PNAS
www.pnas.org

Supplementary Information for

The Evolution and Genomic Basis of Beetle Diversity

Duane D. McKenna[a,b,1,*], Seunggwan Shin[a,b,*], Dirk Ahrens[c], Michael Balke[d], Cristian Beza-Beza[a,b], Dave J. Clarke[a,b], Alexander Donath[e], Hermes E. Escalona[e,f,g], Frank Friedrich[h], Harald Letsch[i], Shanlin Liu[j], David Maddison[k], Christoph Mayer[e], Bernhard Misof[e], Peyton J. Murin[a], Oliver Niehuis[g], Ralph S. Peters[c], Lars Podsiadlowski[e], Hans Pohl[l], Erin D. Scully[m], Evgeny V. Yan[l,n], Xin Zhou[o], Adam Ślipiński[f], Rolf G. Beutel[l]

[a]Department of Biological Sciences, University of Memphis, Memphis, TN, USA; [b]Center for Biodiversity Research, University of Memphis, Memphis, TN, USA; [c]Center for Taxonomy and Evolutionary Research, Arthropoda Department, Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany; [d]SNSB-Bavarian State Collection of Zoology, 81247 Munich, Germany; [e]Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; [f]Australian National Insect Collection, CSIRO, Canberra, 2601, Australia; [g]Department of Evolutionary Biology and Ecology, Institute for Biology I (Zoology), University of Freiburg, 79104 Freiburg (Brsg.), Germany; [h]Institute of Zoology, University of Hamburg, D-20146 Hamburg, Germany; [i]Department of Botany and Biodiversity Research, University of Wien, Wien 1030, Austria; [j]China National GeneBank, BGI-Shenzhen, Guangdong, People's Republic of China 518083; [k]Department of Integrative Biology, Oregon State University, Corvallis, OR, USA; [l]Institut für Zoologie und Evolutionsforschung, Friedrich-Schiller-Universität Jena, D-07743 Jena, Germany; [m]USDA, Agricultural Research Service, Center for Grain and Animal Health, Stored Product Insect and Engineering Research Unit, Manhattan, KS 66502, USA; [n]Borissiak Paleontological Institute, Russian Academy of Sciences, Moscow, 117997 Russia; [o]Department of Entomology, China Agricultural University, Beijing 100193, People's Republic of China 100193

[1]To whom correspondence should be addressed. Email: dmckenna@memphis.edu.

*These authors contributed equally.

www.pnas.org/cgi/doi/10.1073/pnas.1909655116

This PDF file includes:

Supplementary text
Captions for Figs. S1 to S27
Captions for Tables S1 to S8
Captions for Data S1 to S4

Other supplementary materials for this manuscript include the following:

Figs. S1 to S27
Tables S1 to S8
Data S1 to S4 [Pfam Candidate Genes trees and phy',
Blast_10best_hits_Pfam_Candidate_Genes, Pfam_Candidate_Genes_Fas,
supermatrices_partitions]

Supplementary Information Text

Materials and Methods

1. Taxon sampling. For the 4818-gene tree (Fig. 1) we sampled 146 taxa including outgroup species. Outgroup taxa included seven Neuropterida representing all 3 orders (Megaloptera, Neuroptera and Raphidioptera), and four Strepsiptera representing four families. 135 Coleoptera were sampled representing 90 families, including 118 transcriptomes and 17 genomes (or official gene sets; OGS) (Table S3). Beetle diversification rates and the temporal and phylogenetic locations of diversification rate shifts were estimated using a near-comprehensive family-level timetree generated for the same 147 species as above, plus 374 additional species from ref. (1) (521 total species in 143 families; hereafter, the 89-gene tree; Table S6; Fig. S10). See below for more information.

2. Sample Preparation and Sequencing. Samples collected for this study were preserved in RNAlater. See ref. (2) for details relating to extraction of total mRNA and other details (fragmentation, construction of cDNA libraries, and tagging). All mRNA libraries except for those of *Nosodendron* and *Tetraphalerus* were sequenced with Illumina HiSeq 2000 sequencers (Illumina, San Diego, CA, USA) using paired-end (PE) 150 bp and 90 bp reads, following the methods of ref. (3) and ref. (2). The genomes of *Hydroscapha* and *Priacma* were sequenced using PE100 reads, and the genomes of *Car, Mastostethus*, *Nanophyes, Rhynchitomacerinus,* and *Synolabus* were sequenced on an Illumina Hi-Seq 2500 (dual-indexed libraries; PE 250 reads). *Nosodendron* and *Tetraphalerus* were sequenced on an Illumina Hi-Seq 3000 (dual-indexed libraries; PE150 reads), as were the genomes of *Bembidion, Eucinetus, Saphophagus, Sphaerius, Torridincola,* and *Triozocera*.

3. Assembly. Raw RNA-Seq reads of all 1KITE samples were demultiplexed, quality checked and trimmed. Specifically, reads with adapter contamination (minimum length of the alignment: 15 bp; at most 3 mismatches), reads with more than 10 Ns, and reads with more than 50 bp of low-quality sequence data (i.e., Phred quality score: 2, ASCII 66 "B", Illumina 1.5+ Phred+64) were removed following the methods of ref. (3). After this filtering, we performed a *de novo* assembly of the remaining transcript raw reads for each taxon using SOAPdenovo-Trans-31kmer

(version 1.01) (4) following the methods of ref. (2). In the contig forming step, linear k-mers were merged to form edges and different edges were linked by arcs. Arcs with an abundance < 5% of the total out-degrees or < 2% of the total in-degrees were excluded. Subsequently, edges with an average abundance >= 3 were reported as contigs. *De novo* genome assemblies were undertaken using the programs SOAPdenovo2 (5) or the CLC Genomics Workbench (QIAGEN).

4. Contamination check. We used VecScreen (http://www.ncbi.nlm.nih.gov/tools/vecscreen/) and the UniVec database build 7.1 (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) following ref. (2) to identify and remove vector and linker/adapter contamination from our assemblies. Cross-contamination–due to multiplex sequencing of up to 32 libraries on the same lane on an Illumina sequencer–was evaluated using all-versus-all comparison with BLASTn+ (30) (v2.2.29) following the search strategy used by ref. (31) as implemented by ref. (14): contigs were discarded when their coverage did not exceed more than 2x in comparison to all other highly similar blast hits (=at least 98% identity, and a hit length of at least 180 bp). Finally, an independent contamination check was done by NCBI upon database submission (contigs with blast hits to non insect taxa were filtered). Most of the transcriptomes lost not more then 2% of their sequence information through these checks (Table S4)

5. Orthology prediction. We designed a beetle specific orthologous reference gene set using OrthoDB v7 (6, 7). Specifically, we selected the node Endopterygota in OrthoDB and used this for clustering orthologous sequence clusters (i.e., orthologs or ortholog groups) to generate reference gene IDs (e.g., EOG7XXXXX). We then searched for all single copy genes shared across the gene sets for each of four reference taxa: Hymenoptera: *Nasonia vitripennis* (8), Coleoptera: *Tribolium castaneum* (9), Diptera: *Drosophila melanogaster* (10), and Lepidoptera: *Danaus plexippus* (11). These taxa were selected as references because they represent a broad diversity of holometabolous insect orders and have relatively high quality annotated OGS sequences. After obtaining the IDs for the orthologous genes shared across these 4 taxa, we looked for additional single copy genes shared by each possible combination of 3 of the 4 taxa that included the Coleoptera reference (*Tribolium castaneum*). This resulted in OrthoDB single copy ortholog searches involving the following 3 combinations of taxa: (1) Coleoptera, Diptera, Lepidoptera, (2) Coleoptera, Diptera, Hymenoptera, and (3) Coleoptera, Lepidoptera, Hymenoptera. Combining the results from these searches resulted in 6034 unique single copy reference genes. We extracted the reference genes (using EOG7 IDs) from the gene sets of these reference species and downloaded the official gene sets that had been used from OrthoDB v7 (ftp://cegg.unige.ch/OrthoDB7/FASTA/). We manually modified the reference table downloaded from OrthoDB to match the downloaded OGS IDs (12). This resulted in a custom-made ortholog set including 6034 OGs that served as input for Orthograph. See ref. (12) for further details.

We used Orthograph (12) to generate a profile hidden Markov model (pHMM) from the amino acid (AA) sequences of each reference gene based on the OrthoDB7 EOG IDs. We then used the Orthograph pipeline to convert nucleotide (NT) sequences to the corresponding protein sequences using 6-frame translation in the program Exonerate (13). BLASTp was then used to search the translated query protein sequences against the reference database. Both the pHMM and BLAST search results were stored in a database for orthology prediction at the next (orthograph-reporter) stage. The results from the BLAST search were checked against the pHMM hit of the transcripts, this serving to identify the best-hit sequences for each reference gene ID for each target taxon. The same Orthograph pipeline was used to search all genome and

transcriptome (RNASeq) data used in our study (Tables S3 and S6) with non-strict reciprocal searches using the defaults for all Orthograph parameters. After processing in Orthograph, 6033 of the 6034 reference genes were found in our genome and transcriptome data.

6. Multiple sequence alignment (MSA). The fasta files generated for each EOG from the orthology prediction pipeline in Orthograph were summarized using a script from Orthograph (12), and the resulting fasta files were saved in a folder organized by EOG ID. MSAs were then undertaken for each EOG using the L-INS-I algorithm in MAFFT v7.130b (14). We used Pal2Nal v14 (15) as described in ref. (3) and ref. (2) to construct a multiple codon (NT) alignment from the corresponding aligned protein sequences after MSA masking (see below).

7. MSA masking. The alignments were checked to remove outliers following ref. (3). OGs with putative alignment ambiguities or randomized alignment sections were identified on the amino-acid level with Aliscore v1.2 (16-19) using the settings described in ref. (3). After generating corresponding lists for the NT data with custom-made Perl scripts, ambiguously aligned sections were removed from both AA and NT MSAs using the helper script Alicut (https://github.com/PatrickKueck/AliCUT).

1215 of the 6033 genes were identified as misaligned following the methods of ref. (3) and ref. (2) and were removed in their entirety, leaving 4818 genes. After removal of ambiguously aligned sequence sections, alignment gaps (-) at the beginning and end of the sequences were modified to 'X' (translational level) and 'N' (transcriptional level) for further analyses. The resulting masked MSAs were concatenated to form a supermatrix partitioned by gene using FASconCAT v1.0 (20), which at the same time generated information on partitioning (gene boundaries). This was done separately for the AA and NT data, resulting in aligned supermatrices of length 1,907,014 AA sites and 5,721,042 NT sites.

8. Assessing information content, partitioning and model selection. Information content in each AA gene partition was evaluated using the program MARE v 0.1.2-rc (21) with default settings. PartitionFinder 1.1.1 (22) was used to identify the best-fit cluster of partitions for the AA dataset using the rcluster (23) option. We used the -AUTO option in RAxML v8.2.10 (24-26) to select a best-fit model of evolution (the LG+G model), which was used for all partitions in maximum likelihood (ML) phylogenetic analyses. We used the same partitions for the NT dataset. The GTR+G+I model was selected as the best-fit model for additional (separate) analyses of (a) first codon positions (C1), (b) second codon positions (C2) and (c) first and second codon positions combined (C12). We used the corrected Akaike information criterion (AICc) (27) implemented in PartitionFinder to establish clusters of partitions using the following parameters: model_selection = AICC; models = LG+G; branch lengths = linked; search = rcluster, and the options: as rate = 1.0, base = 1.0, model = 0.0, alpha = 1.0 with rcluster-percent = 0.1, using the command line script with the –raxml option. PartitionFinder suggested a total of 793 partitions containing the 4818 genes. NT based analyses were conducted using the same partitions as the AA data

9. Phylogenetic analyses. We used RAxML v8.2.10 (26) to implement ML phylogenetic analyses. All datasets were prepared for analysis following the methods described above and using the partitions identified as described above using PartitionFinder. We used the LG+G and GTR+G+I models for analysis of the AA and NT (C1, C2, C12) datasets, respectively. RAxML does not support the application of different models to different partitions in the same analysis for NT data, so we used the aforementioned models for all partitions in each analysis. We

completed 10 thorough ML tree searches and 100 slow bootstrap replicates on the University of Memphis High Performance Computing (HPC) cluster using 48 cores and 1TB Ram and on the University of Vienna Department of Botany and Biodiversity Research supercomputer using 64 cores and 396GB Ram. The best tree was selected based on likelihood scores.

To further investigate statistical support for nodes in the beetle phylogeny that differed in the 4818-gene and 89-gene phylogenies, or higher-level relationships that lacked strong statistical support in the 4818-gene phylogeny, we performed four-cluster likelihood quartet mapping (FcLM) in IQ-tree v 1.6 (28). Specifically, we addressed fourteen hypotheses relating to higher-level relationships in beetles (Fig. S8). These were: A. monophyly of Coleoptera; B. (Archostemata + Myxophaga + Adephaga), sister to Polyphaga; C. (Archostemata + Myxophaga), sister to Adephaga; D. Gyrinidae sister to the remaining Adephaga; E. Scirtidae sister to the remaining Polyphaga; F. monophyly of (Derodontidae + Clambidae + Eucinetidae), or polyphyly of Scirtoidea; G. monophyly of (Buprestidae + Dryopidae + Heteroceridae); H. monophyly of (Elateroidea +*Notolioon* [Byrrhidae] + *Byrrhus* [Byrrhidae]); I. monophyly of Cucujiformia; J. monophyly of (Scarabaeoidea + Staphylinoidea); K. monophyly of (Cleroidea + Coccinelloidea); L. (Lymexyloidea + Elateroidea + Tenebrionoidea) sister to (Phytophaga + Cucujoidea); M. Erotylidae sister to (Phytophaga + remaining Cucujoidea); N. (Laemophloeidae + Monotomidae + Nitidulidae) sister to (Phytophaga + remaining Cucujoidea). (Fig. S9). For the FcLM analyses, we further condensed each data set to include only genes for which the targeted groups had gene data coverage (decisive data sets). To obtain sufficient coverage for each hypothesis, the total number of quartets calculated was 25 times the number of sequences in the original alignment. We used the same data partitioning scheme and models that were used for phylogenetic analysis of the AA dataset in RAxML (see above). The clustering information is indicated in Fig. S9.

10. Fossils for calibration. We surveyed the literature of fossil Coleoptera for candidate fossil constraints to use in divergence dating analyses, aiming to provide a balanced distribution of calibrations across mainly higher taxa in the tree for each of the two analyses (Figs. 1 and S11). Note that the calibration number refers to the number assigned to the fossil in Figs. 1 and S11, and Table S5. We gathered the oldest fossil representatives of major beetle lineages (families in most cases) and applied a more refined selection process in which we generally screened fossils according to the criteria established by ref. (29). However, we chose to adopt a slightly less conservative approach than the one advocated by these authors as we think that in practice criteria 2 and 3 are difficult to apply in studies of beetle divergence dating (and insects generally) and that in many cases authors of taxonomic descriptions nevertheless provide compelling arguments and evidence justifying the taxonomic placement of fossil taxa. In insects, morphological and molecular phylogenies for similar sets of taxa are rare (more common considering overlapping representation of higher taxa) making it somewhat difficult to apply their criterion 3 as originally described. Moreover, unlike some other groups, it is far less common in insects for newly described fossils to be placed in a phylogeny in the original description, also making it largely impractical at present to strictly apply their criterion 2. Despite these factors, the variation in quality of original descriptions (and of their included fossils) ranges from very poor and unreliable to high quality and thorough analyses (for both fossil preservation and the documentation of characters). We selected fossils from papers providing sufficient descriptions and discussion of characters. Fossils judged as possessing apomorphic characters and/or a diagnostic set of characters of the respective higher taxa were preferentially selected. In this respect we considered an absence of strong disagreement or

uncertainty important in the selection process and made these decisions based solely on published information (*contra* (30)), since publications have at least in theory been vetted by the review process and expert opinion can be misleading if specimens in question have not been physically studied (and, ideally, redescribed/reinterpreted).

For each selected fossil we compiled the following information (Table S5), including information relevant to the five criteria (CR1–5) described by (29): 1) current classification; 2) Museum index numbers (CR1); 3) locality of fossil origin; 4) a summary of the apomorphies and/or phylogenetic justification found in original publications (CR2); 5) agreement of morphological and molecular data when relevant (CR3); 6) details of the geological strata/formation where a fossil was found (CR4); 7) a reference given to the age of the strata (CR5); 8) additional information justifying selection of a constraint; 9) upper and lower bounds on the ages of fossil-bearing strata; notes on fossil-bearing strata gathering largely from The Paleobiology Database (https://paleobiodb.org/). In most cases we standardized the information on geological strata, locality and ages to that listed in The Paleobiology Database, except that the ends of date ranges inferred by direct dating of deposits were used for upper and lower bound values when these were available.

We applied a minimum age constraint to the node immediately descendant of the crown group to which the fossil had been assigned, a generally more conservative approach that also in part accounts for the difficulty in strictly applying criteria 2 and 3. We applied constraints to 18 nodes in the 4818-gene tree (Fig. 1) and 22 nodes in the 89-gene tree (Table S5). The former analysis included one fossil not included in the latter analysis (*Cretomalthus acracrowsonarum*); the latter analysis included the same ones included in the 4818-gene tree analysis (Fig. 1; except *C. acracrowsonarum*) plus five other constraints that we could add as a result of the much greater taxon sample (521 taxa) in that analysis (see Table S5). In our MCMC analyses we also applied a maximum constraint to the root node representing the Devonian-Carboniferous boundary (358.9 Mya) as a more conservative estimate of the origin of Holometabola (the oldest fossil of which is *Srokalarva berthei*, 307–315 Mya; Table S5).

11. Divergence time estimation. The PAML package containing MCMCtree and CODEML was employed to generate timetrees and estimated divergence times. We used the AA sequence data and the partitioned (best) ML tree as input for our analyses. We used the approximate likelihood method (31) in CODEML to generate the Hessian matrix (LG model, default options). Our 4818-gene dataset was too large for divergence time analysis in PAML. Consequently, we reduced the size of the dataset by using a matrix coverage completeness cutoff of >89%; i.e., we used all alignment positions for which 90% or more of taxa had data. This resulted in an AA sequence matrix containing 206,156 AA positions. Following ref. (2) we performed an unpartitioned analysis of the resulting data. We applied fossil calibrations (Table S5) as soft minimum ages (truncated Cauchy distributions) and used the default program settings for the MCMCtree analysis, as recommended in the user manual. The root constraint (Table S5) was applied as a maximum age (358 Mya). For each calibration point we used the following program options: offset 0.1, scale parameter 1, and left tail probability 0.025. We applied a burnin of 100,000 and sampled 1,000,000 MCMC generations, sampling every 500 iterations. Four separate MCMCtree runs were implemented on the University of Memphis HPC cluster using an independent-rates clock with default parameter settings. We determined the resulting effective sample sizes (ESS) for each parameter using the program Tracer 1.7.1 (32). The ESS for all parameters was greater than 200. The correlated-rates clock option has not been tested for our

dataset, since it has already been shown to be unsuitable for similar datasets (2, 33-35). The resulting output files were further checked for convergence using a custom R plotting script. All four runs were determined to have converged, with only minor differences observed between runs.

12. Diversification rate analyses using the 89-gene tree. During the course of our analyses the genome of *Rhinorhipus tamborinensis* (Rhinorhipoidea), a long enigmatic and little-known beetle became available for study via the publication of ref. (36). We handled this genome the same way as our other genomes and transcriptomes (e.g., processing with the Orthograph pipeline) and subsequent new MSA, MSA filtering and partitioning, and an additional (new) ML tree search using the resulting 147-taxon, 804 partitions containing the 4852-gene, 1,942,580 AA alignment. This dataset differed from the original (4818-gene) alignment only by the addition of *Rhinorhipus* and the unique results of the aforementioned data processing steps (resulting in 4852 genes instead of 4818 genes).

