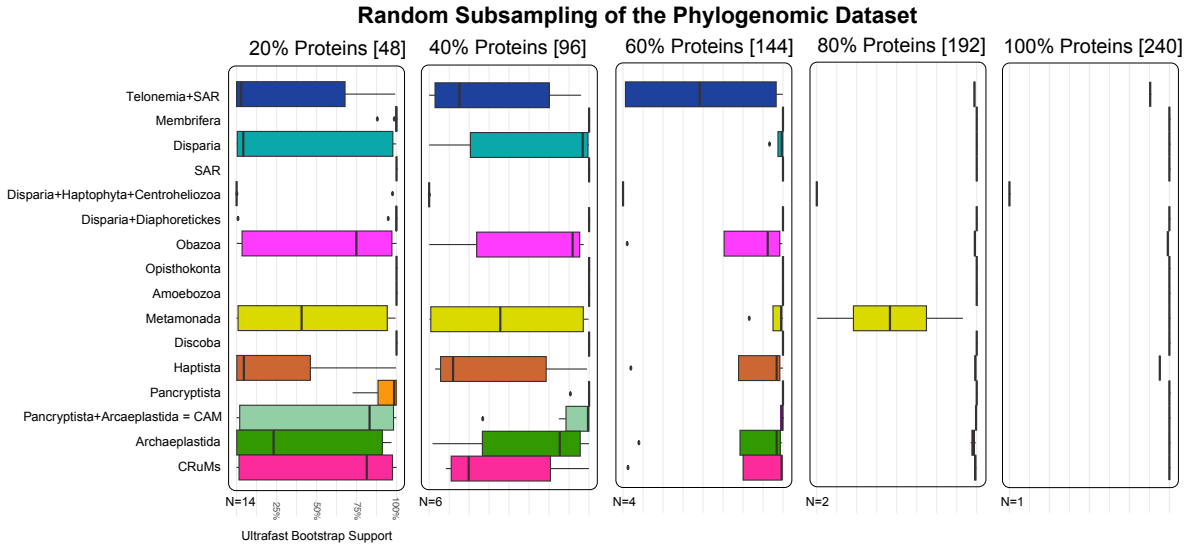
Supplementary Discussion 2: Extended discussion on phylogenomics and tree topology tests.

Phylogenomic matrices at deep phylogenetic timescales may be prone to many potential artifacts that may skew the inferences gleaned from the data. To assess and explore if there are phylogenetic artifacts within the dataset that may provide competing signals, we conducted numerous experiments that further support our phylogenomic conclusions, namely the validity of the Disparia supergroup, and the Membrifera and Provora subclades. Our first approach was to randomly subsample the orthologs within our phylogenomic dataset at various depths and assess the support of each clade of interest. We used several methods to identify outlier orthologs within our phylogenomic matrix, this included both the PhylteR package[101] and the methodology of Shen et al.[102] to calculate the delta site-log likelihood scores of genes (ΔGLS) under various alternative topological hypotheses and examine the impact of removal of genes from the matrix deemed as outliers both in support and in opposition of the alternative hypotheses. We explored concordance vectors[123] of all major clades of eukaryotes and further investigated the Disparia clade, using the discordance trees as additional alternative hypotheses for the ΔGLS analyses. We tested the impact of simple amino acid alphabet recoding to mitigate compositional heterogeneity and substitution saturation in our phylogenetic analyses. We further examined the potential impact of heterotachy, which is variations in lineage-specific evolutionary rates over time, within our phylogenomic matrix using approaches developed in Tice et al.[28]. Taken as a whole, our analyses are in support of Disparia, Membrifera, Provora, and Caelestes as clades. Below, we detail our results for all analyses.

Random subsampling of orthologs

To test the global impact of ortholog selection within the phylogenomic matrix, the 240 proteins were randomly sampled at various percentages (20%-80%) to generate new phylogenomic matrices. Subsampling was conducted using the *random_subsampler.py* utility in the PhyloFisher package. From each matrix, a bootstrapped tree was inferred, and support for the bipartitions of interest was plotted. These analyses show a global trend that when more data is added (sampled), phylogenetic support increases (Suppl. Discussion 2 Fig. 1). A few deep clades of eukaryotes (Opisthokonta, Amoebozoa, Discoba, Pancryptista, and SAR) appear to be robust to the sampling of orthologs. However, most larger clades require a deeper sampling of orthologs, for example for Disparia, 40-60% of orthologs must be sampled to reveal full support of the clade. This is also the phenomenon observed in Obazoa, Archaeplastida, Metamonda, CAM (Pancryptista + Archaeplastida), and CRuMs, all of which are well-supported clades by deep global phylogenomic analyses observed in multiple studies[11] using datasets independent of our own (Suppl. Discussion Fig. 1).



