, J., 2019b. Pancrustacean evolution illuminated by taxon-rich genomic-scale data sets with an expanded Remipede sampling. Genome Biol. Evol. DOI:10.1093/gbe/evz097
Martijn, J., Vosseberg, J., Guy, L., Offre, P., Ettema, T.J., 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature 557, 101–105.
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.
McKenna, D.D., Shin, S., Ahrens, D., Balke, M., Beza-Beza, C., Clarke, D.J., Donath, A., Escalona, H.E., Friedrich, F., Letsch, H., Liu, S., Maddison, D., Mayer, C., Misof, B., Murin, P.J., Niehuis, O., Peters, R.S., Podsiadlowski, L., Pohl, H., Scully, E.D., Yan, E.V., Zhou, X., Ślipiński, A., Beutel, R.G., 2019. The evolution and genomic basis of beetle diversity. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.1909655116
Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.
Philippe, H., Poustka, A.J., Chiodin, M., Hoff, K.J., Dessimoz, C., Tomiczek, B., Schiffer, P.H., Muller, S., Domman, D., Horn, M., Kuhl, H., Timmermann, B., Satoh, N., Hikosaka-Katayama, T., Nakano, H., Rowe, M.L., Elphick, M.R., Thomas-Chollier, M., Hankeln, T., Mertes, F., Wallberg, A., Rast, J.P., Copley, R.R., Martinez, P., Telford M.J., 2019. Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. Curr. Biol. 29, 1818–1826.
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.
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.
Rota-Stabelli, O., Lartillot, N., Philippe, H., Pisani, D., 2013. Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. Syst. Biol. 62, 121–133.
Ruhnau, S., 1986. Phylogenetic relations within the Hydradephaga (Coleoptera) using larval and pupal characters. Entomol. Basil. 11, 231–272.
Schwentner, M., Combosch, D. J., Nelson, J.P., Giribet, G., 2017. A phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. Curr. Biol. 27, 1818–1824.
Sheffield, N.C., Song, H., Cameron, S.L., Whiting, M.F., 2009. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. Syst. Biol. 58(4), 381–394.
Shull, V.L., Vogler, A.P., Baker, M.D., Maddison, D.R., Hammond, P.M., 2001. Sequence alignement of adephagan beetles: evidence for monophyly of aquatic families and the placement of Trachypachidae. Syst. Biol. 50, 945–969.
Si Quang, L., Gascuel, O., Lartillot, N., 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24(20), 2317–2323.

Strassert, J.F., Jamy, M., Mylnikov, A.P., Tikhonenkov, D.V., Burki, F., 2019. New phylogenomic analysis of the enigmatic phylum telonemia further resolves the eukaryote tree of life. Mol. Biol. Evol. 36(4), 757–765.

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

Vasilikopoulos, A., Balke, M., Beutel, R. G., Donath, A., Podsiadlowski, L., Pflug, J.M., Waterhouse, R.M., Meusemann, K, Peters, R.S., Escalona, H., Mayer, C., Liu, S., Hendrich, L., Alarie, Y., Bilton, D.T., Jia, F., Zhou, X., Maddison, D.R., Niehuis, O., Misof, B., 2019. Phylogenomics of the superfamily Dytiscoidea (Coleoptera: Adephaga) with an evaluation of phylogenetic conflict and systematic error. Mol. Phylogenetics Evol. 135, 270–285.

Wang, H.C., Li, K., Susko, E., Roger, A.J., 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol. Biol. 8(1), 331.