Additionally, we processed the data from (1) using the Orthograph pipeline described above. Eighty-nine genes were shared with our 4818-gene genome/transcriptome dataset after orthology assessment and filtering. We used PartitionFinder for partition and model selection, using the same options as in the analysis of the 4818-gene dataset, except model selection. We evaluated all possible models for this analysis, ultimately using IQ-TREE v1.6 (28) for partition specific model-based ML analysis. We executed 25 ML searches and 100 bootstrap replicates using the 89-gene 521-taxon AA dataset in IQ-TREE. Divergence time analyses were performed following the methods used in the 4818-gene analysis, with some differences in fossil calibrations on account of the expanded taxon sample (detailed above) (Table S5). We applied a burn-in of 10,000 and sampled 100,000 MCMC generations, sampling every 2 iterations. Many nodes had ESS values less than 200; however, the CI mean indicated convergence across the 4 runs, so we used the mean values for the tree with the highest overall ESS values in subsequent temporal analyses.

The timetree resulting from analysis of the 89-gene tree was used for diversification rate analyses because it contained exemplars from more beetle families than the 4818-gene tree (19,951 AA sites; matrix coverage completeness cutoff of >50%; 521 species, 149 families) (see Fig. S10 for the full tree with bootstrap support). Notably, higher-level relationships in this tree were almost identical to the 4818-gene tree, differing only in the relative placements of Coccinelloidea and Cleroidea. The program R (37) was employed to evaluate beetle diversification rates using the dependent packages Ape (38, 39) Geiger (40, 41) and Laser (42). We used MEDUSA in the Geiger package (43) to estimate diversification rates and the timing of family-level branching events in the beetle phylogeny. Values for family-level clade species richness are based on published data from (44). Thirty-nine beetle families were added to the 89-gene timetree by rooting each added family at a random position between the proposed subtending and descendant nodes for a total of 188/190 described extant families (Fig. 2). Table S7 indicates which families were added and the rationale for each placement. Only Crowsoniellidae and Jurodidae, each with one extant species known only from their original type series, were missing from the resulting "hybrid" timetree. In cases where a family was non-monophyletic and needed to be split into more than one family-level clade, the number of species was split evenly across all clades for purposes of analyzing net diversification rates.

13. Molecular phylogenetics and evolution of beetle genes encoding PCWDEs and invertases. We used PfamScan (45) to search our transcriptome and genome sequences against a

library of Pfam HMM. We then used SAMtools (46) to extract individual genes corresponding to our target genes (10 families of glycoside hydrolases: GH1, GH5, GH9, GH10, GH28, GH32, GH43, GH44, GH45, GH48; pectinesterase (PE): pectin methylesterase CE8 (carbohydrate esterase family 8), and rhamnogalacturonate lysase (PL4)) from the genome and transcriptome assemblies. Fasta files containing all of the extracted target sequences can be found as data files S3. We used Megan v4 (47) to check for contamination by bacterial or other sequences.

We evaluated the identity of the extracted sequences using BLASTp on the NCBI NR local blast using BLAST+ and further evaluated some sequences using the NAL BLASTp server, which included all of the annotated reference genomes from the Insect 5000 Genomes Project (not all of which are in NCBI). To exclude most (but potentially not all) 'false positives' and spurious matches derived from symbionts or microbes that may have contaminated the samples we retained sequences that had matches to non-microbes and non-plants. Some sequences could not be confidently assigned to the gene family indicated by the pfam search and were excluded from downstream analyses. For example, we omitted 30 sequences with pfam matches to GH5 (mostly from Tenebrionoidea; e.g., the sequences from *Bitoma*) because they had stronger NCBI BLASTp matches to exo-1,3-beta glucanases and better matches to GH1 pfam domains and thus are unlikely cellulases.

We then gathered homologs for each gene of interest from previously published studies by reviewing the published literature and searching the NCBI protein database for genes of interest from each major lineage in the NCBI classification (with a particular focus on bacteria, fungi, plants, and nematodes), and by BLASTp searching our target sequences against the NCBI NR protein database and keeping a subset of the best hits. When possible, we retained sequences that were identified to the species level over those that were not. The resulting sequences were combined with our filtered pfam target sequences to produce one fasta file for each gene. Each gene was then aligned using MAFFT (e-INSI algorithm). The resulting MSAs (one per gene) were trimmed on the 5' and 3' ends to reduce the amount of missing data in the matrix.

Sequences containing unusually long indels or large stretches of low quality or unaligned data were excluded from the matrix unless they represented unique beetle taxa or other lineages that were not otherwise represented. We also removed sequences showing less than 1% pairwise sequence divergence and other highly redundant sequences (i.e., most isoforms) as well as highly incomplete sequences (typically, those with less than 50% of the data present). Therefore, not all genes present for all taxa shown in Fig. 1 were included in phylogenetic analyses. The remaining aligned sequences were collapsed to remove alignment gaps and realigned in MAFFT (e-INSI) in preparation for phylogenetic analysis.

ML phylogenetic analyses were undertaken using the program IQ-TREE (28). We performed 10 ML searches and 100 bootstrap replicates for each gene using the aligned AA dataset after model test in the program IQ-TREE. We calculated transfer bootstrap expectation (TBE) values for each node in the resulting phylogenies using BOOSTER (48) (Figs. S15-26).

Data use statement. Data on genetic material contained in this paper are published for non-commercial use only. Utilization by third parties for purposes other than non-commercial scientific research may infringe the conditions under which the genetic resources were originally accessed, and should not be undertaken without obtaining consent from the corresponding author of the paper and/or obtaining permission from the original provider of the genetic material.

Fig. S1. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the partitioned amino acid supermatrix (147 taxa; 4852 genes, 10 replicate ML searches), including the taxa from Fig. S2 (the 4818 gene tree) + Rhinorhipidae: *Rhinorhipus*.

Fig. S2. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the partitioned amino acid supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

Fig. S3. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the un-partitioned amino acid supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

Fig. S4. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the partitioned nucleotide codon 1 (C1) supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

Fig. S5. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the partitioned nucleotide codon 2 (C2) supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

Fig. S6. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the partitioned nucleotide codon 1 and 2 (C1+C2) supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

Fig. S7. (separate file) Summary tree showing superfamily-level clades and summarizing statistical measures of nodal support from the six above-described ML analyses (Figs. S2-S7).

Fig. S8. (separate file) Testing of alternative phylogenetic hypotheses using four-cluster likelihood mapping (FcLM) was undertaken for nodes representing interrelationships among higher-level taxa in the 4818-gene ML tree (Fig. S2) that had less than maximal statistical support (from ML bootstrapping) and/or recovered unexpected relationships. A. monophyly of Coleoptera; B. (Archostemata + Myxophaga + Adephaga), sister to Polyphaga; C. (Archostemata + Myxophaga), sister to Adephaga; D. Gyrinidae sister to the remaining Adephaga; E. Scirtidae sister to the remaining Polyphaga; F. monophyly of (Derodontidae + Clambidae + Eucinetidae), or polyphyly of Scirtoidea; G. monophyly of (Buprestidae + Dryopidae + Heteroceridae); H. monophyly of (Elateroidea +*Notolioon* [Byrrhidae] + *Byrrhus* [Byrrhidae]); I. monophyly of Cucujiformia; J. monophyly of (Scarabaeoidea + Staphylinoidea); K. monophyly of (Cleroidea + Coccinelloidea); L. (Lymexyloidea + Elateroidea + Tenebrionoidea) sister to (Phytophaga + Cucujoidea); M. Erotylidae sister to (Phytophaga + remaining Cucujoidea); N. (Laemophloeidae

+ Monotomidae + Nitidulidae) sister to (Phytophaga + remaining Cucujoidea). For each analysis, values at the corners indicate the percentage of fully resolved phylogenies for all possible quartets.

Fig. S9. (separate file) Results from FcLM analyses for 14 hypotheses based on the nodes indicated in Fig. S8.

Fig. S10. (separate file) Best tree resulting from maximum likelihood (ML) analysis of the partitioned amino acid supermatrix (521 taxa, 89 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node. Asterisks (*) indicate data generated for the present study.

Fig. S11. (separate file) Chronogram inferred from the 89-gene dataset showing the calibration points used in the time divergence analysis. The divergence time analysis was based on a maximum likelihood (ML) analysis of 521 taxa and 89 genes (Fig. S10), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 19,951 amino acid sites. Nodes constrained by fossil priors are indicated with numbers (see Table S5). The calibration points and fossil ages are shown in red. Asterisks (*) indicate data generated for the present study.

Fig. S12. (separate file) Chronogram inferred from the 89-gene dataset showing 95% CIs for node ages estimates. Based on a maximum likelihood (ML) analysis of 521 taxa 89 genes (Fig. S10), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 19,951 amino acid sites. Asterisks (*) indicate data generated for the present study. Node ages refer to hundreds of millions of years.

Fig. S13. (separate file) Chronogram inferred from the 4818-gene dataset showing (A) mean values and (B) 95% CIs values for node ages estimates. Based on a maximum likelihood (ML) analysis of 146 taxa 4818 genes (Fig. S2), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 206,156 amino acid sites. Node ages refer to hundreds of millions of years.

Fig. S14. (separate file) Chronogram inferred from the 89-gene dataset showing (A) mean values and (B) 95% CIs for node ages estimates. Based on a maximum likelihood (ML) analysis of 521 taxa 89 genes (Fig. S10), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 19,951 amino acid sites. Asterisks (*) indicate data generated for the present study. Node ages refer to hundreds of millions of years.

Fig. S15. (separate file) Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 1 family genes in the program RAxML (10

replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S16. (separate file) Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 5 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S17. (separate file) Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 9 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S18. (separate file) Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 10 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S19. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 28 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S20. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 32 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S21. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 43 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S22. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 44 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S23. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 45 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S24. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 48 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S25. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for CE8 genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S26. **(separate file)** Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for PL4 genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo (10.5281/zenodo.3522944).

Fig. S27. **(separate file)** Schematics showing annotated genomic scaffolds that contain genes inferred to encode putative plant cell wall degrading enzymes and GH32 invertases, including introns (when present), eukaryotic TSS, and polyA signals. The scaffolds are organized by gene family and are shown only for exemplars from beetle families from which these genes have not previously been reported. The scaffolds were annotated using FGENESH version 2.6 (http://www.softberry.com/berry.phtml?topic=fgenesh&group=help&subgroup=gfind). We used the *Tribolium castaneum* (Tenebrionoidea: Tenebrionidae) genome as a reference for gene prediction.

Abbreviations: CDSf - first exon (beginning with start codon), CDSi - internal (internal exon), CDSl - last (ending with stop codon), CDSo - one (only one exon), TSS - position of transcription start (TATA-box position and score), PolA - polyA signal.

Table S1. (separate file) Distribution of endogenous plant cell wall degrading enzymes: GH5, GH9, GH10, GH11 (no compelling matches in the present study), GH28, GH43, GH44, GH45, GH48, CE8, PL4 and GH32 invertases in beetles*.

Table S2. (separate file) Genes implicated in plant cell wall (PCW) degradation in beetles and their enzymatic activities.

Table S3. (separate file) List of taxa used in analyses of the 4818-gene data set, indicating data type (transcriptome or genome) and origin (current study or publicly available data).

Table S4. (separate file) Contamination check results for the 1KITE transcriptomes.

Table S5. (separate file) List of fossils and the nodes they constrained in separate divergence time analyses of the 4818-gene data set and the 89-gene data set.

Table S6. (separate file) List of taxa used in analyses of the 521-taxon 89-gene dataset, indicating data type (transcriptome or genome) and origin (current study or publicly available data).

Table S7. (separate file) List of family-level clade species richness and sources cited for placement of added taxa in the 89-gene ML tree (Fig. S2). These data were used to produce the timetree and in the analysis of diversification rates shown in Fig. 2 in the main paper.

Table S8. (separate file) Summary of pfam results for the gene families studied for the 147 taxa in Fig. 1 in the main paper.

Datasets S1-S4. (separate files)

Available at Zenodo: DOI: 10.5281/zenodo.3522944

Dataset S1.

Gene trees for plant cell wall degrading enzyme phylogenetic analyses in the directory 'Pfam Candidate Genes trees and phy'

1. Phylip formatted files for each gene used in ML analyses.
2. ML tree files for each gene studied showing TBE bootstrap support (100 replicates) (corresponding to Figs. S15-S26).
3. ML tree files for each gene studied showing ML bootstrap support (100 replicates) from IQtree (corresponding to Figs. S15-S26).

Dataset S2.

Directory (Blast_10best_hits_Pfam_Candidate_Genes) including: Blast results (10 best hits) for all sequences extracted from the transcriptome and genome assemblies for the plant cell wall degrading enzyme analysis in the directory 'Pfam_Candidate_Genes_Fas' (before filtering).

Dataset S3.

Directory (Pfam_Candidate_Genes_Fas) including: Candidate genes/transcripts encoding plant cell wall degrading enzymes extracted from the transcriptome and genome assemblies (before filtering).

Dataset S4.

Directory (Supermatrices_partitions) including:

- Supermatrix for Fig. 1 (amino acid and nucleotide level, PHYLIP formats) and Supermatrix for Fig. S10 (amino acid level, PHYLIP format).
- Partition schemes for supermatrix for Fig. 1 (Fig_1_Partition_finder_best_scheme).
- Partition schemes for supermatrix for Fig. 1 prior to Partitionfinder (AA_partitions).
- Partition schemes for supermatrix for NT prior to Partitionfinder (NT_partitions).
- Partition schemes for supermatrix for Fig. S10 (Fig_S10_Partition_finder_best_scheme).

References

1. Zhang SQ, *et al.* (2018) Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nat Commun* 9(1):205.
2. Peters RS, *et al.* (2017) Evolutionary history of the Hymenoptera. *Curr Biol* 27(7):1013-1018.
3. Misof B, *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210):763-767.
4. Xie Y, *et al.* (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30(12):1660-1666.
5. Luo R, *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1(1):18.
6. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, & Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* 41(Database issue):D358-365.
7. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, & Kriventseva EV (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res* 39(Database issue):D283-288.
8. Werren JH, *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327(5963):343-348.
9. Richards S, *et al.* (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452(7190):949-955.
10. Adams MD, *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287(5461):2185-2195.
11. Zhan S, Merlin C, Boore JL, & Reppert SM (2011) The monarch butterfly genome yields insights into long-distance migration. *Cell* 147(5):1171-1185.
12. Petersen M, *et al.* (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* 18(1):111.
13. Slater GS & Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
14. Katoh K & Standley DM (2013) MAFFT Multiple sequence alignment software Version 7: Improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.
15. Suyama M, Torrents D, & Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34:W609-W612.
16. Meusemann K, *et al.* (2010) A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol* 27(11):2451-2464.
17. Li B, Lopes JS, Foster PG, Embley TM, & Cox CJ (2014) Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. *Mol Biol Evol* 31(7):1697-1709.
18. Kuck P, *et al.* (2010) Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool* 7:10.
19. Misof B & Misof K (2009) A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst Biol* 58(1):21-34.

20. Kuck P & Meusemann K (2010) FASconCAT: Convenient handling of data matrices. *Mol Phylogenet Evol* 56(3):1115-1118.
21. Misof B*, et al.* (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics* 14:348.
22. Lanfear R, Calcott B, Ho SY, & Guindon S (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* 29(6):1695-1701.
23. Lanfear R, Calcott B, Kainer D, Mayer C, & Stamatakis A (2014) Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol* 14:82.
24. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688-2690.
25. Stamatakis A, Ludwig T, & Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21(4):456-463.
26. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312-1313.
27. Hurvich CM & Tsai CL (1989) Regression and time series model selection in small samples. *Biometrika* 76:297-307.
28. Nguyen LT, Schmidt HA, von Haeseler A, & Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32(1):268-274.
29. Parham JF*, et al.* (2012) Best practices for justifying fossil calibrations. *Syst Biol* 61(2):346-359.
30. Toussaint EFA*, et al.* (2017) The peril of dating beetles. *Syst Entomol* 42:1-10.
31. dos Reis M & Yang Z (2011) Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol Biol Evol* 28(7):2161-2172.
32. Rambaut A, Drummond AJ, Xie D, Baele G, & Suchard MA (2018) Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst Biol* 67(5):901-904.
33. Linder M, Britton T, & Sennblad B (2011) Evaluation of Bayesian models of substitution rate evolution-parental guidance versus mutual independence. *Syst Biol* 60(3):329-342.
34. Lepage T, Bryant D, Philippe H, & Lartillot N (2007) A general comparison of relaxed molecular clock models. *Mol Biol Evol* 24(12):2669-2680.
35. dos Reis M, Donoghue PC, & Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet* 17(2):71-80.
36. Kusy D*, et al.* (2018) Genome sequencing of *Rhinorhipus* Lawrence exposes an early branch of the Coleoptera. *Front Zool* 15:21.
37. R_Core_Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
38. Paradis E & Schliep K (2018) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3):526-528
39. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289-290.
40. Pennell MW*, et al.* (2014) geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 30(15):2216-2218.
41. Harmon LJ, Weir JT, Brock CD, Glor RE, & Challenger W (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics* 24(1):129-131.

42. Rabosky DL (2007) LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evol Bioinform Online* 2:273-276.
43. Harmon LJ, Rabosky DL, FitzJohn RG, & Brown JW (2011) MEDUSA: Modeling Evolutionary Diversification Using Stepwise AIC. Version: 0.12.
44. Ślipiński SA, Leschen RAB, & Lawrence JF (2011) Order Coleoptera Linnaeus, 1758. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (ed. by Z.-Q. Zhang). *Zootaxa* 3148:203-208.
45. Finn RD*, et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue):D222-230.
46. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
47. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, & Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 21(9):1552-1560.
48. Lemoine F*, et al.* (2018) Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* 556(7702):452-456.
49. McKenna DD*, et al.* (2016) Genome of the asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biol* 17(1):227.
50. Zhao C, Doucet D, & Mittapalli O (2014) Characterization of horizontally transferred beta-fructofuranosidase (ScrB) genes in *Agrilus planipennis*. *Insect Mol Biol* 23(6):821-832.
51. Ohio Agricultural Research and Development Center, (OARDC), "Horizontal Gene Transfer and Gene Duplication of Plant Cell Wall Degrading Enzyme Genes in an Invasive Insect Pest" 2014; https://kb.osu.edu/bitstream/handle/1811/60352/OARDCposter2014-2.pdf). [the easiest access to this source is via the URL].
52. Eyun SI*, et al.* (2014) Molecular evolution of glycoside hydrolase genes in the western corn rootworm (*Diabrotica virgifera virgifera*). *Plos One* 9(4):e94052
53. Antony B, Johny J, Aldosari SA, & Abdelazim MM (2017) Identification and expression profiling of novel plant cell wall degrading enzymes from a destructive pest of palm trees, *Rhynchophorus ferrugineus*. *Insect Mol Biol* 26(4):469-484.
54. Evans JD*, et al.* (2018) Genome of the small hive beetle (*Aethina tumida*, Coleoptera: Nitidulidae), a worldwide parasite of social bee colonies, provides insights into detoxification and herbivory. *GigaScience* 7(12).
55. Chang CJ*, et al.* (2012) A novel exo-cellulase from white spotted longhorn beetle (*Anoplophora malasiaca*). *Insect Biochem Mol Biol* 42(9):629-636.
56. Pauchet Y, Wilkinson P, Chauhan R, & Ffrench-Constant RH (2010) Diversity of beetle genes encoding novel plant cell wall degrading enzymes. *PLoS One* 5(12):e15635.
57. Lee SJ*, et al.* (2005) A novel cellulase gene from the mulberry longicorn beetle, Apriona germari: Gene structure, expression, and enzymatic activity. *Comp Biochem Phys B* 140(4):551-560.
58. Wei YD*, et al.* (2006) Molecular cloning, expression, and enzymatic activity of a novel endogenous cellulase from the mulberry longicorn beetle, Apriona germari. *Comp Biochem Phys B* 145(2):220-229.