Suppl. Discussion 2 Fig. 1| Examination of the effect of random subsampling of proteins in the PhyloFisher dataset on the bipartitions of interest. The number of replicates (n) necessary for a 95% probability of sampling every protein when subsampling 20, 40, 60 and 80% of proteins was calculated using the formula: $0.95 = 1 – (1 – x/100)^n$, where x is the percentage of proteins subsampled. ML analyses were conducted under ELM+C60+G using the PMSF method in IQ-TREE v2.3.4 with 1,000 Ultrafast bootstrap replicates. The ultrafast bootstrap support values for all nodes of interest with the variability of support values illustrated by box-and-whisker plots.

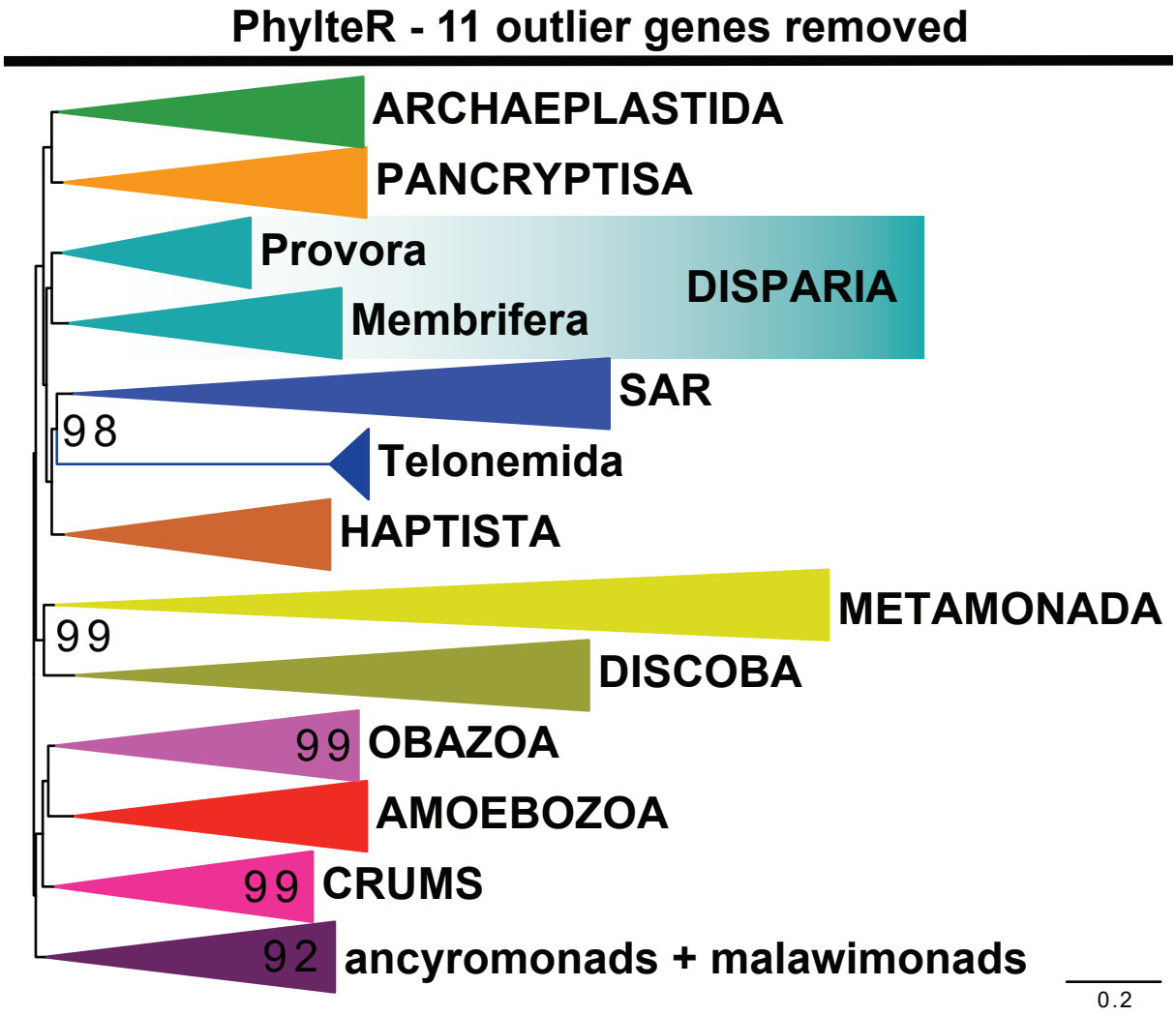
Outlier gene removal identified through PhylteR