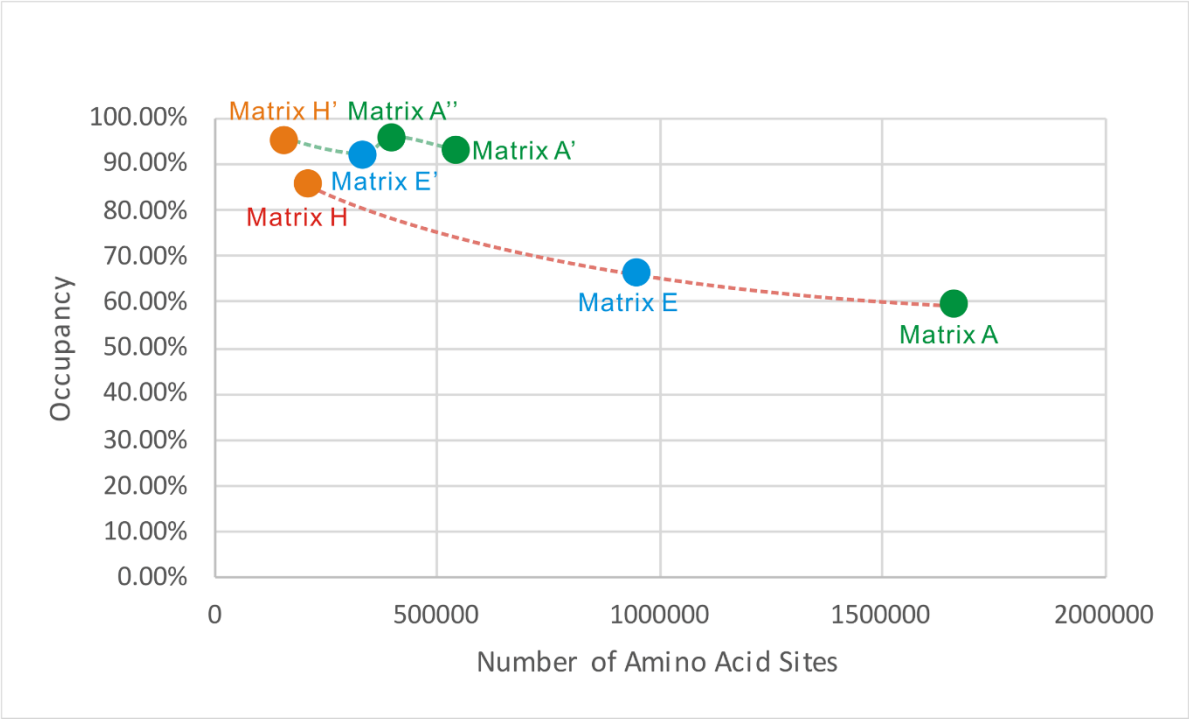
Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J., Embley, T.M., 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat. Ecol. Evol. 4(1), 138–147.

Wolfe, J. M., Breinholt, J. W., Crandall, K. A., Lemmon, A. R., Lemmon, E. M., Timm, L. E., Siddall, M.E., Bracken-Grissom, H.D., 2019. A phylogenomic framework, evolutionary timeline and genomic resources for comparative studies of decapod crustaceans. Proc. R. Soc. B Biol. Sci. 286, 20190079.Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Kiley, W., Seitz, Anantharaman, K., Starnawski, P., Kjeldsen, K.U., Stott, M.B., Nunoura, T., Banfield, J.F., Schramm, A., Baker, B.J., Spang, A., Stott, M.B., 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353–358

514

515 **[Captions]**

516



517

518 **Fig. 1.** Data occupancies and amino acid site numbers of original (Matrices A, E and H) and trimmed
519 (Matrices A', A'', E' and H') supermatrices that were used in the present study.