59. Pauchet Y, Kirsch R, Giraud S, Vogel H, & Heckel DG (2014) Identification and characterization of plant cell wall degrading enzymes from three glycoside hydrolase families in the cerambycid beetle *Apriona japonica*. *Insect Biochem Mol Biol* 49:1-13.
60. Liu J*, et al.* (2015) Endogenous cellulolytic enzyme systems in the longhorn beetle *Mesosa myops* (Insecta: Coleoptera) studied by transcriptomic analysis. *Acta Biochim Biophys Sin* 47(9):741-748.
61. Calderon-Cortes N, Watanabe H, Cano-Camacho H, Zavala-Paramo G, & Quesada M (2010) cDNA cloning, homology modelling and evolutionary insights into novel endogenous cellulases of the borer beetle *Oncideres albomarginata chamela* (Cerambycidae). *Insect Mol Biol* 19(3):323-336.
62. Noda H, Watanabe H, & Scrivener AM (1997) Diet and carbohydrate digestion in the yellow-spotted longicorn beetle *Psacothea hilaris*. *J Insect Physiol* 43(11):1039-1052.
63. Sugimura M, Watanabe H, Lo N, & Saito H (2003) Purification, characterization, cDNA cloning and nucleotide sequencing of a cellulase from the yellow-spotted longicorn beetle, *Psacothea hilaris*. *Eur J Biochem* 270(16):3455-3460.
64. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, & Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42(Database issue):D490-495.
65. Busch A, Kunert G, Heckel DG, & Pauchet Y (2017) Evolution and functional characterization of CAZymes belonging to subfamily 10 of glycoside hydrolase family 5 (GH5_10) in two species of phytophagous beetles. *PLoS One* 12(8):e0184305.
66. Salem H*, et al.* (2017) Drastic genome reduction in an herbivore's pectinolytic symbiont. *Cell* 171(7):1520-1531 e1513.
67. Fujita K, Shimomura K, Yamamoto K, Yamashita T, & Suzuki K (2006) A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. *Biochem Biophys Res Commun* 345(1):502-507.
68. Kirsch R*, et al.* (2012) Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. *BMC Genomics* 13:587.
69. Pauchet Y & Heckel DG (2013) The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proc Biol Sci* 280(1763):20131021.
70. Girard C & Jouanin L (1999) Molecular cloning of cDNAs encoding a range of digestive enzymes from a phytophagous beetle, Phaedon cochleariae. *Insect Biochem Mol Biol* 29(12):1129-1142.
71. Kirsch R, Kunert G, Vogel H, & Pauchet Y (2018) Pectin digestion in herbivorous beetles: Impact of pseudoenzymes exceeds that of their active counterparts. *bioRxiv*. doi: https://doi.org/10.1101/462531
72. Nakayama DG*, et al.* (2017) A transcriptomic survey of Migdolus fryanus (sugarcane rhizome borer) larvae. *Plos One* 12(3):e0173059.
73. Schoville SD*, et al.* (2018) A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae). *Sci Rep* 8(1):1931.
74. Aw T*, et al.* (2010) Functional genomics of mountain pine beetle (*Dendroctonus ponderosae*) midguts and fat bodies. *BMC Genomics* 11:215.

75. Keeling CI, *et al.* (2013) Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biol* 14(3):R27.
76. Doostdar H, McCollum TG, & Mayer RT (1997) Purification and characterization of an endo-polygalacturonase from the gut of west indies sugarcane rootstalk borer weevil (*Diaprepes abbreviatus* L.) Larvae. *Comp Biochem Physiol B Biochem Mol Biol* 118:861-867.
77. Acuna R, *et al.* (2012) Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci USA* 109(11):4197-4202.
78. Padilla-Hurtado B, *et al.* (2012) Cloning and expression of an endo-1,4-beta-xylanase from the coffee berry borer, *Hypothenemus hampei*. *BMC Res Notes* 5:23.
79. Vega FE, *et al.* (2015) Draft genome of the most devastating insect pest of coffee worldwide: the coffee berry borer, *Hypothenemus hampei*. *Sci Rep* 5:12525.
80. Shen ZC, Reese JC, & Reeck GR (1996) Purification and characterization of polygalacturonase from the rice weevil, *Sitophilus oryzae* (Coleoptera: Curculionidae). *Insect Biochem Mol Biol* 26(427-433).
81. Shen ZC, Manning G, Reese JC, & Reeck GR (1999) Pectin methylesterase from the rice weevil, *Sitophilus oryzae* (L.) (Coleoptera: Curculionidae): purification and characterization. *Insect Biochem Mol Biol* 29:209-214.
82. Shen Z, *et al.* (2003) Polygalacturonase from *Sitophilus oryzae*: possible horizontal transfer of a pectinase gene from fungi to weevils. *J Insect Sci* 3:24.
83. Shen Z, *et al.* (2005) Pectinmethylesterase from the rice weevil, *Sitophilus oryzae*: cDNA isolation and sequencing, genetic origin, and expression of the recombinant enzyme. *J Insect Sci* 5:21.
84. Kirsch R, Heckel DG, & Pauchet Y (2016) How the rice weevil breaks down the pectin network: Enzymatic synergism and sub-functionalization. *Insect Biochem Mol Biol* 71:72-82.
85. Pedezzi R, *et al.* (2014) A novel beta-fructofuranosidase in Coleoptera: Characterization of a beta-fructofuranosidase from the sugarcane weevil, *Sphenophorus levis*. *Insect Biochem Mol Biol* 55:31-38.
86. Evangelista DE, de Paula FF, Rodrigues A, & Henrique-Silva F (2015) Pectinases from *Sphenophorus levis* Vaurie, 1978 (Coleoptera: Curculionidae): putative accessory digestive enzymes. *J Insect Sci* 15:168.
87. Genta FA, Bragatto I, Terra WR, & Ferreira C (2009) Purification, characterization and sequencing of the major beta-1,3-glucanase from the midgut of *Tenebrio molitor* larvae. *Insect Biochem Mol Biol* 39(12):861-874.
88. Willis JD, Oppert B, Oppert C, Klingeman WE, & Jurat-Fuentes JL (2011) Identification, cloning, and expression of a GHF9 cellulase from *Tribolium castaneum* (Coleoptera: Tenebrionidae). *J Insect Physiol* 57(2):300-306.
89. Chang WH & Lai AG (2018) Mixed evolutionary origins of endogenous biomass-depolymerizing enzymes in animals. *BMC Genomics* 19(1):483.
90. Danchin EG, *et al.* (2010) Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proc Natl Acad Sci USA* 107(41):17651-17656.
91. Pauchet Y, Kirsch R, Giraud S, Vogel H, & Heckel DG (2014) Identification and characterization of plant cell wall degrading enzymes from three glycoside hydrolase families in the cerambycid beetle *Apriona japonica*. *Insect Biochem Mol Biol* 49:1-13.

92. Davison A & Blaxter M (2005) Ancient origin of glycosyl hydrolase family 9 cellulase genes. *Mol Biol Evol* 22(5):1273-1284.
93. Kirsch R*, et al.* (2014) Horizontal gene transfer and functional diversification of plant cell wall degrading polygalacturonases: Key events in the evolution of herbivory in beetles. *Insect Biochem Mol Biol* 52:33-50.
94. Palomares-Rius JE*, et al.* (2014) Distribution and evolution of glycoside hydrolase family 45 cellulases in nematodes and fungi. *Bmc Evol Biol* 14:69.
95. Scully ED, Hoover K, Carlson JE, Tien M, & Geib SM (2013) Midgut transcriptome profiling of *Anoplophora glabripennis*, a lignocellulose degrading cerambycid beetle. *BMC Genomics* 14:850.
96. Vasilikopoulos A*, et al.* (2019) Phylogenomics of the superfamily Dytiscoidea (Coleoptera: Adephaga) with an evaluation of phylogenetic conflict and systematic error. *Mol Phylogenet Evol* 135:270-285.
97. Sharkey CR*, et al.* (2017) Overcoming the loss of blue sensitivity through opsin duplication in the largest animal group, beetles. *Sci Rep* 7(1):8.
98. Pauli T*, et al.* (2016) Transcriptomic data from panarthropods shed new light on the evolution of insulator binding proteins in insects : Insect insulator proteins. *BMC Genomics* 17(1):861.
99. Seppey M*, et al.* (2019) Genomic signatures accompanying the dietary shift to phytophagy in polyphagan beetles. *Genome Biol* 20(1):98.
100. Santos MFA, Mermudes JRM, & Fonseca VMM (2011) A specimen of Curculioninae (Curculionidae, Coleoptera) from the Lower Cretacecous, Araripe Basin, north-eastern Brazil. *Palaeontology* 54:807-814.
101. Oberprieler RG & Oberprieler SK (2012) *Talbragarus averyi* gen. et sp. n., the first Jurassic weevil from the southern hemisphere (Coleoptera: Curculionoidea: Nemonychidae). *Zootaxa* 3478:256-266.
102. Arnoldi LV (1977) Eobelidae. *Mesozoic Coleoptera*, eds Arnoldi L, Zherikhin V, Nikitrin L, & Ponomarenko A (Nauka Publishers, Moscow), pp 144-176.
103. Wang B*, et al.* (2014) The earliest known longhorn beetle (Cerambycidae: Prioninae) and implications for the early evolution of Chrysomeloidea. *J Syst Palaeontol* 12:565-574.
104. Cai CY, Slipinski A, & Huang DY (2015) The oldest root-eating beetle from the Middle Jurassic of China (Coleoptera, Monotomidae). *Alcheringa* 39:488-493.
105. Chang SC, Gao KQ, & Zhou CF (2017) New chronostratigraphic constraints on the Yixian Formation with implications for the Jehol Biota. *Palaeogeogr Palaeoclimatol Palaeoecol* 487(1):399-406.
106. Wang B & Zhang HC (2011) The Oldest Tenebrionoidea (Coleoptera) from the Middle Jurassic of China. *J Paleontol* 85(2):266-270.
107. Kirejtshuk AG, Nabozhenko MV, & Nel A (2012) First Mesozoic representative of the subfamily Tenebrioninae (Coleoptera, Tenebrionidae) from the Lower Cretaceous of Yixian (China, Liaoning). *Entomol Rev* 92:97-100.
108. Liu ZH, Slipinski A, Leschen RAB, Ren D, & Pang. H (2015) The oldest Prionoceridae (Coleoptera: Cleroidea) from the Middle Jurassic of China. *Annales Zoologici* 65:41-52.
109. Fikacek M*, et al.* (2014) Modern hydrophilid clades present and widespread in the Late Jurassic and Early Cretaceous (Coleoptera: Hydrophiloidea: Hydrophilidae). *Zool J Linnean Soc* 170:710-734.

110. Cai C, Huang D, Thayer MK, & Newton AF (2012) Glypholomatine Rove Beetles (Coleoptera: Staphylinidae): a Southern Hemisphere Recent Group Recorded from the Middle Jurassic of China. *J Kansas Entomol Soc* 85(3):239-244.
111. Nikolajev GV (2010) On the Mesozoic taxa of scarabaeoid beetles of the Family Hybosoridae (Coleoptera: Scarabaeoidea). *Paleontol J* 44(6):649-653.
112. Cai C, Lawrence JF, Ślipiński A, & Huang D (2014) First fossil tooth-necked fungus beetle (Coleoptera: Derodontidae): *Juropeltastica sinica* gen. n. sp. n. from the Middle Jurassic of China. *Eur J Entomol* 111(2):299-302.
113. Jin ZY, Slipinski A, Pang H, & Ren D (2013) A new Mesozoic species of soft-bodied plant beetle (Coleoptera: Dascillidae) from the Early Cretaceous of Inner Mongolia, China with a review of fossil Dascillidae. *Annales Zoologici* 63:501-509.
114. Chang HL, Zhao YY, & Ren D (2009) New fossil elaterids (Insect: Coleoptera: Polyphaga: Elateridae) from the Middle Jurassic of Inner Mongolia, China. *Prog Nat Sci* 19:1433-1437.
115. Peris D, C. , Maier A, & Sánchez-García A (2015) Taxonomic names, in The oldest known riffle beetle (Coleoptera: Elmidae) from Early Cretaceous Spanish amber. *Comptes Rendus Palevol* 14(181-186).
116. Yan EV, Beutel RG, & Lawrence JF (2018) Whirling in the late Permian: ancestral Gyrinidae show early radiation of beetles before Permian-Triassic mass extinction. *BMC Evol Biol* 18(33):1-10.
117. Ponomarenko AG (1977) Suborder Adephaga, Polyphaga Incertae Sedis, Infraorder Staphyliniformia, in Mezozoiskie zhestkokryiye [Mesozoic Coleoptera]. Akademiya Nauk SSSR. *Trudy Paleontol Inst* 161:17-119.
118. Ponomarenko AG (1969) Istoricheskoe Razvitie Zhestkokrylykh-Arkhostemat [Historical Development of the Archostomate Beetles]. *Trudy Akademiya Nauk SSSR* 125:1-240.
119. Henderickx H & Bosselaers J (2013) Taxonomic names, in X-ray micro-CT reconstruction reveals eight antennomeres in a new fossil taxon that constitutes a sister clade to Dundoxenos and Triozocera (Strepsiptera: Corioxenidae). *Palaeontol Electron* 16(3.29A):1-16.
120. Haug JT, Labandeira CC, Santiago-Blay JA, Haug C, & Brown S (2016) Erratum to: Life habits, hox genes, and affinities of a 311 million-year-old holometabolan larva. *BMC Evol Biol* 16(169):1-6.
121. Kirejtshuk AG, Poschmann M, & Nel A (2014) Taxonomic names, in Evolution of the elytral venation and structural adaptations in the oldest Palaeozoic beetles (Insecta: Coleoptera: Tshekardocoleidae). *J Syst Palaeontol* 12:575-600.
122. Tillyard RJ (1922) Some new Permian insects from Belmont, NSW, in the collection of Mr. John Mitchell. *P Linn Soc NSW* 47:279-292.
123. Pan XX, Chang HL, & Ren D (2011) Taxonomic names, in The first fossil buprestids from the Middle Jurassic Jiulongshan Formation of China (Coleoptera: Buprestidae). *Zootaxa* 2745:53-62.
124. Kirejtshuk AG & Azar D (2008) New taxa of beetles (Insecta, Coleoptera) from Lebanese amber with evolutionary and systematic comments. *Alavesia* 2:15-46.
125. Lawrence JF*, et al.* (2011) Phylogeny of the Coleoptera Based on Morphological Characters of Adults and Larvae. *Annales Zoologici* 61(1):1-217.
126. Gunter NL, Oberprieler RG, & Cameron SL (2016) Molecular phylogenetics of Australian weevils (Coleoptera: Curculionoidea): exploring relationships in a

hyperdiverse lineage through comparison of independent analyses. *Aust Entomol* 55:217-233.
127. Königer S, Lorenz V, Stollhofen H, & Armstrong RA (2002) Origin, age and stratigraphic significance of distal fallout ash tuffs from the Carboniferous-Permian continental Saar-Nahe Basin (SW Germany). *Int J Earth Sci* 91:341-356.
128. Martill D (2007) The age of the Cretaceous Santana Formation fossil Konservat Lagerstätte of north-east Brazil: a historical review and an appraisal of the biochronostratigraphic utility of its palaeobiota. *Cretac Res* 28:895-920.
129. Bean LB (2006) The leptolepid fish Cavenderichthys talbragarensis (Woodward, 1895) from the Talbragar Fish Bed (Late Jurassic) near Gulgong, New South Wales. *Rec West Aust Mus* 23:43-76.
130. Turner S*, et al.* (2009) Australian Jurassic sedimentary and fossil successions: current work and future prospects for marine and non-marine correlation. *GFF* 131(1-2):49-70.
131. Ji Q*, et al.* (2004) *The Mesozoic Jehol Biota of western Liaoning and a synthetic study of related stratigraphic sequences (in Chinese)* (Geological Publishing House, Beijing).
132. Zhang HC, Wang B, & Fang Y (2010) Evolution of insects in diversity through the Jehol Biota. *Sci China Earth Sci* 53:1908-1917.
133. Huang DY, Nel A, Shen Y, Selden PA, & Lin Q (2006) Discussions on the age of the Daohugou fauna—evidence from invertebrates. *Pro Nat Sci* 16:308-312.
134. Liu YQ*, et al.* (2012) Timing of the earliest known feathered dinosaurs and transitional pterosaurs olderthan the Jehol Biota. *Palaeogeogr Palaeoclimatol Palaeoecol* 323-325:1-12.
135. Najarro M*, et al.* (2009) Unusual concentration of Early Albianarthropod bearing amber in the Basque-Cantabrian Basin (El Soplao,Cantabria, northern Spain): Palaeoenvironmental and palaeobiological implications. *Geol Acta* 7:363-387.
136. Knight O & Le M (1950) Fossil insect beds of Belmont, N.S.W. *Rec Aust Mus* 22(3):251-253.
137. Metcalf I, Crowley JL, Nicoll RS, & Schmitz M (2015) High-precision U-Pb CA-TIMS calibration of Middle Permian to Lower Triassic sequences, mass extinction and extreme climate-change in eastern Australian Gondwana. *Gondwana Res* 28(1):61-81.
138. Hind MC & Helby RJ (1969) VII The Great Artesian Basin in New South Wales. *J Geol Soc Aust* 16(1):481-497.
139. Yang W, Li S, & Jiang B (2007) New evidence for Cretaceous age of the feathered dinosaurs of Liaoning: zircon U-Pb SHRIMP dating of the Yixian Formation in Sihetun, northeast China. *Cretac Res* 28:177-182.
140. Chang SC, Zhang HC, Renne PR, & Fang Y (2009) High-precision 40Ar/39Ar age for the Jehol Biota. *Palaeogeogr Palaeoclimatol Palaeoecol* 280:94-104.
141. Aleksandrova GN & Zaporozhets NI (2008) Palynological Characteristics of Upper Cretaceous and Paleogene Deposits on the West of the Sambian Peninsula (Kaliningrad Region), Part 2. *Stratigr Geol Correl* 16(5):528-539.
142. Châteauneuf JJ & Gruas-Cavagnetto C (1978) (Les zones de Wetzeliellaceae (Dinophyceae) du basin de Paris. Comparaison et correlations aves les zones du Paleogene des basins du Nord-Ouest de l'Europe. Bull. BRGM (2-eme serie), Section IV, pp 59-93.
143. Powell AJ (1992) Dinoflagellate Cysts of the Tertiary System. *A Stratigraphic Index of Dinoflagellate Cysts*, ed Powell AJ (Chapman and Hall, London), pp 155-251.