Several tools and approaches have been developed to identify orthologs that have discordant phylogenetic histories when compared to an overall set of ortholog trees or to a species tree. PhylteR is a tool that detects outliers in phylogenomic dataset by iteratively removing taxa from ortholog trees and optimizing a score of tree concordance factors for matrices. This approach is independent of the taxa of interest and aims to detect global outliers within the data. Using default parameters with setting changes, we tested several k (k: Strength of outlier detection) thresholds to detect outliers, the larger k value, the fewer outliers will be detected. To aggressively detect more outliers in the dataset, we tested multiple "k" settings and found that lowering the k value to 1 detected 11 "complete outlier" orthologs (i.e., BTUB, C3H4, CAPZ, FTSJ1, GCST, H2A, IPO4, RAD51A, RICTOR, RPS16, and TMS). To test the impact of these outliers, we constructed a new supermatrix without these proteins, yielding a dataset consisting of 229 proteins. Removal of these proteins did not affect the monophyly of Disparia or other major clades of eukaryotes (Suppl. Discussion 2 Fig. 2).

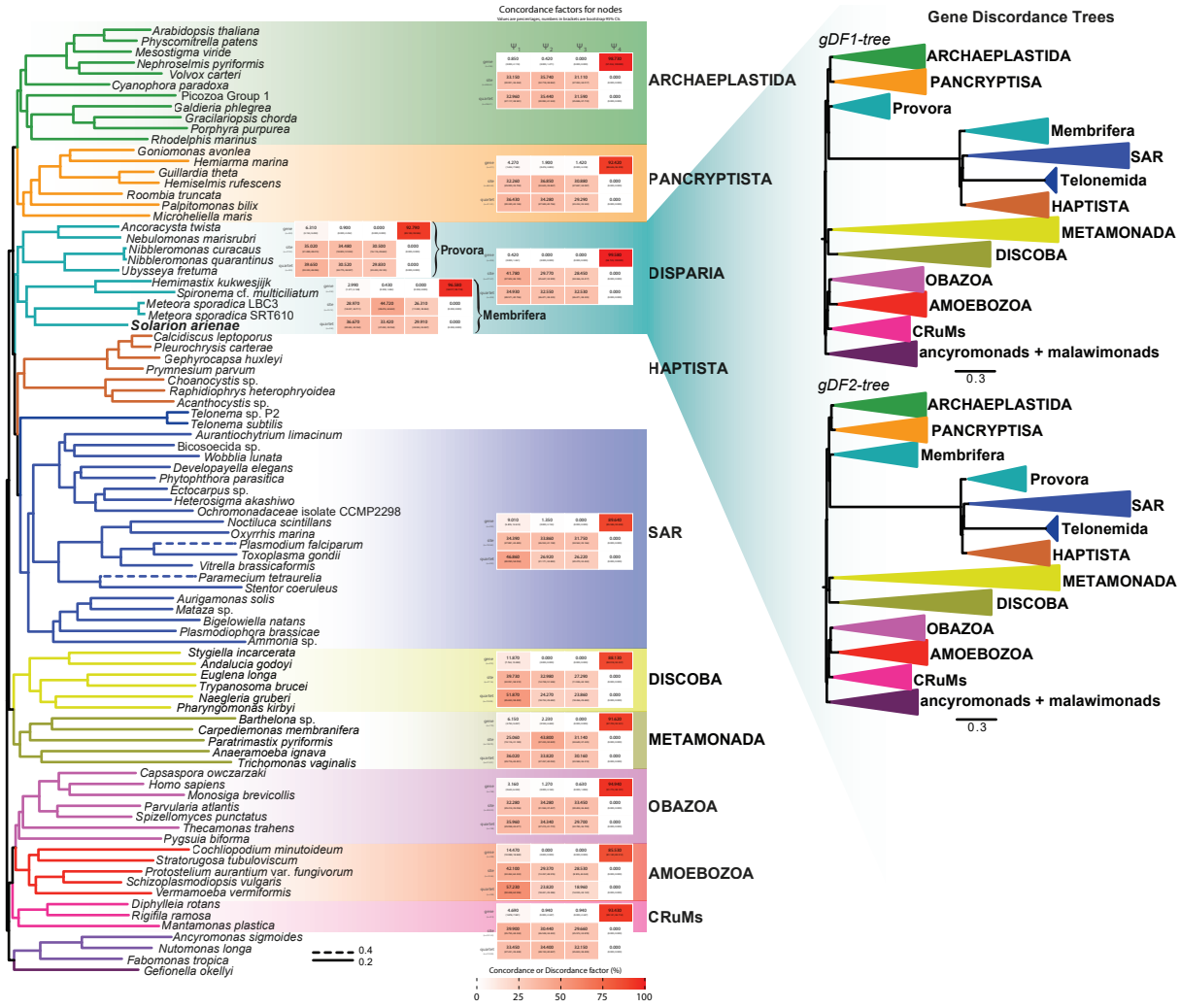
Tree concordance vector analyses

Using the approaches developed in Lanfear & Hahn[123], concordance vectors for the nine major clades of eukaryotes were calculated. These vectors generally show poor concordance despite maximal or near-maximal statistical support in Fig. 2 of the main text. Concordance vectors were also calculated for Membrifera and Provora. For the Disparia clade, 41.8% of sites support

