science.sciencemag.org/content/372/6542/eabe0511/suppl/DC1

Supplementary Materials for

A rooted phylogeny resolves early bacterial evolution

Gareth A. Coleman†, Adrián A. Davín†, Tara A. Mahendrarajah, Lénárd L. Szánthó, Anja Spang, Philip Hugenholtz‡*, Gergely J. Szöllősi‡*, Tom A. Williams‡*

†These authors contributed equally to this work.
‡These authors contributed equally to this work.
*Corresponding author. Email: p.hugenholtz@uq.edu.au (P.H.); ssolo@elte.hu (G.J.Sz.); tom.a.williams@bristol.ac.uk (T.A.W.)

Published 7 May 2021, *Science* **372**, eabe0511 (2021)
DOI: 10.1126/science.abe0511

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/372/6542/eabe0511/suppl/DC1)

MDAR Reproducibility Checklist (.pdf)
Tables S4 and S7 to S9 (.xlsx)

Materials and Methods

Taxon sampling for focal analysis

To obtain a representative taxon sampling from across known bacterial diversity, we sampled taxa according to the classification provided by the Genome Taxonomy Database (GTDB r89)(*13*) as follows. First, we removed genomes with Quality < 0.75 (Quality is defined as Completeness - (5*Contamination)(*14*)), and filtered out all phyla subsequently left with fewer than 10 species. Genomes were sampled from the remaining taxa on a per-class basis: for classes containing a single order, the genome with the highest quality score was sampled; for classes containing multiple orders, the highest quality genome from each of two randomly chosen orders was sampled. This protocol ensured that every class in the GTDB is represented in the final tree. We then manually added the genome of *Gloeomargarita litophora* given its importance in constraining the phylogeny and timing of chloroplast evolution. The list of genomes can be found in table S9.

Unrooted species tree inference

We used Orthologous Matrix (OMA) 2.1.1 (*73*) to identify candidate single-copy bacterial orthologs, and retained those with at least 75% of all species represented in each family. Sequences were aligned in MAFFT *(74)* using the -auto option, and trimmed in BMGE 1.12 (*75*) using the BLOSUM30 model. Initial trees were inferred for each candidate marker gene under the LG+G+F model in IQ-TREE 1.6.10 *(76)*. The trees were manually inspected, and we selected orthologues where the monophyly of 14 pre-defined major lineages was not violated with bootstrap support >70%, resulting in 62 final orthologues. Concatenation of this marker set resulted in an alignment of 18,234 amino acids. We inferred an unrooted phylogeny from this concatenate under the LG+C60+R8+F model, which was chosen as the best-fitting model by the BIC criterion in IQ-TREE (*76*). We additionally removed the most compositionally heterogeneous sites from the sequence alignment using Alignment Pruner (*81*) (https://github.com/novigit/davinciCode/blob/master/perl) (20%, 40%, 60% and 80% respectively) and inferred trees using the same procedure described above in order to compare the resulting topologies.

Outgroup rooting

To root the bacterial tree using an archaeal outgroup, we used a representative sampling of 148 archaeal genomes and inferred the ML tree in IQ-TREE under the best-fitting LG+C60+R8+F model. The concatenated alignment included a subset of 29 out of the 62 bacterial orthologs that were shared between bacteria and archaea, as determined by by Hidden Markov Model (HMM) searches and manual inspection of single gene trees. ML trees for these 29 genes recovered the clanhood (*82*) of Bacteria and Archaea. We performed approximately-unbiased (AU) tests *(28)* to determine whether a range of published alternative rooting hypotheses (table S2) could be rejected, given the model and data (AU p-value > 0.05).

Gene family clustering and ALE analysis

We used the protein annotations provided by GTDB, which were originally obtained using Prodigal. To infer homologous gene families for amalgamated likelihood estimation (ALE), we performed an all vs all similarity search using Diamond (*77*) with an E-value threshold of $<10^{-7}$ to avoid distant hits and k = 0 to report all the relevant hits. Current clustering methods are not consummate and the parameters that determine the granularity of clustering do not have a direct biological motivation. Setting the value of the Markov Cluster (MCL) algorithm *(83)* inflation parameter therefore involves a trade-off between inferring large, inclusive clusters that will contain false positives (sequences that are not part of the real gene family) and small, conservative clusters that may divide real gene families into several subclusters. An additional practical concern for phylogenomics is that overly large clusters may align poorly and result in low-quality single protein trees. In our rooting analysis, we experimented with a range of values for the mcl inflation parameter, and chose 1.2 because the clusters were inclusive without a substantial reduction in post-masking alignment length compared to more granular settings.