144. Luterbacher HP*, et al.* (2004) The Paleogene Period. *A Geologic Time Scale 2004*, eds Gradstein FM, Ogg JG, & Smith AG (Cambridge Univ. Press, Cambridge), pp 384-408.
145. Cherchi A & Schroeder R (1999) Late Barremian orbitolinid foraminifera from northern Somalia. *B Soc Paleontol Ital* 38(1):3-13.
146. Schroeder R*, et al.* (2010) Revised orbitolinid biostratigraphic zonation for the Barremian - Aptian of the eastern Arabian Plate and implications for regional stratigraphic correlations. *GeoArabia Special Publication 4* v. 1:49-96.
147. Clavel B*, et al.* (2007) Dating and progradation of the Urgonian limestone from the Swiss Jura to southeast France. *Z Dtsch Ges Geowiss* 158(4):1025-1062.
148. Maksoud S*, et al.* (2016) Revision of "Falaise de BLANCHE" (Lower Cretaceous) in Lebanon, with the definition of a Jezzinian Regional Stage. *Carnets de Géologie [Notebooks on Geology]* 14(18):401-427.
149. Ballerio A*, et al.* (2014) An Enigma of Scarabaeidology Revealed: the Re-discovery of *Belohina inexpectata* in Southern Madagascar, and its placement in Scarabaeoidea. *Scarabs*. Issue 76.
150. Shockley FW, Tomaszewska KW, & McHugh JV (2009) An annotated checklist of the handsome fungus beetles of the world (Coleoptera: Cucujoidea: Endomychidae). *Zootaxa* 1999:1-113.
151. McKenna DD*, et al.* (2015) The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Syst Entomol* 40(4):835-880.
152. Fikáček M & Trávníček D (2009) Order Coleoptera, family Georissidae. *Arthropod Fauna of the UAE*, ed Van HA (Dar Al Ummah, Abu Dhabi), Vol 2, pp 145-148.
153. Fikáček M (2009) Order Coleoptera, family Helophoridae. *Arthropod fauna of the UAE* 2:142-144.
154. Lawrence JF & Ślipiński A (2013) *Globorentonium*, a new genus of rentoniine Trogossitidae (Coleoptera: Cleroidea) from Australia and Brazil. *Zootaxa* 3710(3):257-270.
155. Fikáček M (2011) Spercheidae Erichson 1837. Spercheus Kugelann 1798. Filterfeeding water scavenger beetles. http://tolweb.org/Spercheus/ in The Tree of Life Web Project, http://tolweb.org/.
156. Gunter NL*, et al.* (2013) A molecular phylogeny of the checkered beetles and a description of Epiclininae a new subfamily (Coleoptera: Cleroidea: Cleridae). *Syst Entomol* 38:626-636.
157. Robertson JA*, et al.* (2015) Phylogeny and classification of Cucujoidea and the recognition of a new superfamily Coccinelloidea (Coleoptera: Cucujiformia). *Syst Entomol* 40:745-778.
158. Baca SM, Alexander A, Gustafson GT, & Short AE (2017) Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of 'Hydradephaga'. *Syst Entomol* 42:786-795.
159. Ślipiński SA, Leschen RAB, & Lawrence JF (2011) Order Coleoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa* 3148:203-208.
160. Haddad S*, et al.* (2018) Anchored hybrid enrichment provides new insights into the phylogeny and evolution of longhorned beetles (Cerambycidae). *Syst Entomol* 43(1):68-89.

161. Shin S*, et al.* (2018) Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. *Mol Biol Evol* 35(4):823-836.

Fig. S1. Best tree resulting from maximum likelihood (ML) analysis of the partitioned amino acid supermatrix (147 taxa; 4852 genes, 10 replicate ML searches), including the taxa from Fig. S2 (the 4818 gene tree) + Rhinorhipidae: *Rhinorhipus*.

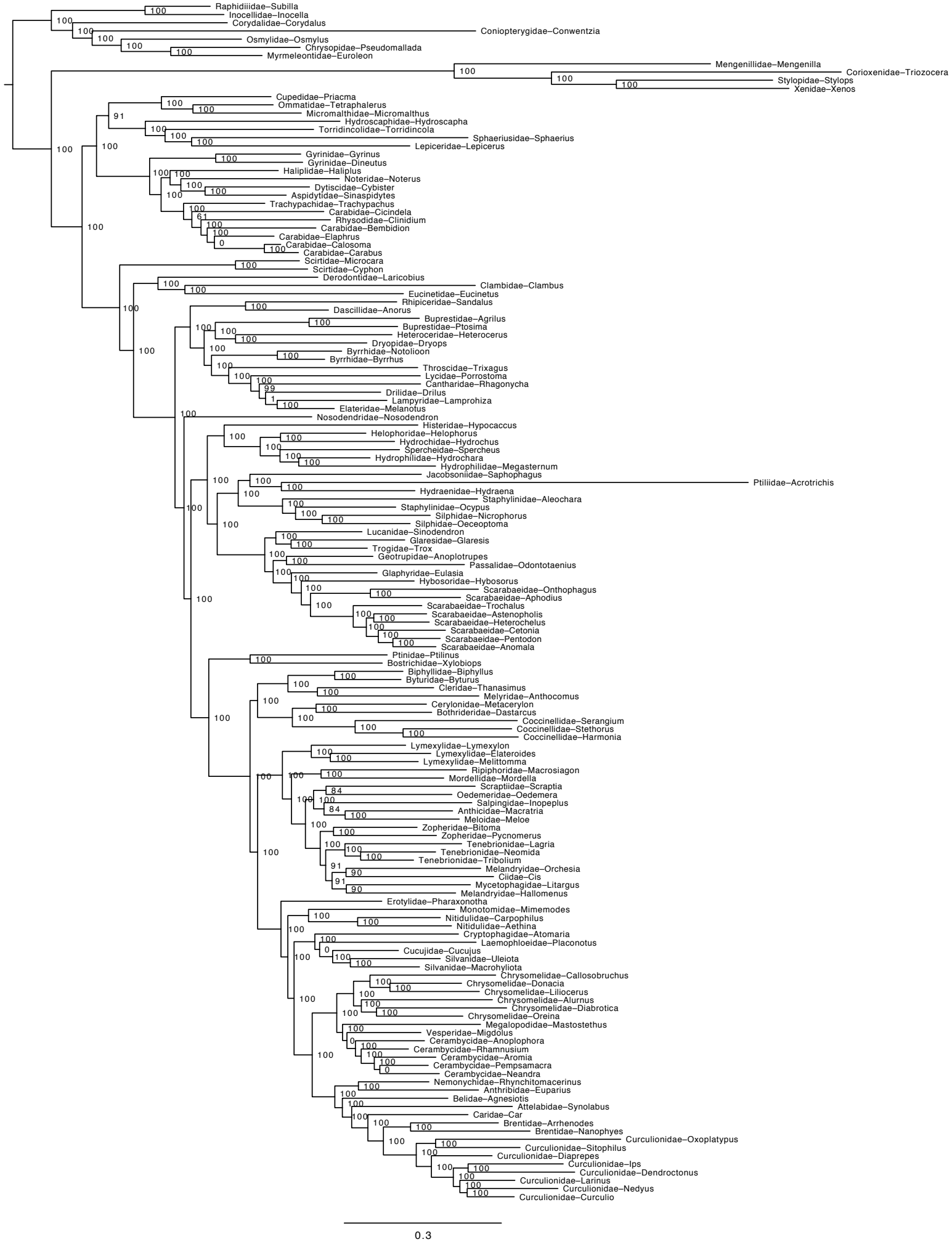


Fig. S2. Best tree resulting from maximum likelihood (ML) analysis of partitioned amino acid supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

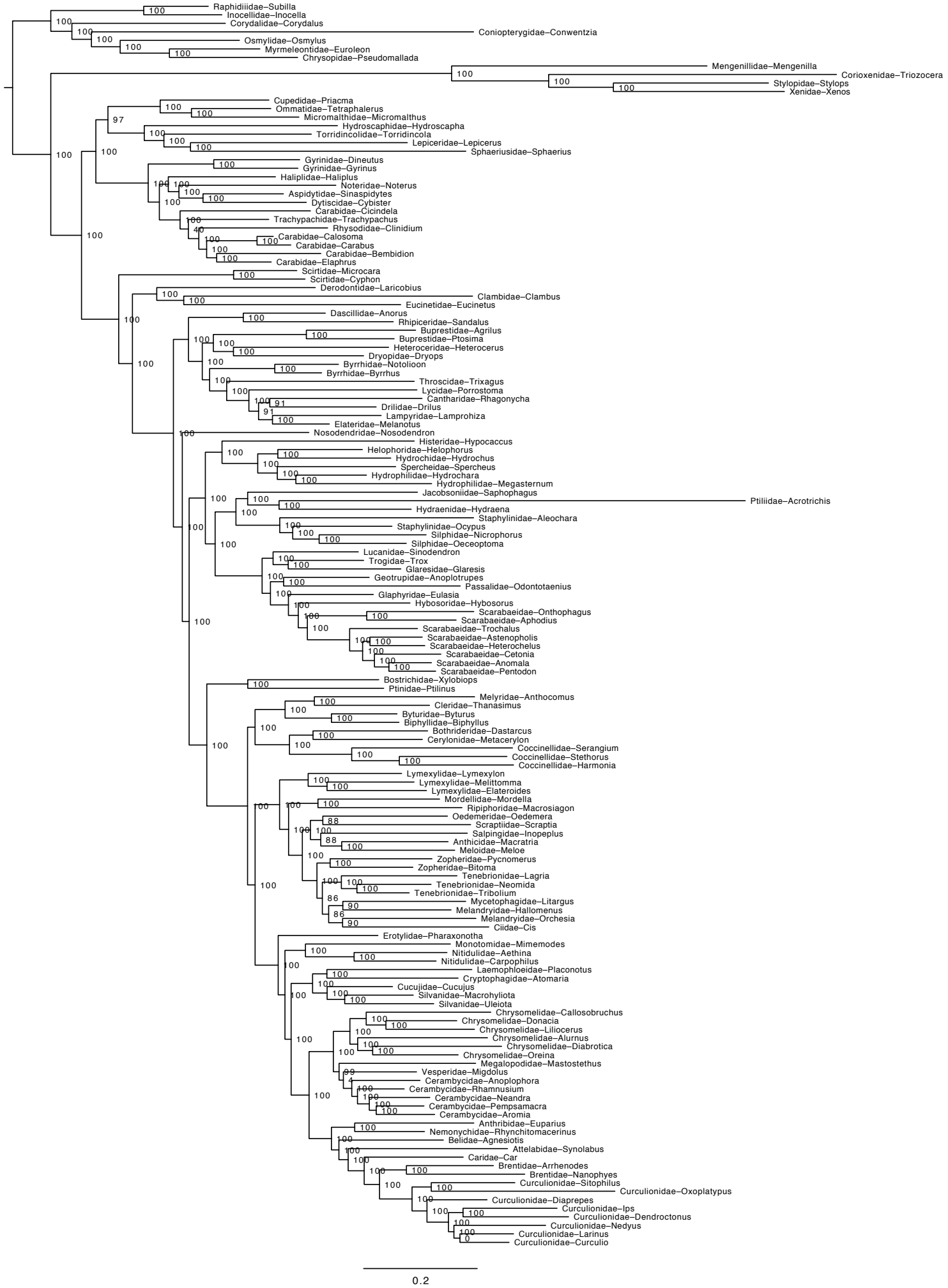


Fig. S3. Best tree resulting from maximum likelihood (ML) analysis of un-partitioned amino acid supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

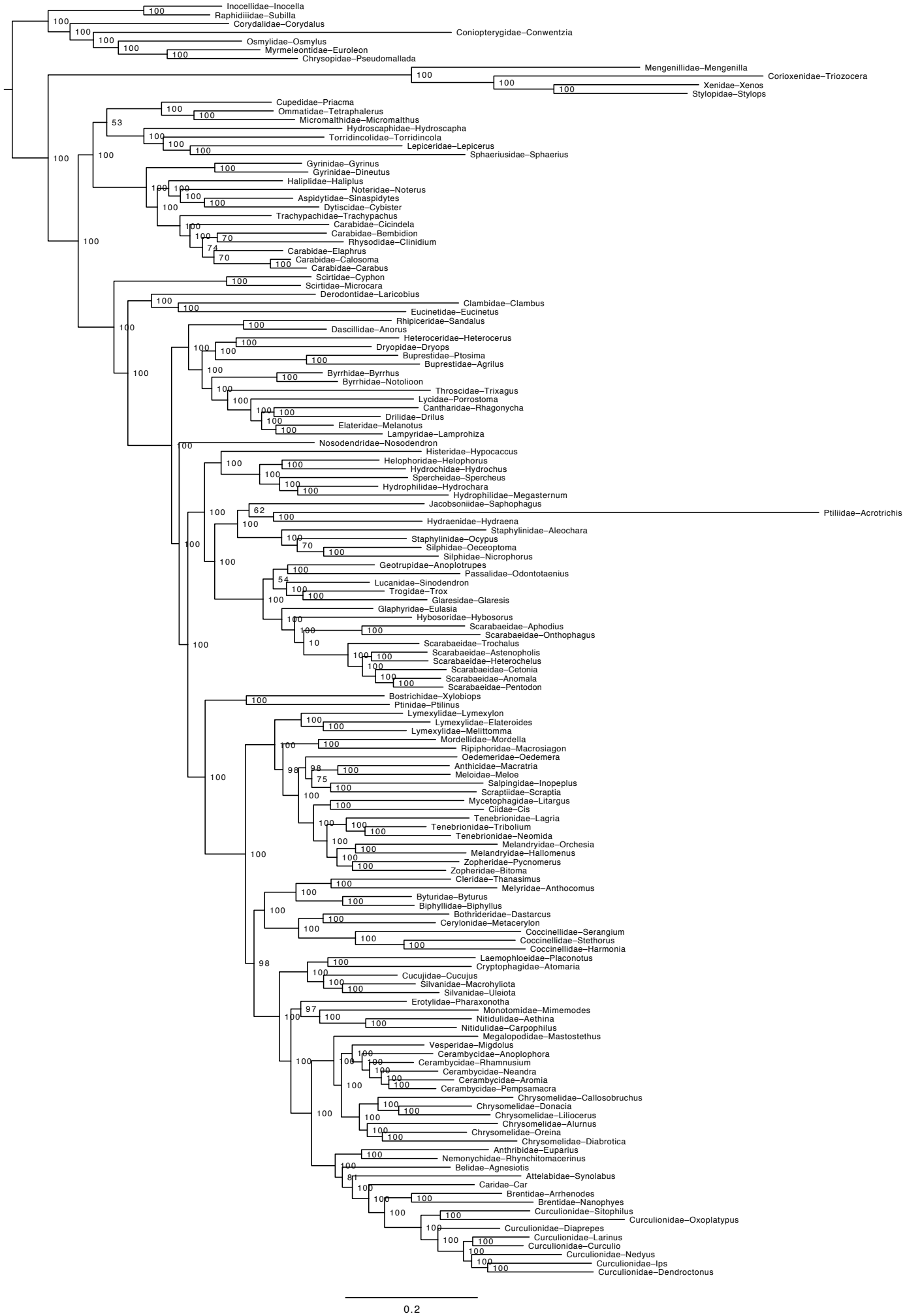


Fig. S4. Best tree resulting from maximum likelihood (ML) analysis of partitioned nucleotide codon 1 (C1) supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

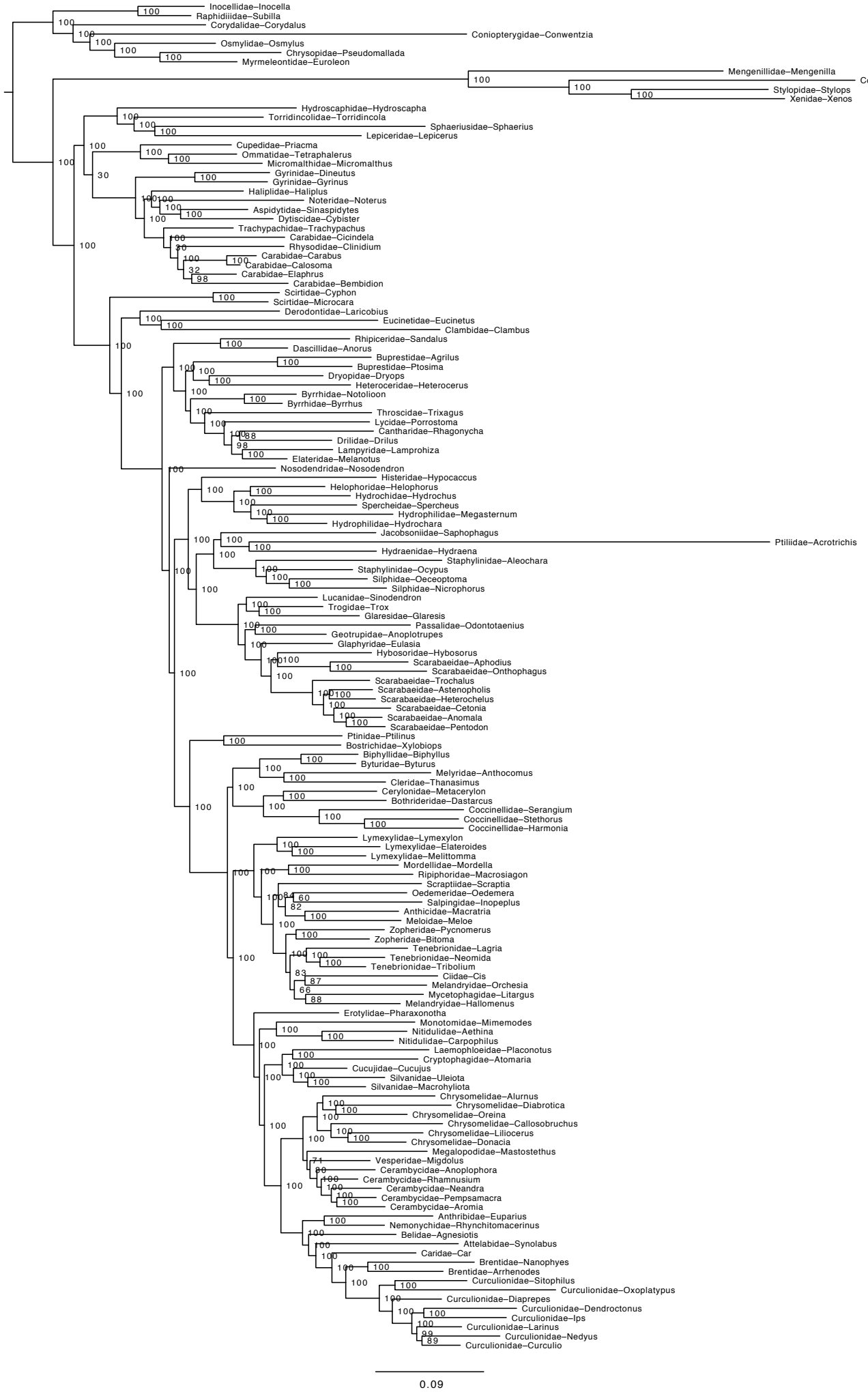


Fig. S5. Best tree resulting from maximum likelihood (ML) analysis of partitioned nucleotide codon 2 (C2) supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

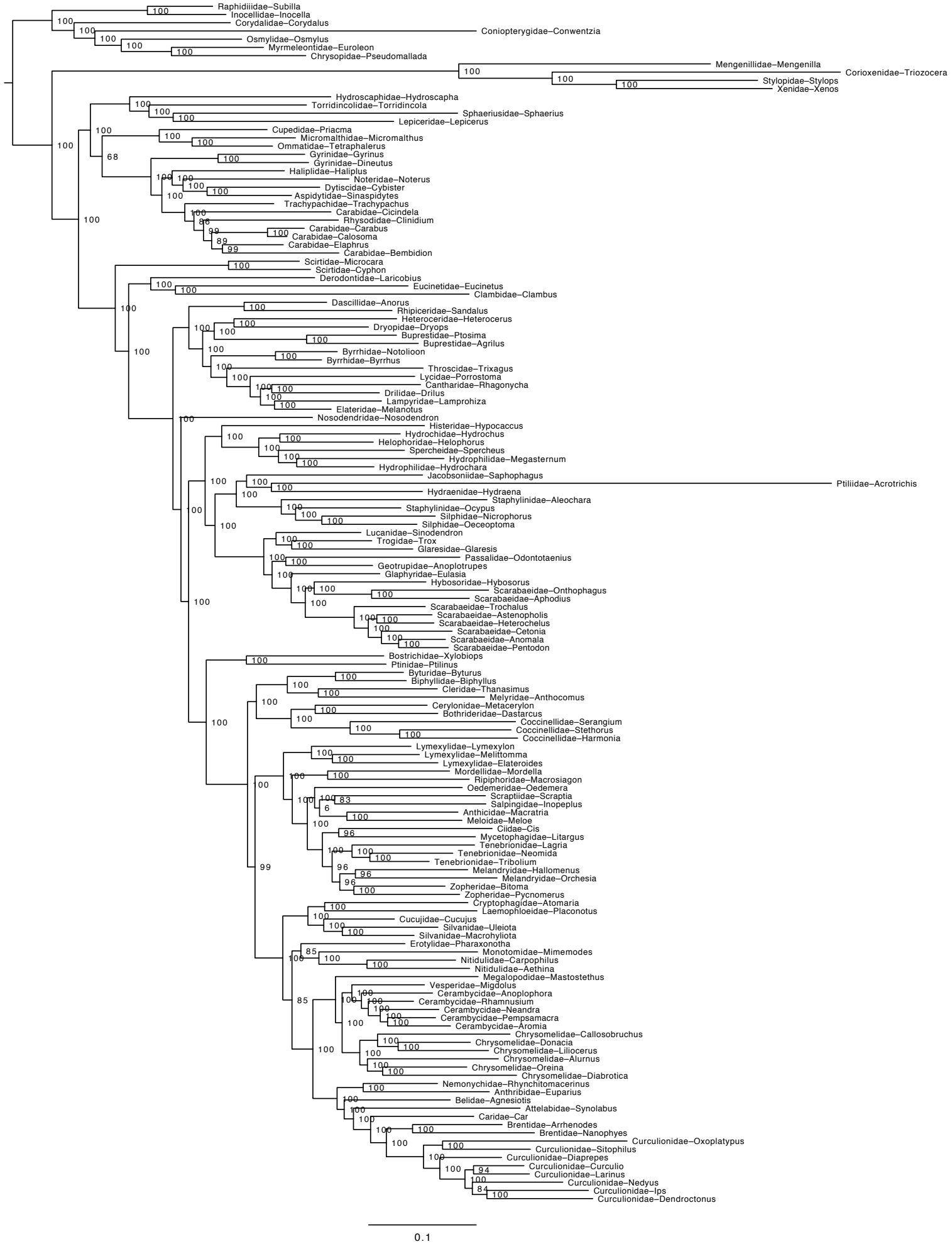


Fig. S6. Best tree resulting from maximum likelihood (ML) analysis of partitioned nucleotide codon 1 and 2 (C12) supermatrix (146 taxa, 4818 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node.