topology Ψ1 versus alternative topologies (Suppl. Discussion 2 Fig. 3). Concordance factors are difficult to interpret at this level of the phylogenetic depth; however, using this approach, we were able to identify discordance trees for the Disparia clade. These trees were used as alternative topologies for examining using ΔGLS tests below.



Suppl. Discussion 2 Fig 2| Phylogenetic tree with removed genes identified as outliers by PhylteR. PhylteR was used with default parameters and an aggressive K number (k=1), identifying 11 genes as outliers. A new matrix with these outlier genes (BTUB, C3H4, CAPZ, FTSJ1, GCST, H2A, IPO4, RAD51A, RICTOR, RPS16, and TMS) excluded was generated (74,159 AA sites), and an ML tree was inferred and bootstrapped with ultrafast bootstrapping (1,000 replicates) under the ELM+C60+G using the PMSF method in IQ-TREE v2.3.4. Nodes without support values represent 100% ultrafast bootstrap support.



Suppl. Discussion 2 Fig. 3| Concordance vectors for each clade reveal that overall concordance at the deepest levels of the tree is similar across the deep tree of eukaryotes, when evaluated from the PhyloFisher dataset. Concordance vectors were calculated per node of interest in the species tree (to the left, same as Fig. 2) using the recipe in http://www.iqtree.org/doc/recipes/concordance-vector. Concordance factor vectors were mapped to the ML tree inferred under ELM+C60+G from the phylogenomic matrix. To calculate the vectors, an ML tree was inferred individually from each single protein under the model ELM+C60+G in IQ-TREE v2.3.4, provided as input for the recipe. Opposite each clade name is a matrix of the gene, site, and quartet concordance vectors. The numbers in each cell are percentages, with higher percentages colored in darker shades of red. From these analyses, we used IQ-TREE to identify discordant trees associated with gene discordance factors (gDF), which in essence, is an alternative discordant topology. IQ-TREE outputs two trees labeled as gDF1-tree and gDF2-tree in the concordance factor statistics file. [iqtree2 -te astral_species_annotated.tree -p partition_file.nex --df-tree --scfl 100 --prefix scfl-dftree -T 126] and [iqtree2 -t scfl-dftree.cf.tree --df-tree --gcf alltrees.treefile --prefix gcf-dftree-again -T 126]. We identified the node ID for Disparia (node 126) in the vector analyses file (id_gcf_scf_qcf.nex) by viewing the tree in Dendroscope3[103]. Using the output from IQ-TREE (gcf-dftree-again.cf.stat), we collected the gDF1-tree and gDF2-tree (displayed as a collapsed tree to the right). From gDF1-tree and gDF2-tree, a loose constraint phylogeny was generated, with two alternatives of Disparia as a