520

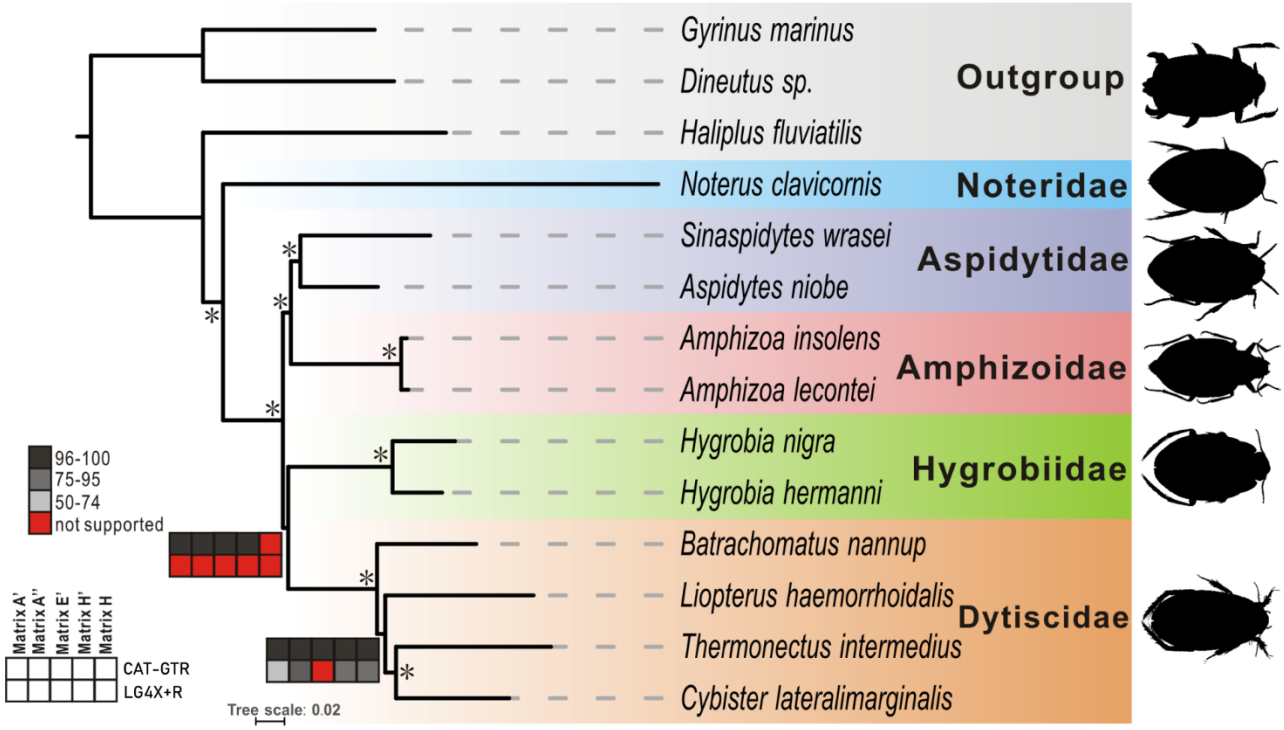


Fig. 2. Phylogenetic tree based on the PhyloBayes analysis of supermatrix A' with the site-heterogeneous CAT-GTR model. Supermatrix A' comprises 14 taxa (11 in-group taxa) and 542,493 amino acid positions. Support values for all analyses are plotted below respective branches as specified in the legend at the bottom-left corner. * denotes strongly supported clades in all analyses (BPP > 0.98 or MLB > 95).