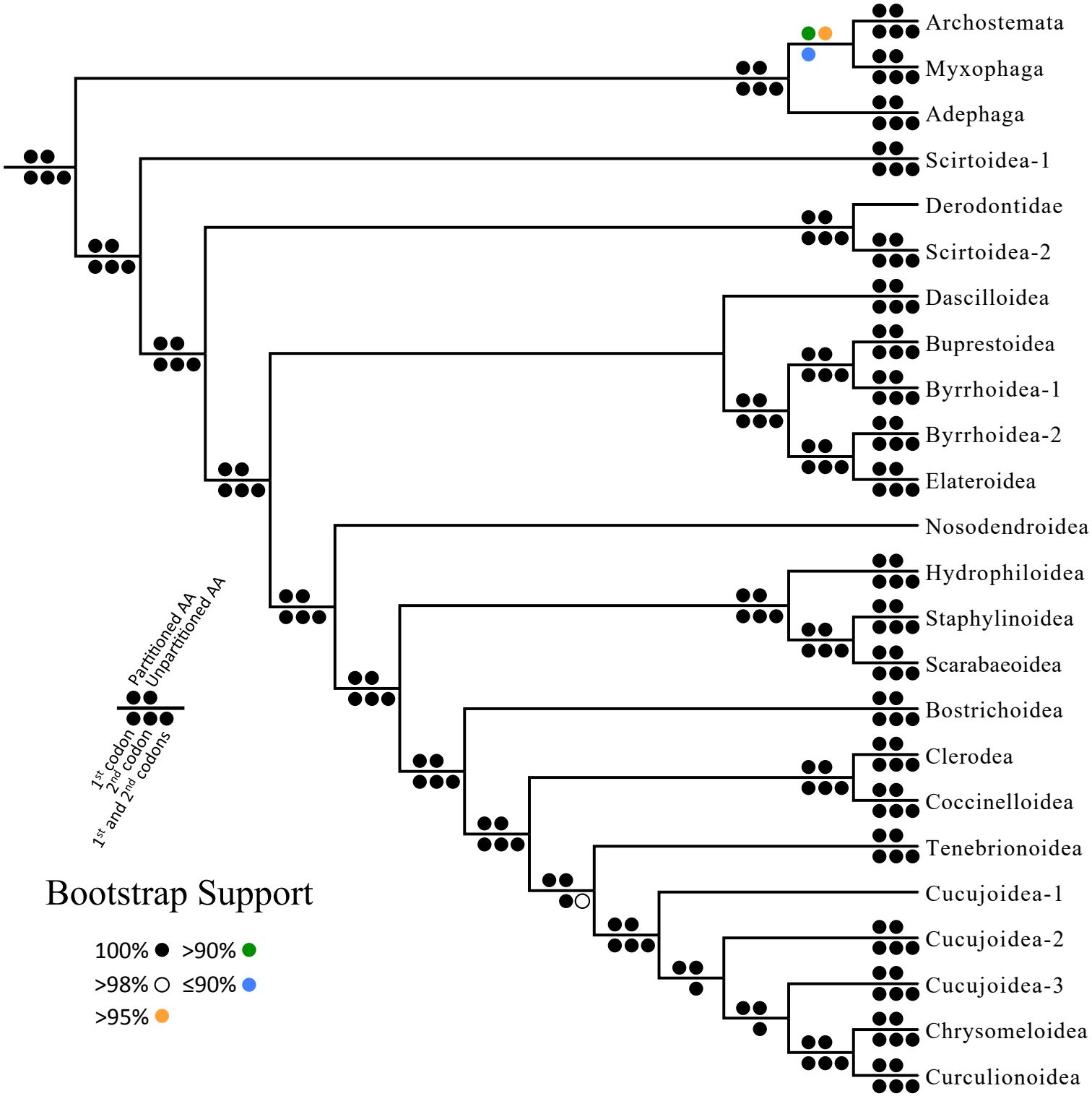


Fig. S7. Summary tree showing superfamily-level clades and summarizing statistical measures of nodal support from the six above-described ML analyses (Figs S2-6).

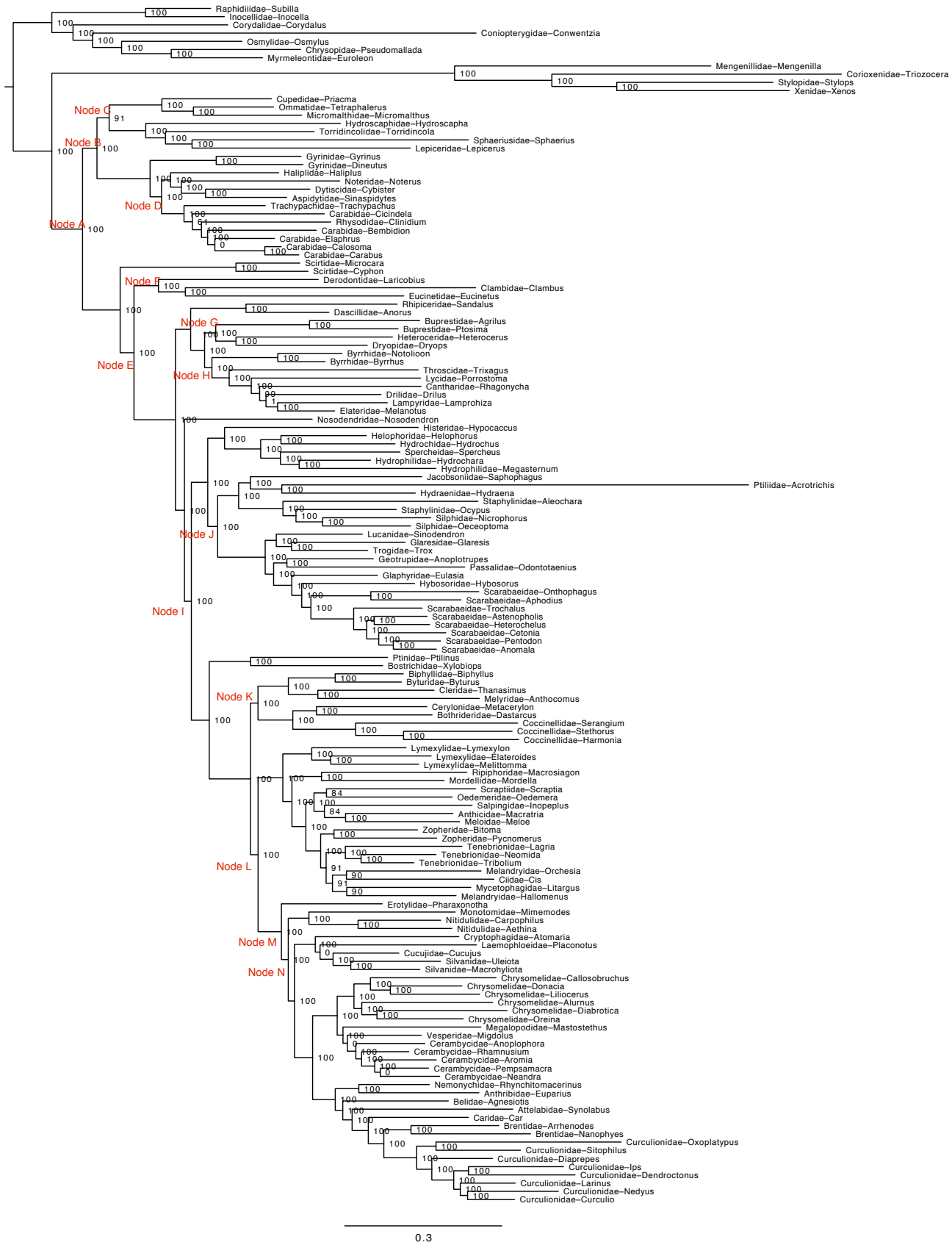
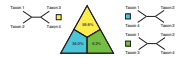


Fig. S8. Testing of alternative phylogenetic hypotheses using four-cluster likelihood mapping (FcLM) was undertaken for nodes representing interrelationships among higher-level taxa in the 4818-gene ML tree (Fig. S2) that had less than maximal statistical support (from ML bootstrapping) and/or recovering unexpected relationships. A. monophyly of Coleoptera; B. (Archostemata + Myxophaga + Adephaga), sister to Polyphaga; C. (Archostemata + Myxophaga), sister to Adephaga; D. Gyrinidae sister to the remaining Adephaga; E. Scirtidae sister to the remaining Polyphaga; F. monophyly of (Derodontidae + Clambidae + Eucinetidae), or polyphyly of Scirtoidea; G. monophyly of (Buprestidae + Dryopidae + Heteroceridae); H. monophyly of (Elateroidea +*Notolioon* [Byrrhidae] + *Byrrhus* [Byrrhidae]); I. monophyly of Cucujiformia; J. monophyly of (Scarabaeoidea + Staphylinoidea); K. monophyly of (Cleroidea + Coccinelloidea); L. (Lymexyloidea + Elateroidea + Tenebrionoidea) sister to (Phytophaga + Cucujoidea); M. Erotylidae sister to (Phytophaga + remaining Cucujoidea); N. (Laemophloeidae + Monotomidae + Nitidulidae) sister to (Phytophaga + remaining Cucujoidea). For each analysis, values at the corners indicate the percentage of fully resolved phylogenies for all possible quartets.

A



Hypothesis 1: Monophyly of Coleoptera

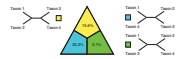
Taxon 1:
Adephaga/Archostemata/Myxophaga

Taxon 2:
Polyphaga

Taxon 3:
Strepsiptera

Taxon 4:
Neuropteroida

B



Hypothesis 2: (Archostemata + Myxophaga + Adephaga), sister to Polyphaga

Taxon 1:
Archostemata/Myxophaga

Taxon 2:
Adephaga

Taxon 3:
Polyphaga

Taxon 4:
Strepsiptera

C

Hypothesis 3: (Archostemata + Myxophaga), sister to Adephaga

Taxon 1:
Archostemata

Taxon 2:
Myxophaga

Taxon 3:
Adephaga

Taxon 4:
Polyphaga

D

Hypothesis 4: Gyrinidae sister to the remaining Adephaga

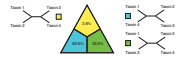
Taxon 1:
Carabidae/Rhysodidae/Trachypachidae

Taxon 2:
Dytiscidae/Haliplidae/Noteridae/Aspidytidae

Taxon 3:
Gyrinidae

Taxon 4:
Archostemata/Myxophaga

E



Hypothesis 5: Scirtidae sister to the remaining Polyphaga

Taxon 1:
Derodontoidea/Clambidae/Eucinetidae

Taxon 2:
Polyphaga

Taxon 3:
Scirtidae

Taxon 4:
Adephaga/Archostemata/Myxophaga

F

Hypothesis 6: Monophyly of (Derodontoidea + Clambidae + Eucinetidae)

Taxon 1:
Derodontoidea

Taxon 2:
Clambidae/Eucinetidae

Taxon 3:
Scirtidae

Taxon 4:
Polyphaga

G

Hypothesis 7: Monophyly of (Buprestidae + Heteroceridae + Dryopidae)

Taxon 1:
Buprestidae

Taxon 2:
Heteroceridae/Dryopidae

Taxon 3:
Byrrhidae/Elateroidea

Taxon 4:
Dascilloidea

H

Hypothesis 8: Monophyly of (Elateroidea + Notoliooon[Byrrhidae] + Byrrhus[Byrrhidae])

Taxon 1:
Byrrhidae

Taxon 2:
Elateroidea

Taxon 3:
Buprestidae/Heteroceridae/Dryopidae

Taxon 4:
Dascilloidea

I

Hypothesis 9: Monophyly of Cucujiformia

Taxon 1:
Scarabaeoidea/Staphylinoidea/Histeroidea/ Hydrophiloidea

Taxon 2:
Lymexyloidea/Melandryidae/Tenebrionoidea/Cucujoidea/Chrysomeloidea/Curculionoidea/ Coccinelloidea/ Bostrichoidea

Taxon 3:
Nosodendridae

Taxon 4:
Dascilloidea/Buprestoidea/Byrrhoidea/ Elateroidea

J

Hypothesis 10: Monophyly of (Scarabaeoidea + Staphylinoidea)

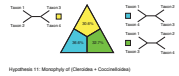
Taxon 1:
Scarabaeoidea

Taxon 2:
Staphylinoidea

Taxon 3:
Histeroidea/Hydrophiloidea

Taxon 4:
Lymexyloidea/Melandryidae/Tenebrionoidea/Cucujoidea/Chrysomeloidea/Curculionoidea/ Coccinelloidea/ Bostrichoidea

K



Hypothesis 11: Monophyly of (Cleroidea + Coccinelloidea)

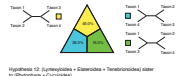
Taxon 1:
Coccinelloidea

Taxon 2:
Cleroidea

Taxon 3:
Bostrichoidea

Taxon 4:
Lymexyloidea/Melandryidae/Tenebrionoidea/Cucujoidea/Chrysomeloidea/Curculionoidea

L



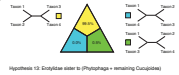
Hypothesis 12: (Lymexyloidea + Elateroidea + Tenebrionoidea) sister to (Phytophaga + Cucujoidea)
Taxon 1:
Lymexyloidea/Melandryidae/Tenebrionoidea

Taxon 2:
Cucujoidea/Chrysomeloidea/Curculionoidea

Taxon 3:
Coccinelloidea/Cleroidea

Taxon 4:
Bostrichoidea

M



Hypothesis 13: Erotylidae sister to (Phytophaga + remaining Cucujoidea)

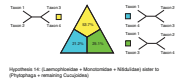
Taxon 1:
Erotylidae

Taxon 2:
"Cucujoidea" /Chrysomeloidea/Curculionoidea

Taxon 3:
Lymexyloidea/Melandryidae/ Tenebrionoidea

Taxon 4:
Coccinelloidea/Cleroidea

N



Hypothesis 14: (Laemophloeidae + Monotomidae + Nitidulidae) sister to (Phytophaga + remaining Cucujoidea)

Taxon 1:
Laemophloeidae/Monotomidae/Nitidulidae

Taxon 2:
"Cucujoidea" /Chrysomeloidea/Curculionoidea

Taxon 3:
Erotylidae

Taxon 4:
Lymexyloidea/Melandryidae/Tenebrionoidea

Fig. S5. Results from IQ-M analyses for 14 hypotheses based on the nodes indicated in Fig. S8.

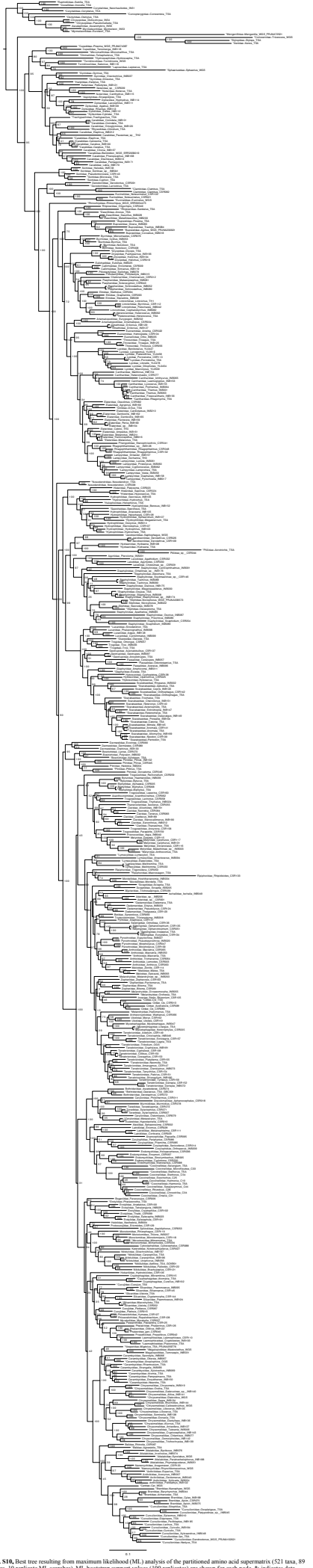


Fig. S10, Best tree resulting from maximum likelihood (ML) analysis of the partitioned amino acid supermatrix (521 taxa, 89 genes, 10 replicate ML searches). ML bootstrap support values (100 replicates) are shown for each node. *: indicates data generated for the present study.