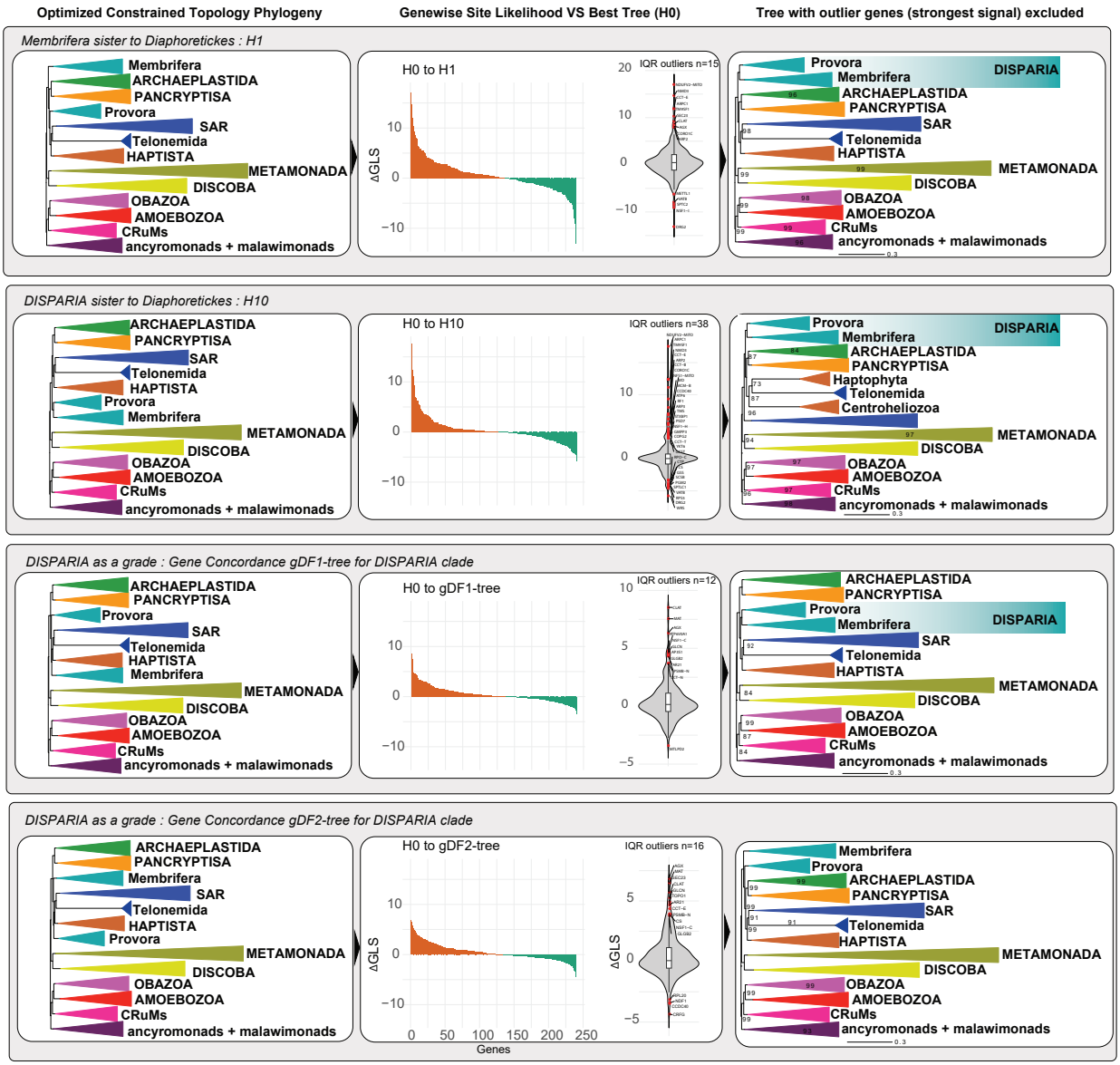
grade, which was optimized in IQ-TREE under ELM+C60+G and used for deltaGLS analyses below. Support values for the trees are not presented in these figures.

Impact of removal of outlier genes (orthologs) in support or opposition to alternative hypotheses

In a landmark study of Shen et al.[102], it was shown that some inferred phylogenomic relationships hinge on data from just a single gene or a few genes. The study demonstrated that removing these small data subsets can significantly alter the phylogenetic tree's topology, indicating that minimal data can disproportionately influence results. To test this phenomenon on our data, we used the approach therein. We examined four alternative hypotheses two of which had the best log-likelihood in our Approximately Unbiased test (H1 and H10 in Supplementary Table 1) as well as the two discordance trees identified in the concordance vector analyses (Suppl. Discussion 2 Fig. 3). Using the differences in the sum of the site log-likelihood per gene in comparisons between the best tree and each alternative tree, we removed genes identified as outliers in their support or opposition of either topology to generate a new matrix for each comparison (Suppl. Discussion 2 Fig. 4). Removing these outliers for H1, H10, and gDF1 had no impact on the support for Disparia, whereas the gDF2 outliers result in Disparia as a grade. These results mostly support the monophyly of Disparia but show some conflicting signal within the clade and its placement in the tree of eukaryotes.