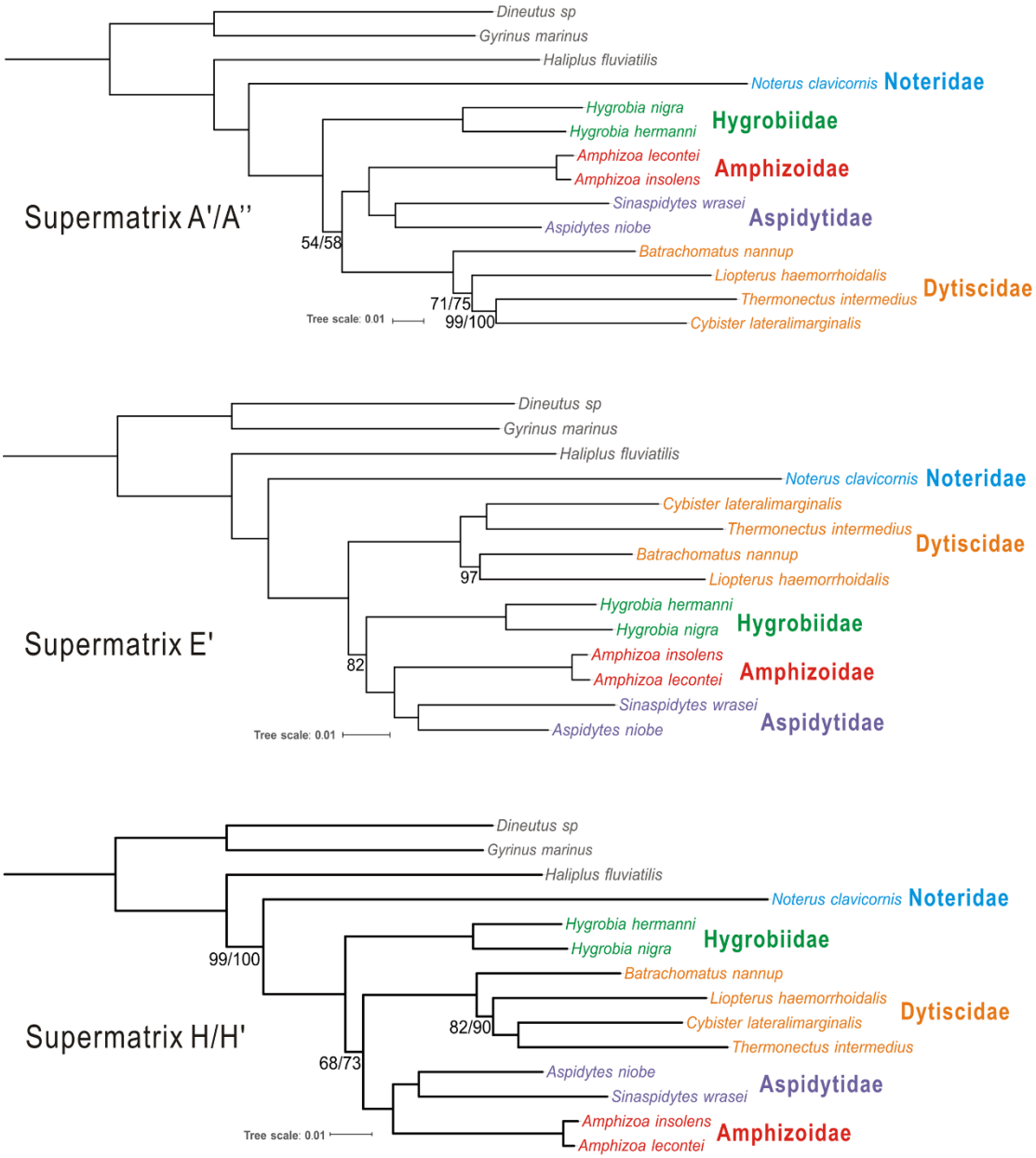


Fig. 3. Different phylogenetic hypotheses deduced from the analysis of amino-acid sequence data (Supermatrices A', E', H' and H) under the simplistic LG4X+R model. Branch support (MLB) is denoted based on 1,000 ultrafast bootstrap replicates; MLB values equal to 100 are not shown).

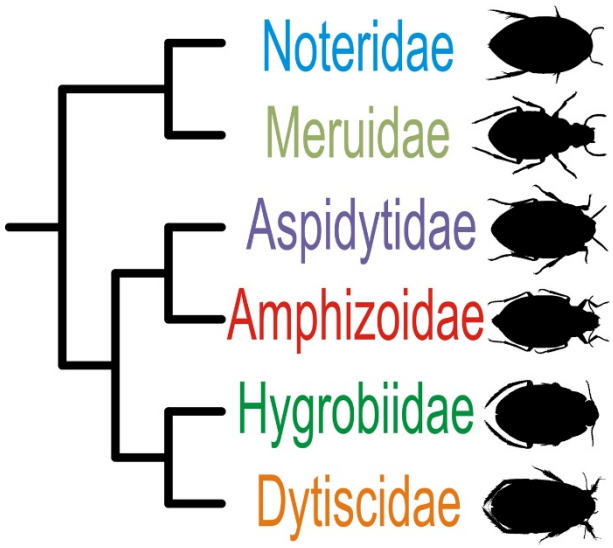


Fig. 4. Phylogenetic hypothesis on family phylogenetic relationships among Dytiscoidea based on the present study and previously published data.