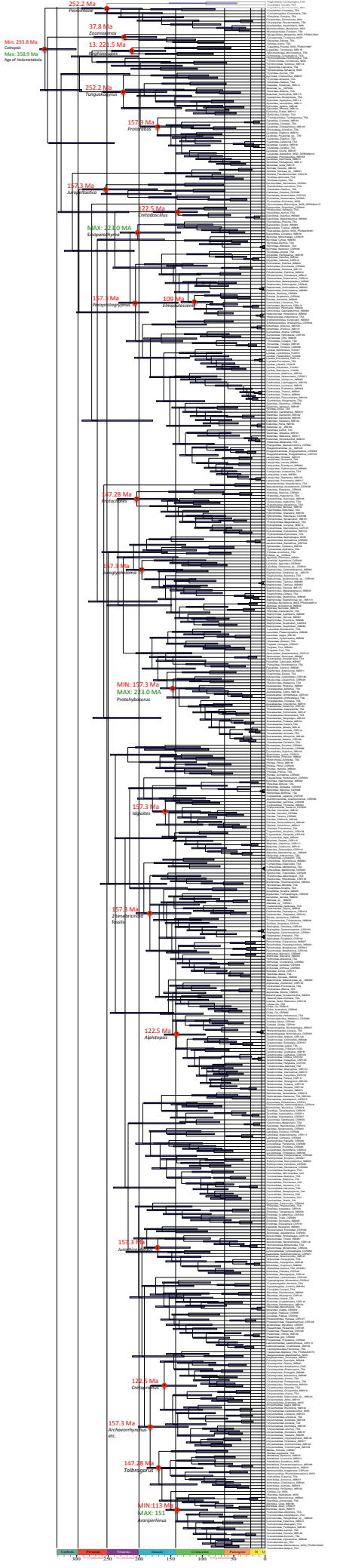


Fig. S11, Chronogram inferred from the 89-gene dataset showing the calibration points used in the time divergence analysis. The divergence time analysis was based on a maximum likelihood (ML) analysis of 521 taxa and 89 genes (Fig. S 10), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 19,951 amino acid sites. Nodes constrained by fossil priors are indicated with numbers (see Supplementary Table 5). The calibration points and fossil ages are shown in red. *: indicates data generated for the present study.

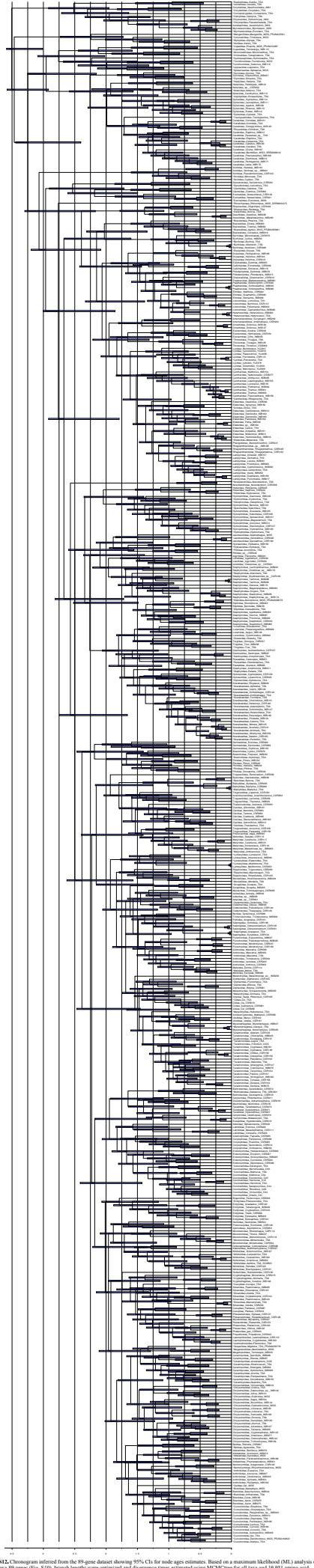


Fig. S12. Chronogram inferred from the 89-gene dataset showing 95% CIs for node ages estimates. Based on a maximum likelihood (ML) analysis of 521 taxa 89 genes (Fig. S10), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 19,951 amino acid sites. *: indicates data generated for the present study. Node ages refer to hundreds of millions of years.

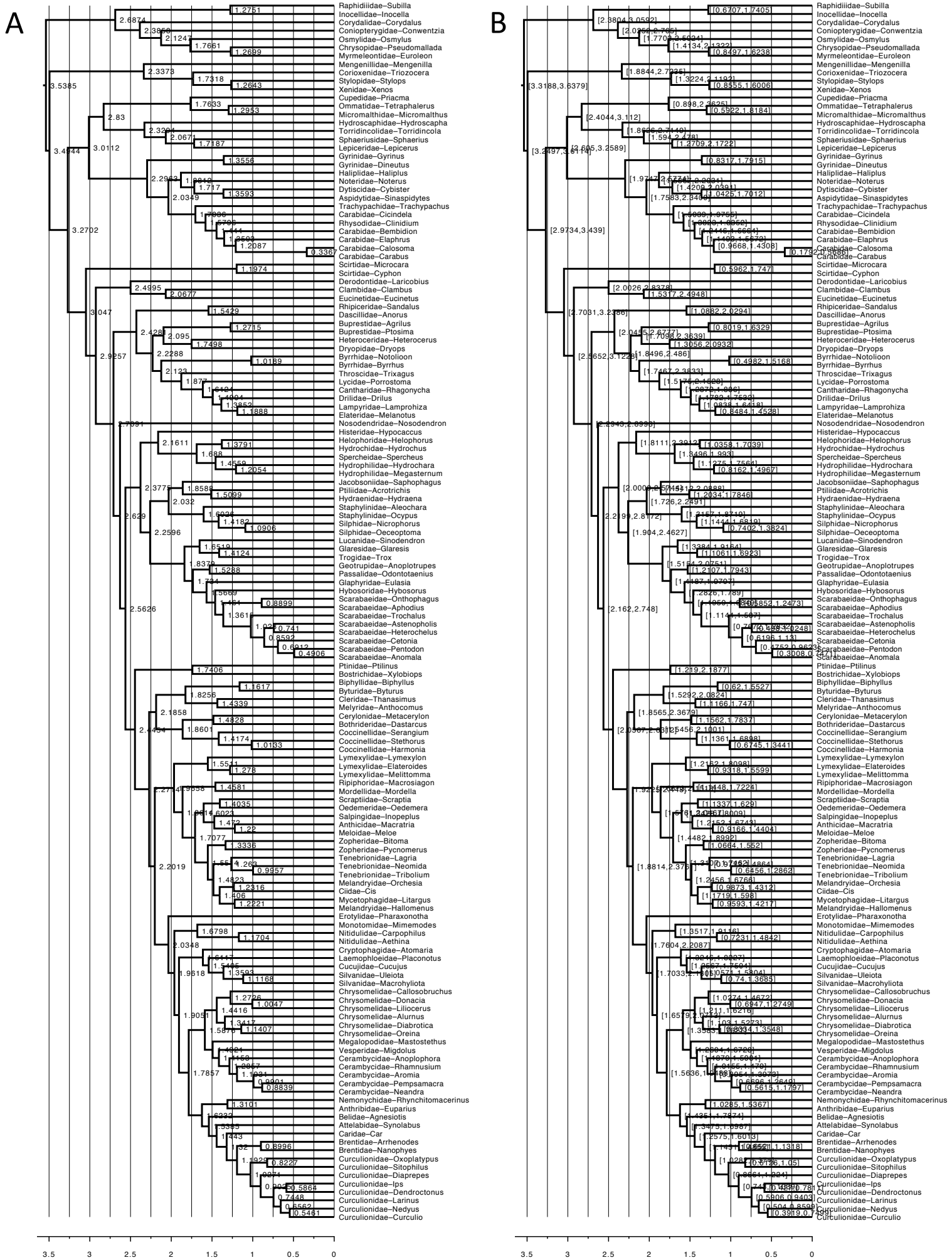


Fig. S13. Chronogram inferred from the 4818-gene dataset showing (A) mean values and (B) 95% CIs values for node ages estimates. Based on a maximum likelihood (ML) analysis of 146 taxa 4818 genes (Fig. S2), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 206,156 amino acid sites. Node ages refer to hundreds of millions of years.

Fig. S14, Chronogram inferred from the 89-gene dataset showing (A) mean values and (B) 95% CIs for node ages estimates. Based on a maximum likelihood (ML) analysis of 521 taxa 89 genes (Fig. S 10), branch lengths were optimized and divergence times estimated using MCMCtree for all taxa and 19,951 amino acid sites. *: indicates data generated for the present study. Node ages refer to hundreds of millions of years.

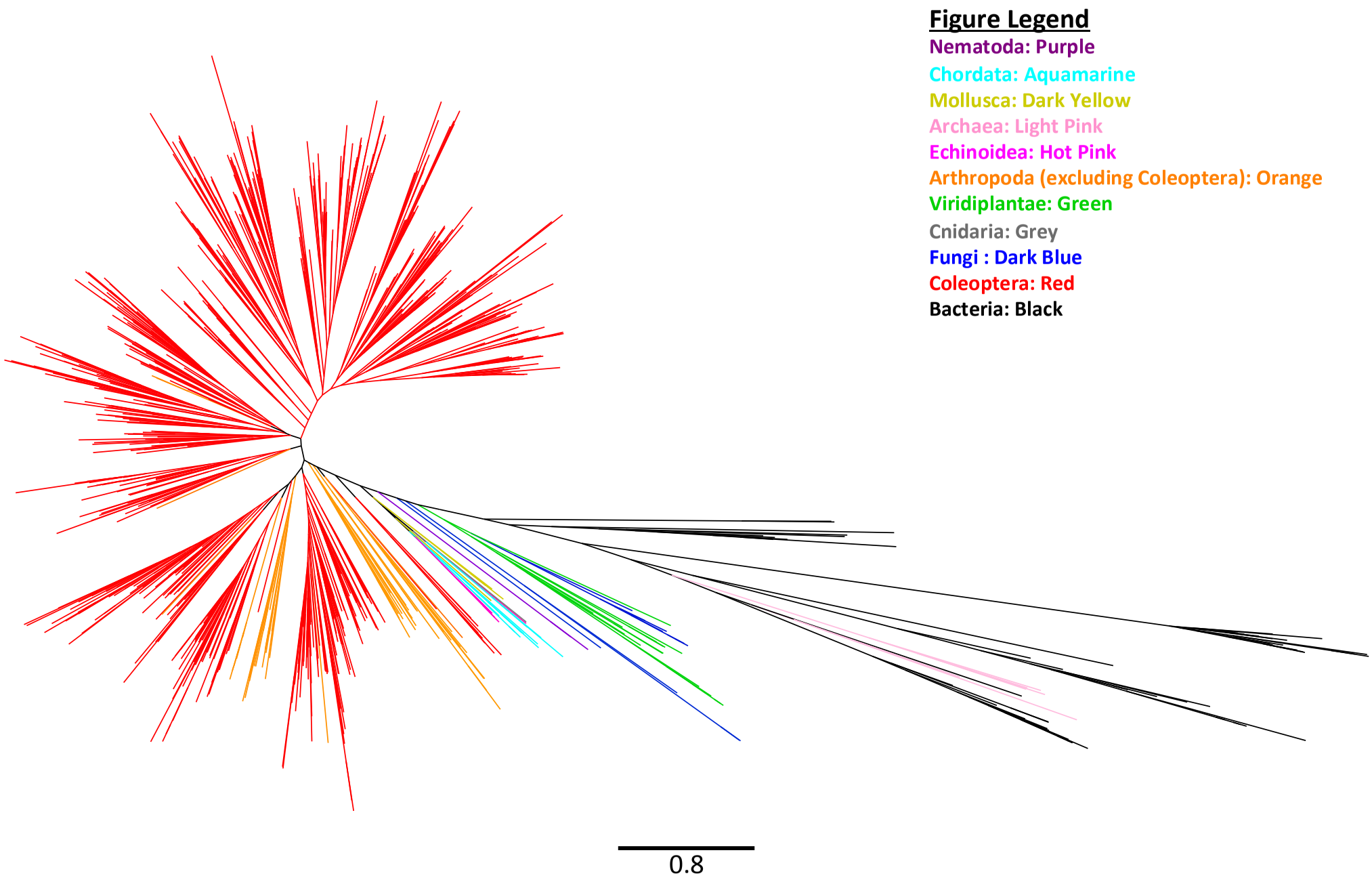


Fig. S15. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 1 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

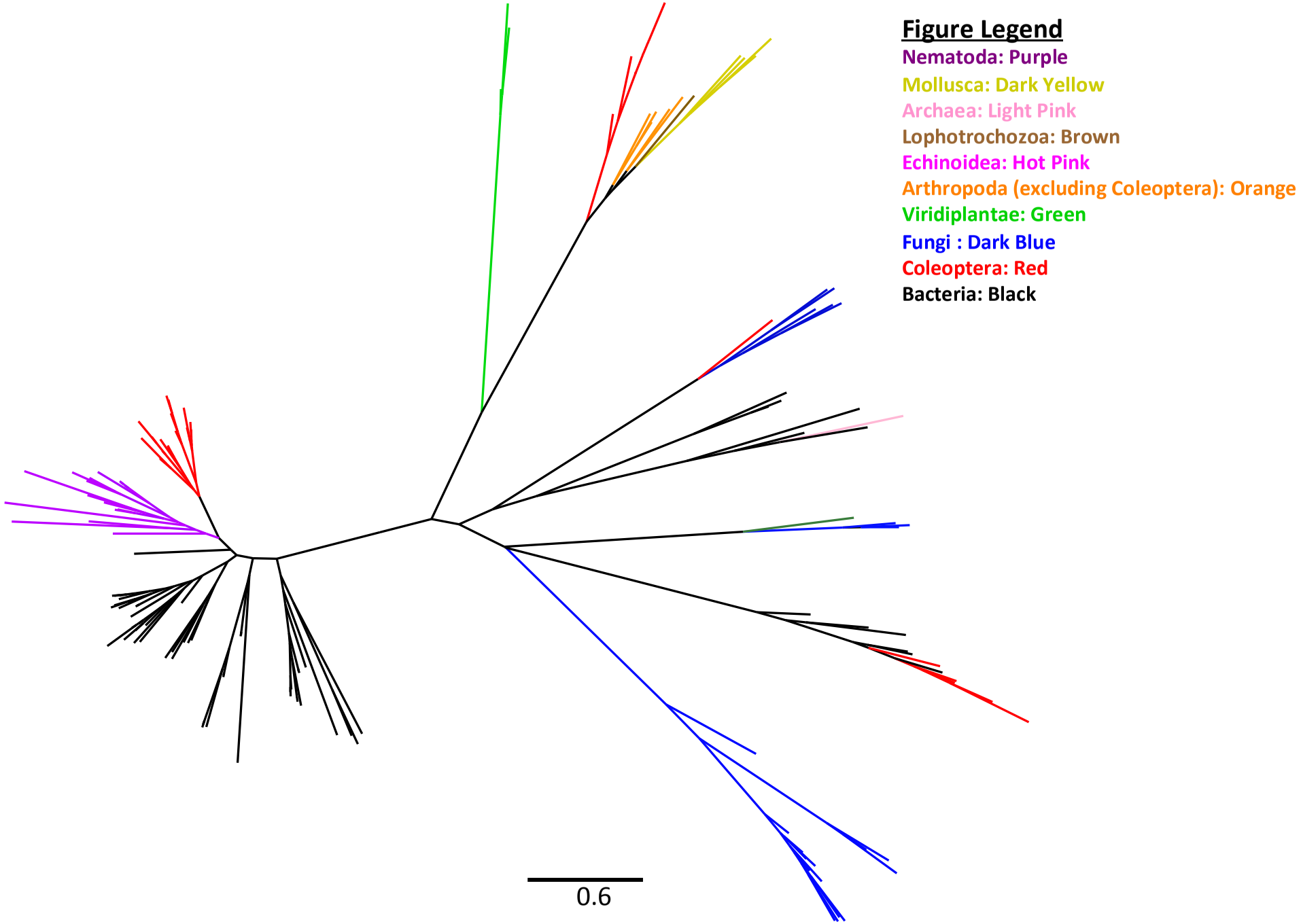


Fig. S16. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 5 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

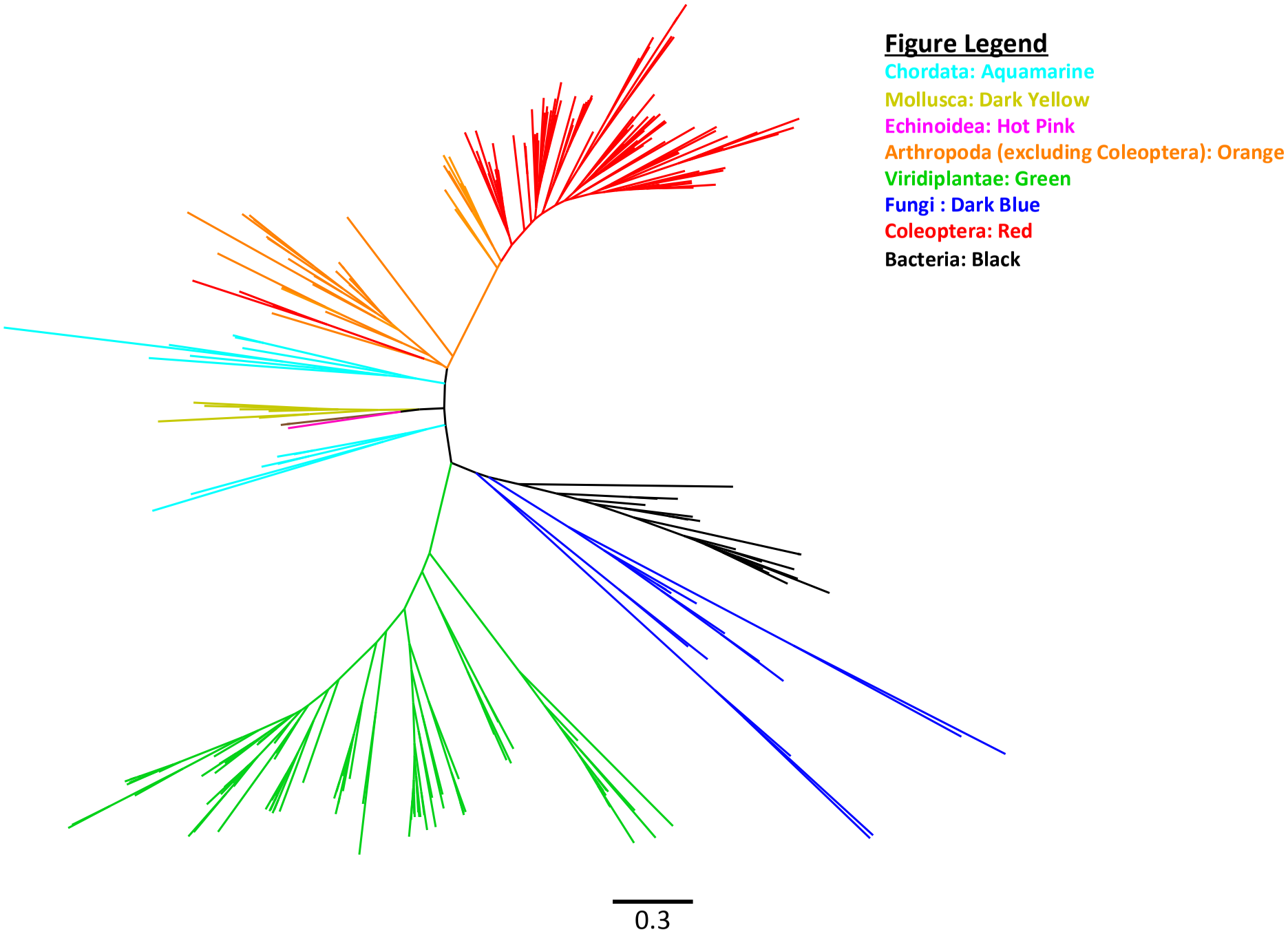


Fig. S17. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 9 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