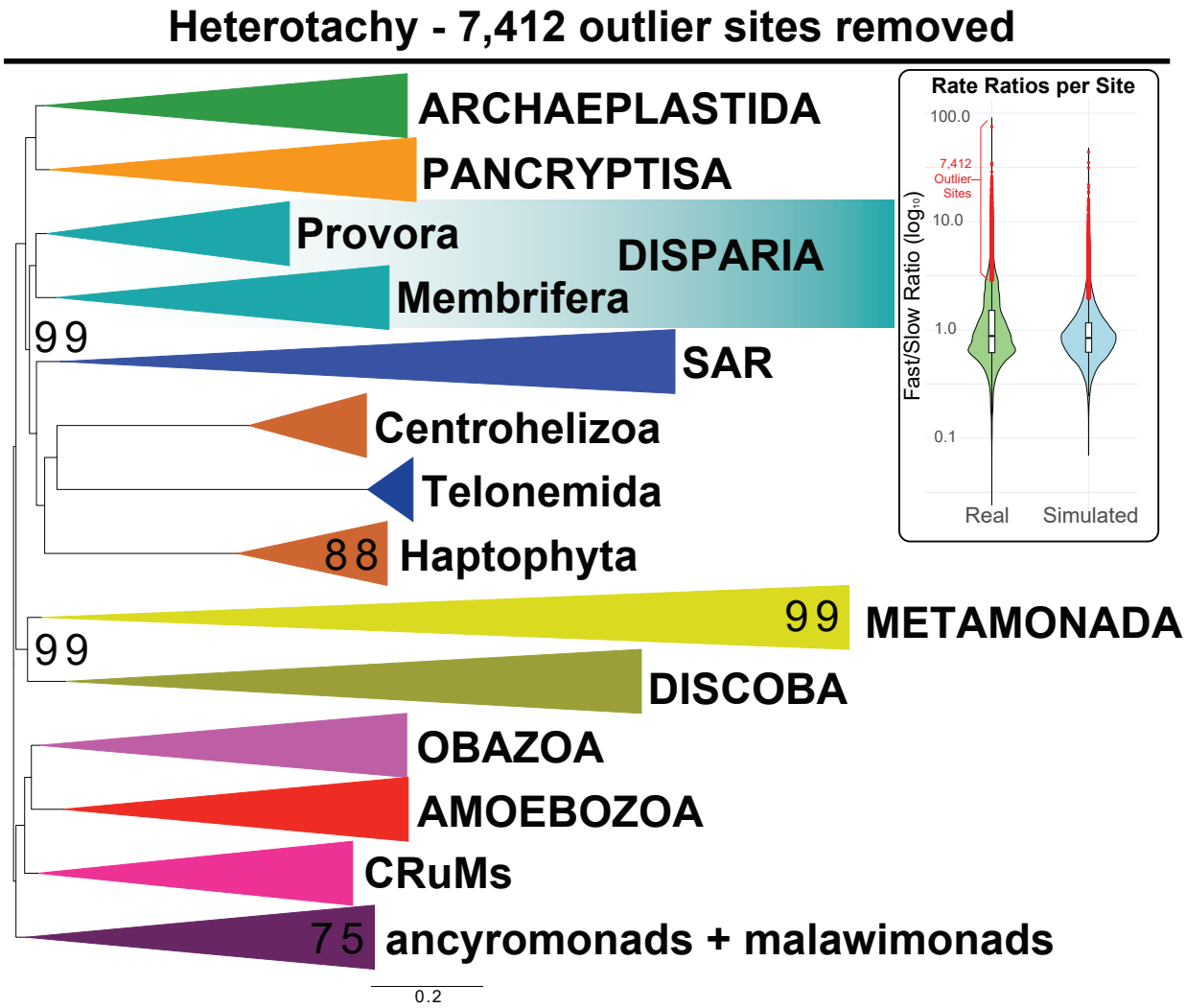
Identification and removal of heterotachious sites

Much like fast-evolving sites within a phylogenomic matrix, heterotachious sites, those that are evolving at differing rates across branches in the tree, are susceptible to generating phylogenetic reconstruction artifacts. To examine the impact of potentially heterotachious sites, we simulated datasets under the ELM+C60+G model and compared the heterotachy metrics between simulated and real datasets. Using the heterotachy metrics generated by the *heterotachy.py* utility in the PhyloFisher software package, we calculated the most heterotachious sites using an interquartile range strategy (Suppl. Discussion 2 Fig. 5). We then removed the outlier heterotachious sites (7,412 sites). Removal of these sites had no impact on the support of the Disparia clade, suggesting that heterotachy is not an artifact contributing to this clade.



Suppl. Discussion 2 Fig. 4| Our approach for quantifying and visualizing phylogenetic signal of potentially conflicting genes within the phylogenomic dataset using the approach presented in Shen et al.[102]. Four alternative phylogenetic hypotheses were explored. Constrained trees from Approximately Unbiased tests for two alternative hypotheses (H1, H10) with the best log likelihood scores and gDF1-tree and gDF2-tree in the concordance factor experiments for the Disparia node. These constraints were provided as loose constraints to IQ-TREE v2.3.4 using the "-g" function {iqtree2 -m ELM+C60+G -mwopt -T 64 -s matrix.fas -fs PF.Dec102023.SUMK.87.ELMC60G-PMSF.sitefreq -pre PF.Dec102023.SUMK.87.ELMC60G-PMSF.constraint.treeDF2 -g treeDF2.tre} under the ELM+C60+G using the PMSF method to infer the optimized constraint tree (left column of figure). Site log likelihood scores were inferred under the same model for each optimized constraint tree and the unconstrained best ML tree of Fig. 2 (here referred to as H0, null hypothesis) in IQ-TREE {iqtree2 -m ELM+C60+G -s matrix.fas -pre PF.Dec102023.SUMK.87.ELMC60G_mwopt.t-PMSF -fs PF.Dec102023.SUMK.87.ELMC60G-PMSF.sitefreq -mem 980G -T 60 -mwopt -g PF.Dec102023.SUMK.87.ELMC60G-PMSF.constraint.treeDF2.treefile --sitelh}. Using the

indices for each gene of the 240 in the PhyloFisher supermatrix (available at FigShare https://doi.org/10.6084/m9.figshare.27182820), the overall site log likelihood score (lnL) was generated as the sum of lnL for each site within a gene. This was calculated for each gene under the constraint hypothesis tree and H0. On a genewise basis, summed lnL scores were compared between H0 and each constrained tree, providing a difference in gene likelihood score (i.e., ΔGLS). The ΔGLS scores for each gene were plotted in descending order of their ΔGLS values for each constraint (orange are in favor of H0, green are in favor of the constraint tree). The ΔGLS values are plotted as violin plots, and outliers were identified using the interquartile range (IQR) method. For each constraint tree, IQR gene outliers (both above and below the IQR threshold, i.e., those genes with the strongest phylogenetic signal) were excluded in the creation of a new matrix from which an ML tree was inferred and bootstrapped with ultrafast bootstrapping (1,000 replicates) under the ELM+C60+G using the PMSF method in IQ-TREE v2.3.4. For the final tree (right side of the figure), nodes without support values indicated represent 100% ultrafast bootstrap support.

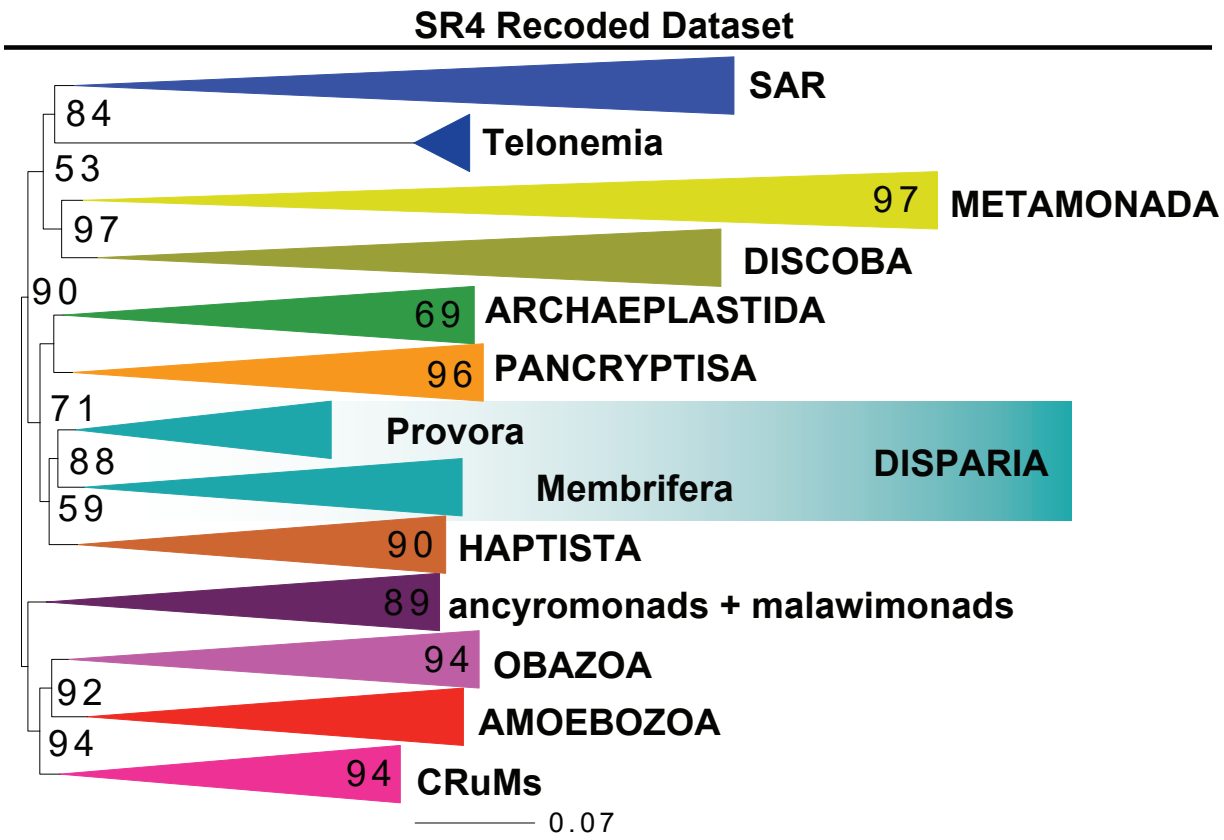


Suppl. Discussion 2 Fig. 5| Phylogenetic tree with the most heterotachious sites removed. The ratio of fast to slow taxa site rates, on a per-site basis, estimated from the real dataset (green) and nine simulated datasets (blue) (right inset violin plots). Datasets were randomly simulated