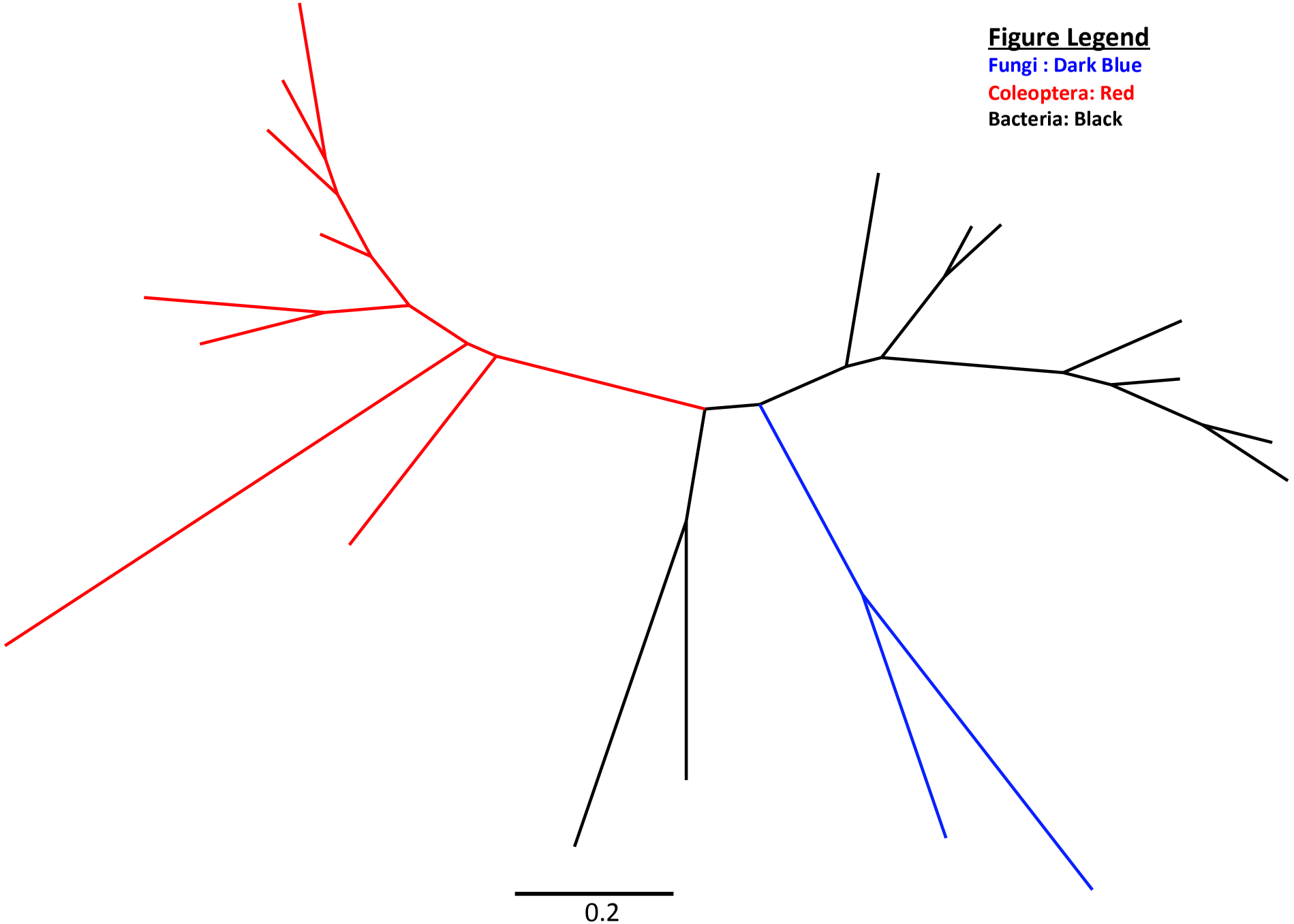


Fig. S18. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 10 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

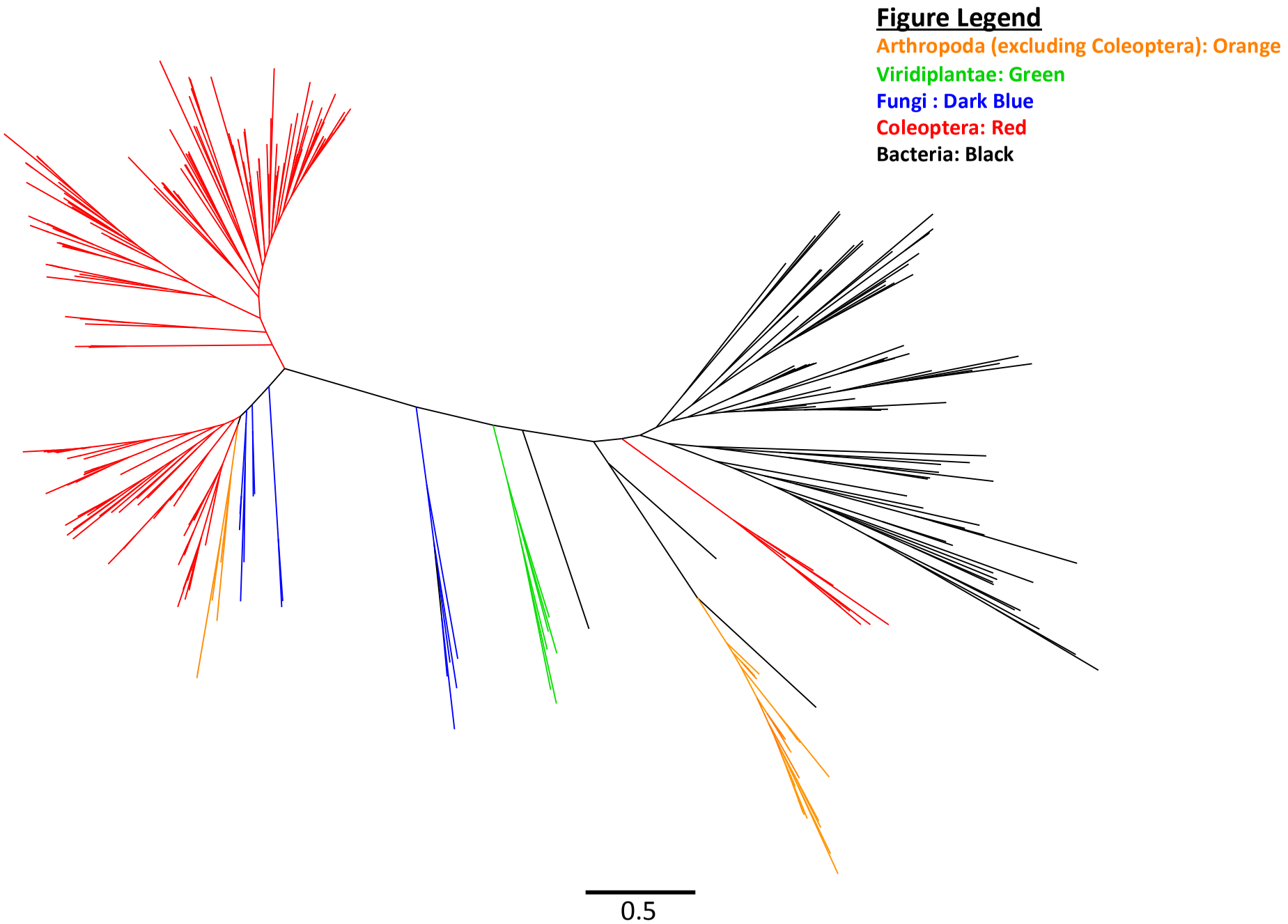


Fig. S19. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 28 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

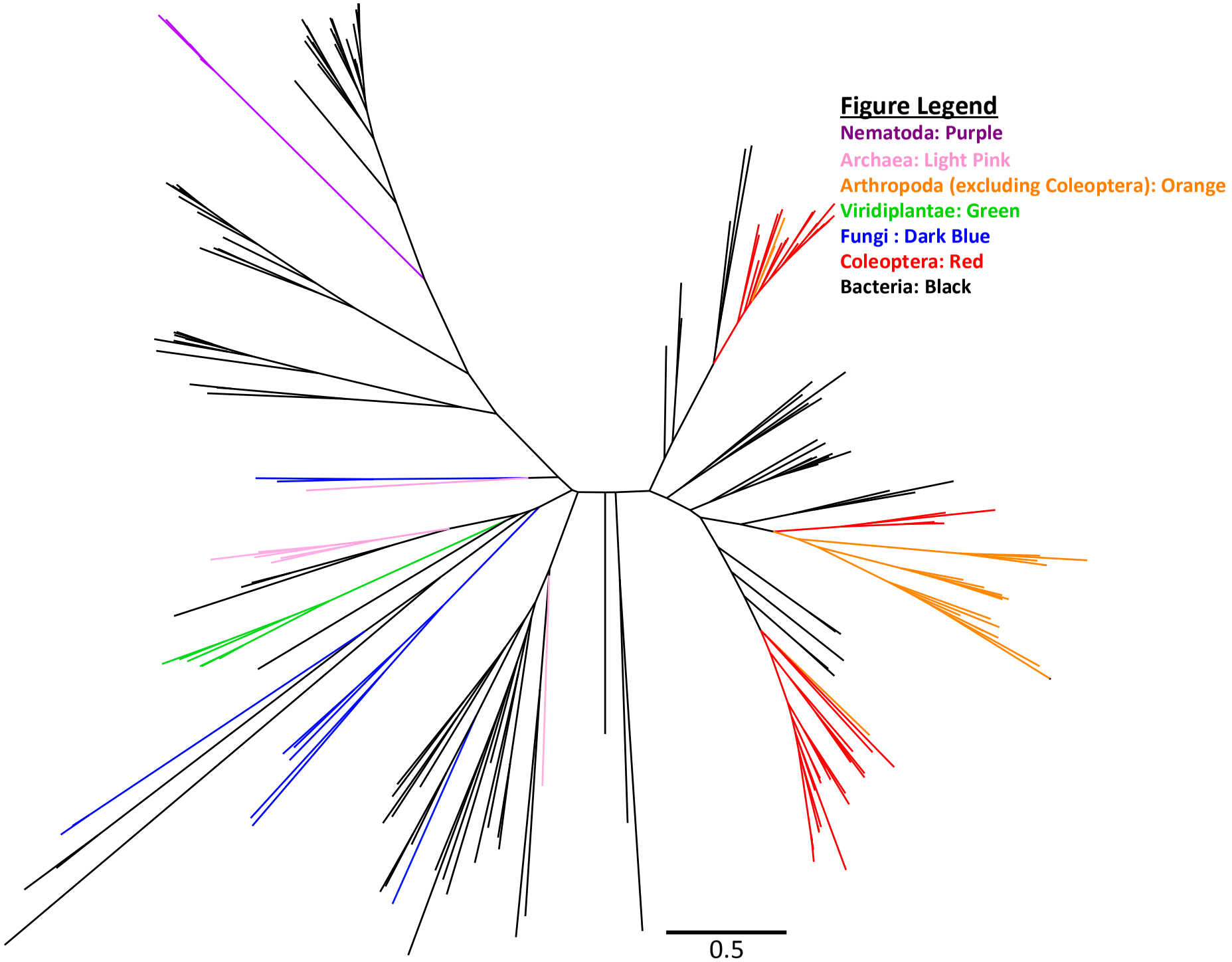


Fig. S20. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 32 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

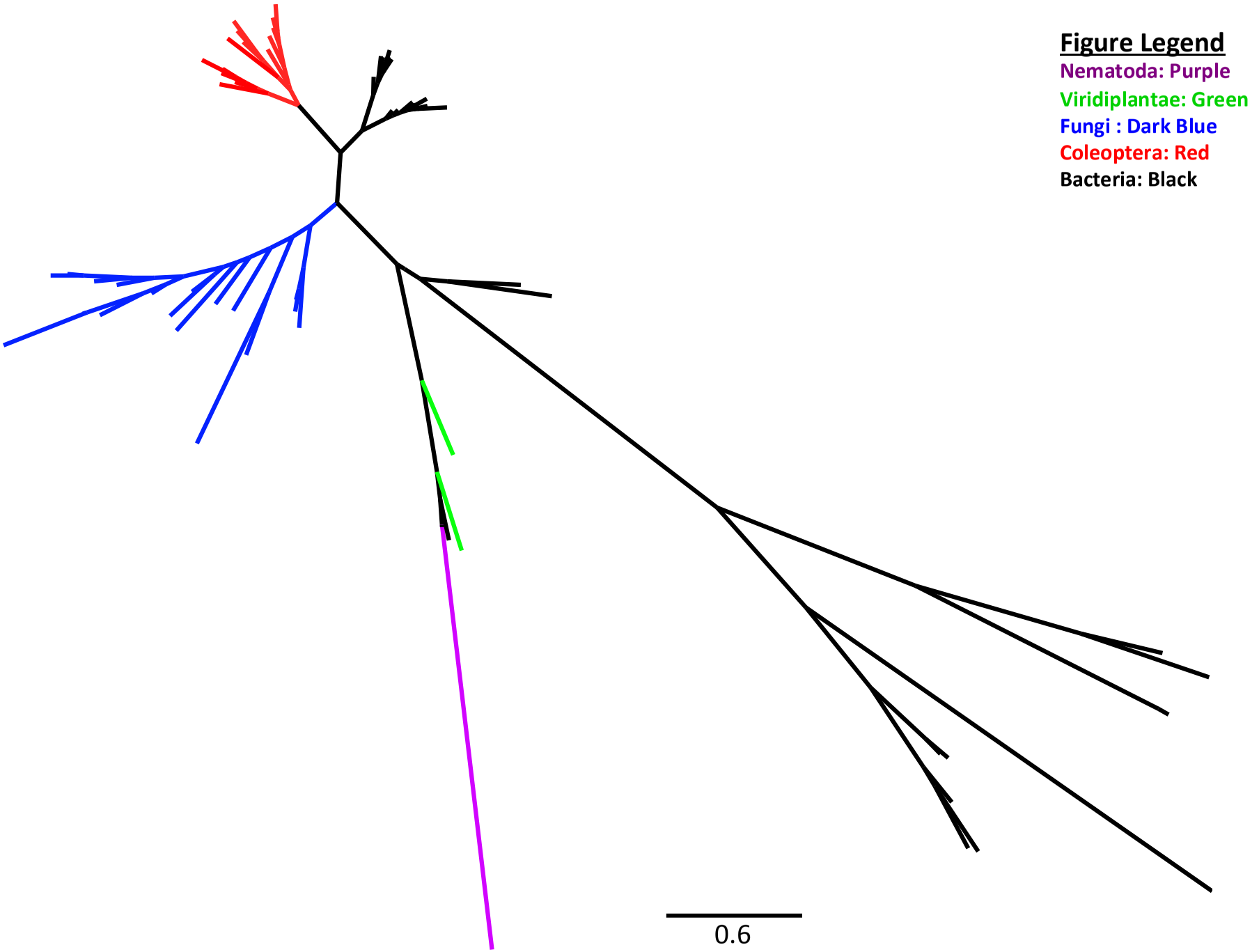


Fig. S21. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 43 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

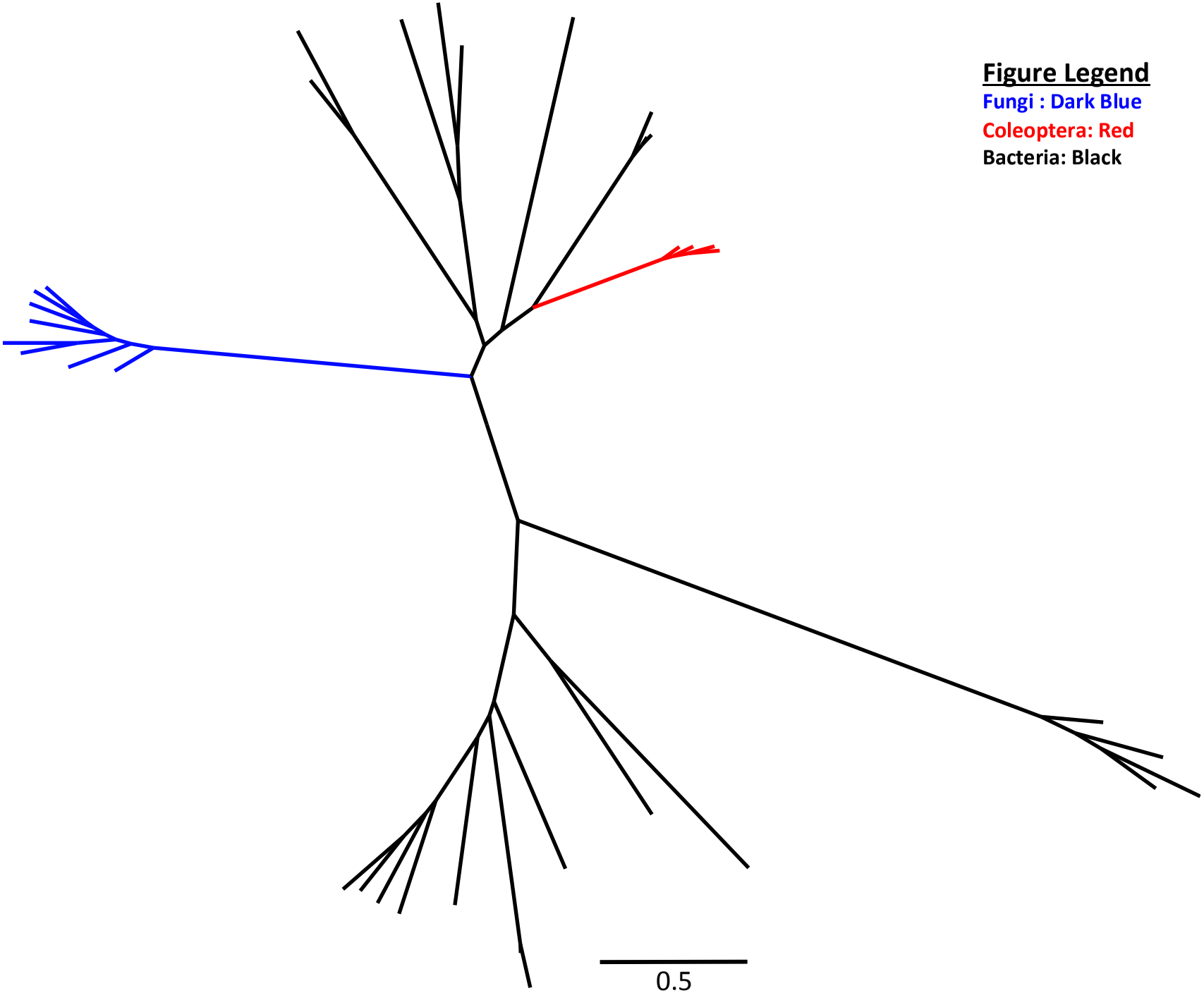


Fig. S22. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 44 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