nine times under the ELM+C60+G model of evolution using our primary output tree under this model (Fig. 2) with the primary dataset. Fast/slow taxa site ratios were estimated using the *heterotachy.py* utility. To examine the expected among-site rate variation, given our ML tree as accounted for within the ELM+C60+G model of evolution, a ratio of fast to slow taxa site rates was inferred from the simulations under the ELM+C60+G model. The nine simulated datasets were generated using the *alisim* function in IQ-TREE v2.3.4. To do this, we provided *alisim* our phylogenomic ML tree inferred under the same model from the real dataset with fixed branch lengths, generating alignments of the same size as the real dataset (77,133 AA) {iqtree2 --alisim alisimELMC60G_1 -m ELM+C60+G -t ../PF.Dec102023.SUMK.87.ELMC60G.treefile -T 2 --seqtype AA -af fasta -blfix --length 77133}. To reduce any biases generated by missing data, the pattern of gaps and missing data was transferred to the simulated data from the real dataset using a custom Python script (https://github.com/TheBrownLab/rjones-scripts/blob/main/scripts/replicate_missing_data.py). From hence simulated datasets and the real data, site rates were calculated using the *heterotachy.py* utility using the methodology described within Tice et al.[28], which resulted in evolutionary rates calculated per site in the dataset for slow (slowest 1/3rd) and fast (fastest 1/3rd) taxa, based on their branch lengths in our phylogenomic ML tree inferred from the real dataset. Using the output site rates for Fast and Slow taxa (i.e., fast.rate_est.dat and slow.rate_est.dat), a ratio of Fast to Slow was computed for each site in both real and simulated data. These ratios were sorted and used to calculate outlier heterotachious site ratios using the interquartile range method. In the real data, 7,412 sites were identified as outliers and removed from the dataset. From this reduced dataset (69,721 sites), an ML tree was inferred and bootstrapped with ultrafast bootstrapping (1,000 replicates) under the ELM+C60+G using the PMSF method in IQ-TREE v2.3.4. Nodes without support values indicated represent 100% ultrafast bootstrap support.

Amino Acid alphabet recoding

Amino acid recoding has been proposed as a strategy to improve the effects of phylogenetic reconstruction artifacts rooted in compositional heterogeneity and substitutional saturation of a supermatrix. Even though a recent study showed that recoding is not a recommended method to offset these biases, outside of the most extreme cases[124], a previous work has suggested that *Meteora* + Hemimastigophora are perhaps not the sister group to Provora (*Ancoracysta*) due to recoding analyses' inability to recover the topology[23]. In this, the most recent study on taxa relevant to this manuscript, recoding the phylogenomic matrix using the SR4 alphabet[125] did not recover *Ancoracysta* (Provora) + *Meteora* + *Hemimastigophora* and placed *Ancoracysta* elsewhere. To examine if the same observation held true in our data, we recoded the dataset into an SR4 alphabet and inferred a phylogenetic tree. The clade Disparia is recovered even when SR4 recoding is applied (Suppl. Discussion 2 Fig. 6).



Suppl. Discussion 2 Fig. 6| Phylogenomic tree of our PhyloFisher generated dataset recoded from 20 amino acids into four categories, helping to preserve more of the phylogenetic signal over long evolutionary timescales. The dataset was recoded using the *aa_recoder.py* utility in the PhyloFisher software package. From this, an ML tree was inferred and bootstrapped with ultrafast bootstrapping (1,000 replicates) under the GTR+G+F model in IQ-TREE v2.3.4. Nodes without support values indicated represent 100% ultrafast bootstrap support.

References:

123. Lanfear, R. & Hahn, M., W. The Meaning and Measure of Concordance Factors in Phylogenomics. **41**, (2024).

124. Hernandez, A. M. & Ryan, J. F. Six-State Amino Acid Recoding is not an Effective Strategy to Offset Compositional Heterogeneity and Saturation in Phylogenetic Analyses. *Syst. Biol.* **70**, 1200–1212 (2021).

125. Susko, E. & Roger, A. J. On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).