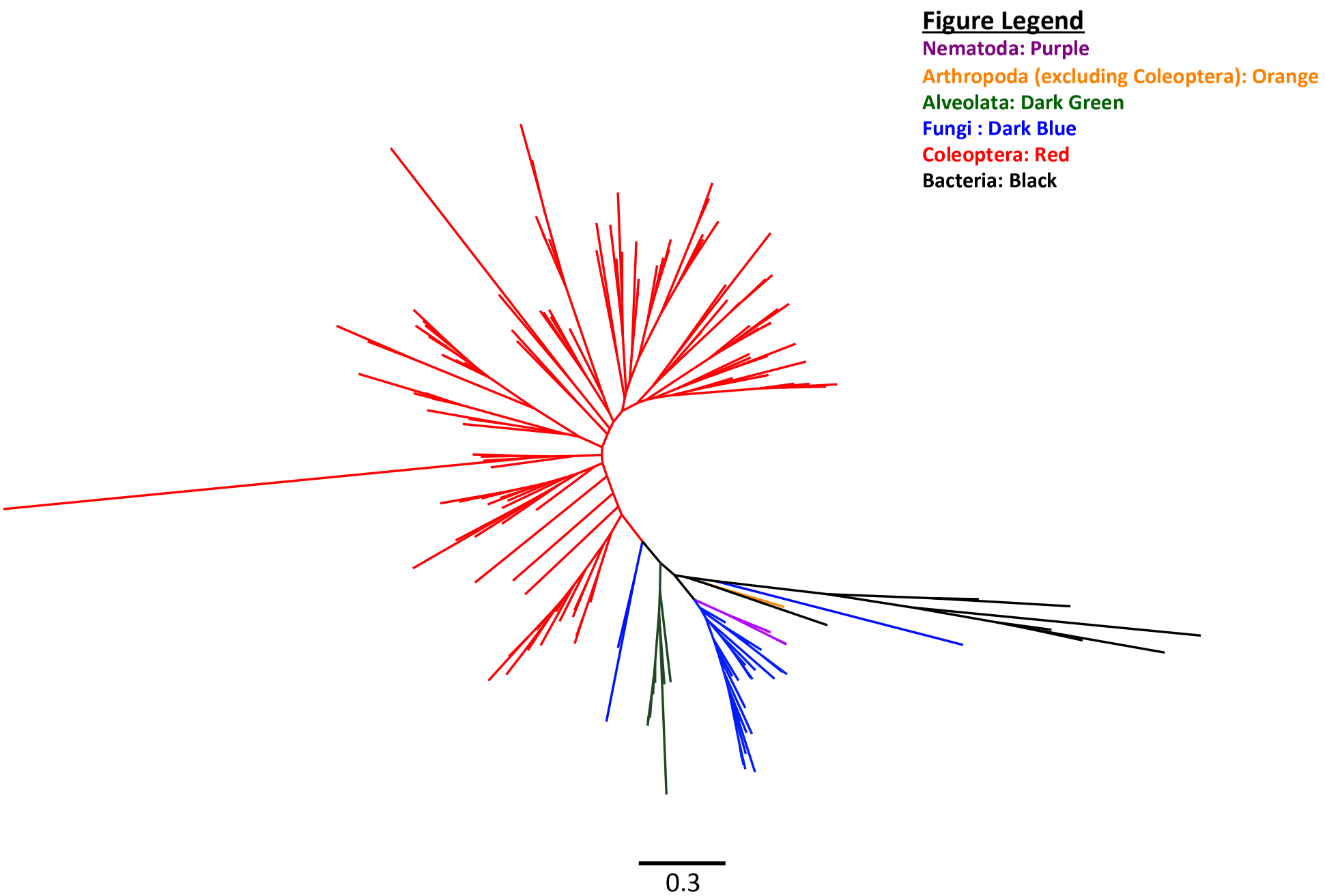


Fig. S23. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 45 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

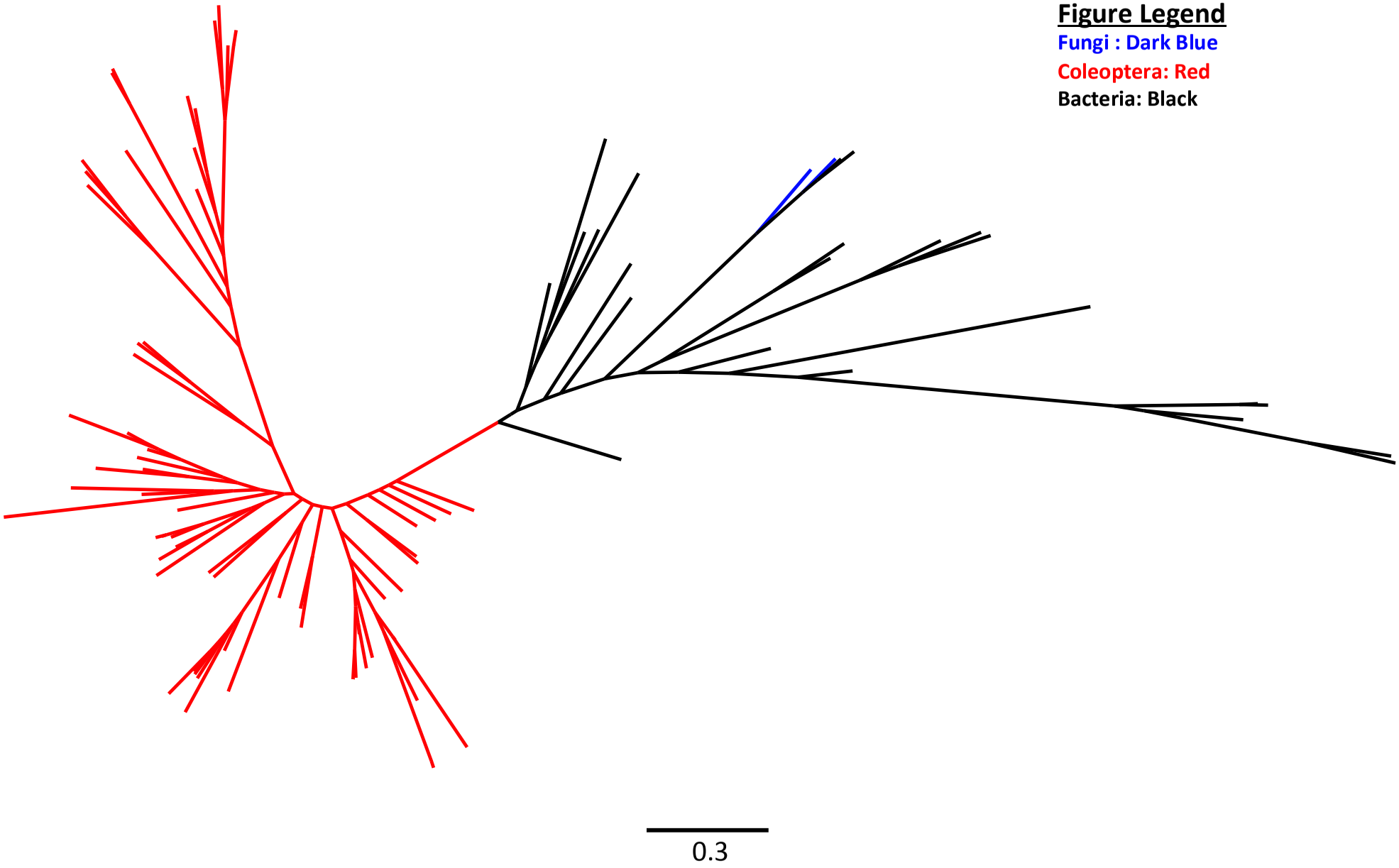


Fig. S24. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for glycoside hydrolase 48 family genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

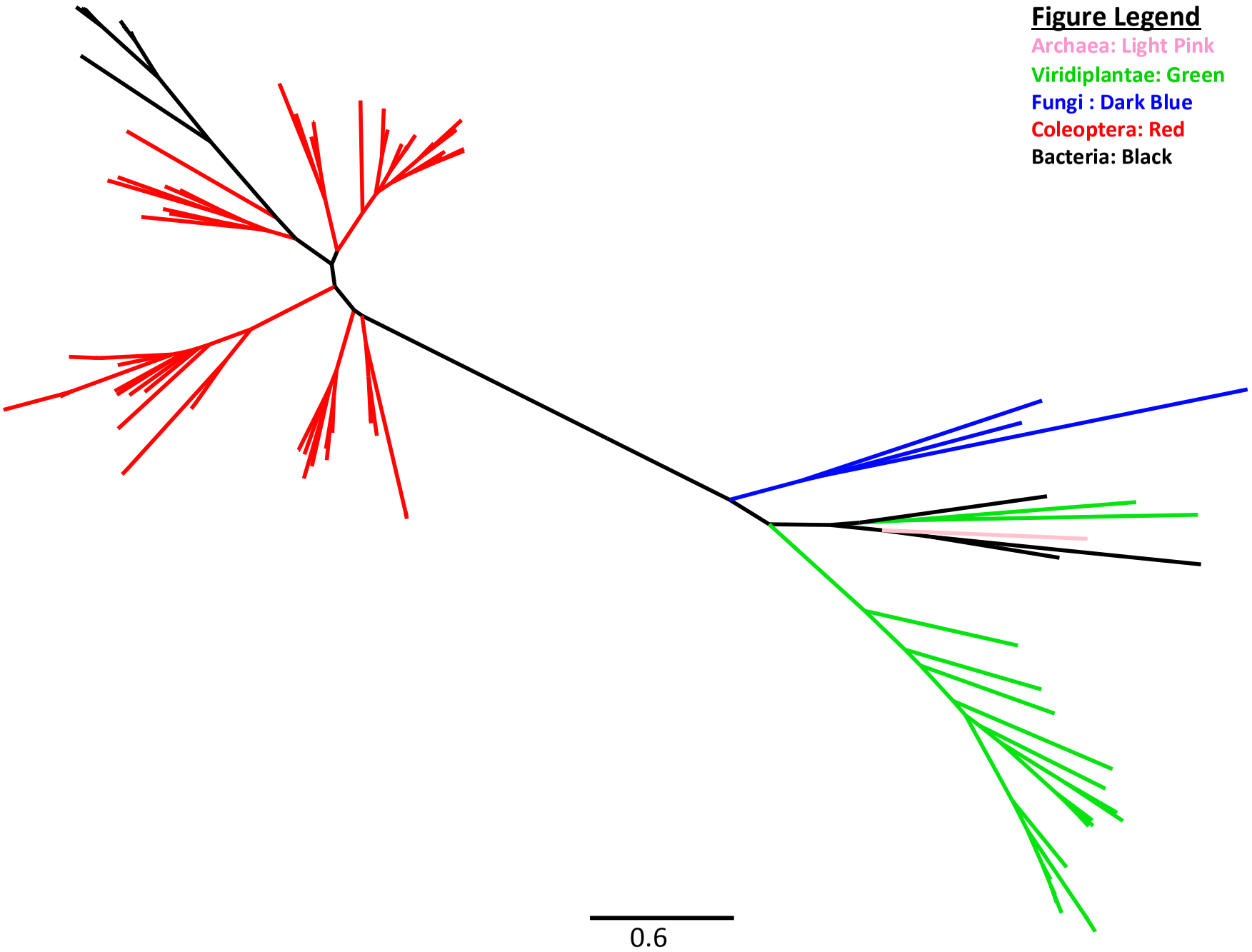


Fig. S25. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for CE8 genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.

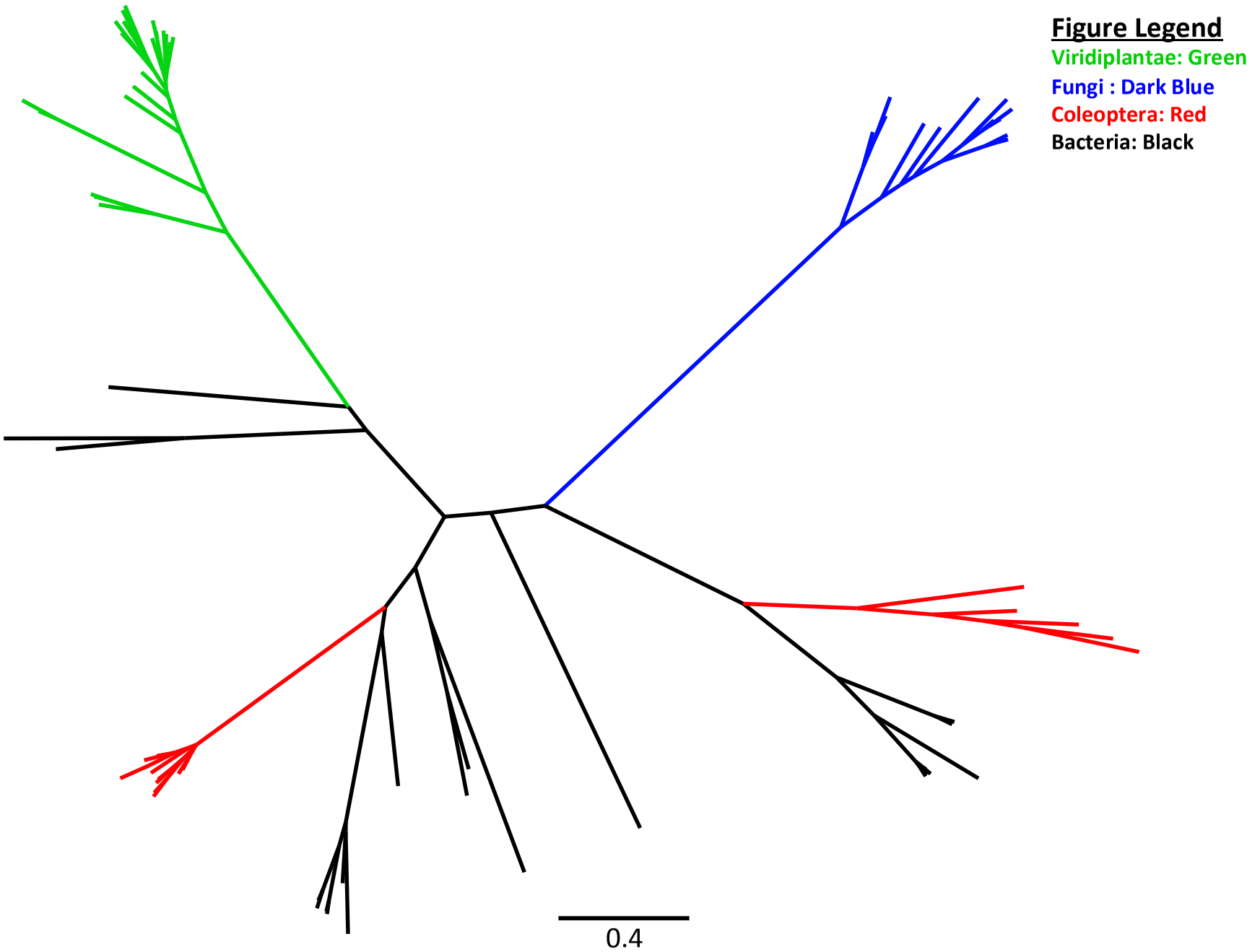
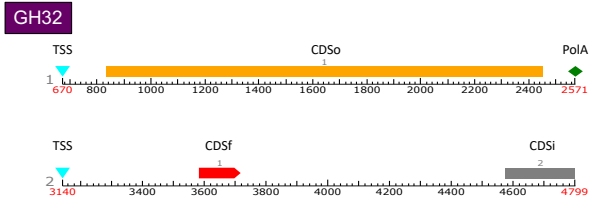
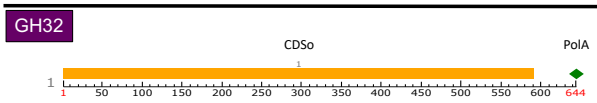


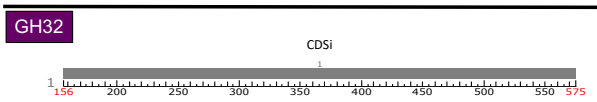
Fig. S26. Best tree resulting from maximum likelihood (ML) analysis of aligned amino acid sequence data for PL4 genes in the program RAxML (10 replicate ML searches). Taxon names, ML bootstrap support values (100 replicates) and transfer bootstrap expectation (TBE) support values (100 replicates) are available in the corresponding .tre files submitted to Zenodo.



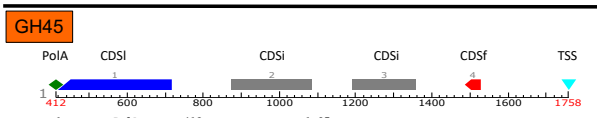
A. GH32_23_Mastostethus (Chrysomeloidea: Megalopodidae): There are two genes on this scaffold. One is a single exon gene with coding boundaries from 831 to 2450 in the +1 frame (length 1620) that is predicted to code for a GH32. It is a full length gene model with both start and stop codons, and it contains a eukaryotic TSS and a polyA signal. In our phylogenetic analyses it forms a clade with other beetle GH32 sequences. The second gene model codes for a partial pantothenate kinase with two exons. This insect gene has the start codon (CDSf; first exon), but no stop codon.



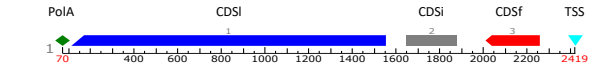
B. C28768136_GH32_24_Nanophyes (Curculionoidea: Brentidae): This scaffold contains a complete single exon gene with a polyA signal. In our phylogenetic analyses it forms a clade with other beetle GH32 sequences.



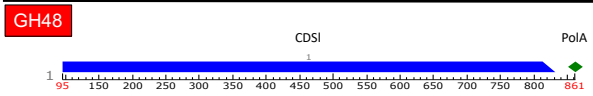
C. C26549544_GH32_47_Rhynchitomacerinus (Curculionoidea: Nemonychidae): This scaffold contains a partial GH32 sequence (internal CDS). In our phylogenetic analyses it forms a clade with other beetle GH32 sequences.



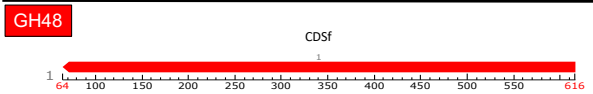
D. C3526741_GH45_104_Synolabus (Curculionoidea: Attelabidae): This scaffold contains a 4-exon (full-length) GH45 sequence. It also has a eukaryotic TSS and a polyA signal. In our phylogenetic analyses it forms a clade with other beetle GH45 sequences.



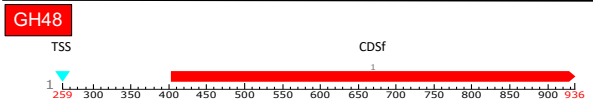
E. C3541577_GH48_1_Synolabus (Curculionoidea: Attelabidae): This scaffold codes for a 3-exon (full-length) GH48 and it has a eukaryotic TSS and a polyA signal. In our phylogenetic analyses it forms a clade with other beetle GH48 sequences.



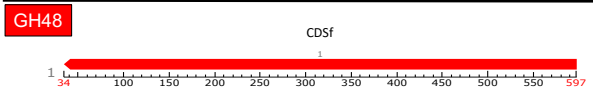
F. C29039538_GH48_Nanophyes (Curculionoidea: Brentidae): This scaffold contains a partial GH48 sequence with a stop codon and a polyA signal. In our phylogenetic analyses it forms a clade with other beetle GH48 sequences.



G. C28588156_GH48_Nanophyes (Curculionoidea: Brentidae): This scaffold contains a partial GH48 sequence containing a single internal exon. In our phylogenetic analyses it forms a clade with other beetle GH48 sequences.



H. C26661175_GH48_44_Rhynchitomacerinus (Curculionoidea: Nemonychidae): This scaffold contains a partial GH48 sequence and a eukaryotic TSS. In our phylogenetic analyses it forms a clade with other beetle GH48 sequences.



I. C26570892_GH48_43_Rhynchitomacerinus (Curculionoidea: Nemonychidae): This scaffold contains a partial GH48 sequence. In our phylogenetic analyses it forms a clade with other beetle GH48 sequences.

Figure S27. Schematics showing annotated genomic scaffolds that contain genes inferred to encode putative plant cell wall degrading enzymes and GH32 invertases, including introns (when present), eukaryotic TSS, and polyA signals. The scaffolds are organized by gene family and are shown only for exemplars from beetle families from which these genes have not previously been reported. The scaffolds were annotated using FGENESH version 2.6 (http://www.softberry.com/berry.phtml?topic=fgenesh&group=help&subgroup=gfind). We used the *Tribolium castaneum* (Tenebrionoidea: Tenebrionidae) genome as a reference for gene prediction.

Abbreviations: CDSf - first exon (beginning with start codon), CDSi - internal (internal exon), CDSl - last (ending with stop codon), CDSo - one (only one exon), TSS - position of transcription start (TATA-box position and score), PolA - polyA signal.