Clustering using MCL (*83*) with an inflation parameter of 1.2 resulted in 186,827 gene families and a total of 11,765 families with 4 or more sequences. We aligned the 11,765 gene families using MAFFT (*74*) (with the --auto option) and filtered with BMGE (*75*) (using bmge -t AA -m BLOSUM30) After filtering, 260 alignments contained no high-quality columns and were discarded. We filtered out sequences comprising more than 80% gaps to produce the final set of alignments. We also discarded all alignments with less than 30 columns, leaving a total of 11,272 families. The gene trees were computed using IQ-TREE v 1.6.10 using the following command:
iqtree -m TEST -s FAMXXX.faa.aln.trimmed -bb 10000 -wbtl -nt AUTO -madd LG4X,LG4M,LG+C10,LG+C20,LG+C30,LG+C40,LG+C50,LG+C60,C10,C20,C30,C40,C50,C60.
Conditional clade probabilities (CCPs) were computed using ALEobserve and the resulting ALE files were reconciled with the species tree. Loss rates were corrected by genome completeness, estimated using CheckM (*84*). We tested 62 roots (Online Data Supplement *(80)*).

Simulations to evaluate the performance of ALE

Performance of ALE for species tree rooting

The ability of the ALEml_undated algorithm to infer the correct gene tree root in the presence of gene duplications, transfers and losses was previously investigated using simulations (*1*). Briefly, gene families were simulated on a rooted species tree using a continuous-time origination, duplication, transfer and loss (ODTL) process (that is, a more complex model of genome evolution than that implemented in ALEml_undated), and ALEml_undated was used to estimate the root from subsamples of the simulated families. The maximum likelihood root according to ALE was the correct root in 95/100 replicates, and the log likelihood of alternative roots decreased with nodal distance from the correct root (as observed in our empirical data, see Fig. 1). In the remaining 5 cases, the maximum likelihood root

was one branch away from the true root. Analysis of empirical data suggested that ALE root inferences are robust to (that is, consistent across) subsets of the data that vary in terms of the rate of horizontal gene transfer or species representation in gene families (*1*). These properties make the ALE approach appropriate for inferring the root of Bacteria.

Accuracy of duplication, transfer and loss rates inferred by ALE

To test the accuracy of ALE at correctly inferring duplication, transfer and losses, we simulated a species tree of 265 leaves (the same size as the focal dataset) using Zombi (*78*). A simulation approach is necessary because, for empirical data, we do not know the true gene family history and so cannot evaluate method performance directly. The empirical realism of simulations is often an issue, and it is not always clear how best to accommodate the complexities of real genome evolution, including heterogeneity of DTL rates across families and, indeed, biases in rates (for instance, the high lineage-specific rate of gene loss that appears to characterise CPR). To make our simulations as realistic as possible, we first simulated the evolution of gene families using the Gm mode in Zombi, which assumes that every family has its own and independent rates of D, T and L, with the rates sampled from distinct gamma distributions (D~G(0.2,0.5), T~G(2,05), L~G(2.2, 0.5)). We simulated a total of 97929 families. We computed gene tree-species tree reconciliations with ALEml_undated for all families. Then, to ensure that the simulated dataset was as similar as possible to the real dataset, we sampled 2000 families at random from the real dataset. For each of those families, we selected the simulated family most similar in terms of DTL events (similarity was computed as one minus the squared sum of differences between the different inferred events by the reconciliations and the size of the families). This procedure resulted in a set of simulated families that were closely similar to the real data in terms of DTL events (fig. S3). These families recapitulate the gene family- and lineage-specific heterogeneity of DTL rates observed in the real data, and so provide the best possible basis for evaluating the performance of ALE. We used this set of simulated families in the analyses described below.

Comparison of the real (that is, simulated) and inferred numbers of D, T and L events on these data suggest that ALE accurately estimates the numbers of all three kinds of events (fig. S4), with mean errors close to 0 for all three types of events (D ~ 0.005, T ~ 0.048, L ~ 0.019). A detailed examination of the errors that do occur indicated that errors are most common in small families (fig. S5), and that the number of DTL events tends to be under-estimated when the true number of events is high (fig. S6). These observations motivate some of the sensitivity analyses of the empirical data described below, in which gene families with high inferred rates and, in a separate analysis, small sizes were excluded from the root calculation (see “Testing the robustness of the inferred root region” below); the Gracilicutes-Terrabacteria root was robust to all of these treatments.

Different combinations of DTL events can give rise to the same gene tree topology. For example, genes that are patchily distributed across species might be explained by a series of gene transfers, ancestral presence followed by independent losses, or a combination of processes. To investigate whether ALE can distinguish between

different kinds of DTL events based on gene tree topologies, we examined the correlations in inference errors for different kinds of events. Negative correlations (for example, over-estimation of transfer associated with under-estimation of losses) would suggest that the method can mistake one kind of DTL event for another. No correlations of this type were obtained (fig. S7), suggesting that ALE can distinguish the history of DTL events giving rise to a given gene tree topology. To further investigate whether ALE overestimates the number of gene transfers compared to duplications and losses, we specifically examined the inference results for the subset of simulated families with 0 transfers but one or more duplication and loss events (2429 of 97292 families). Of these 2,429 families, ALE correctly inferred that 2,332 (96%) had no transfers.

Next, we investigated whether biases in DTL rates across the tree - which result in variation in genome sizes, such as the small genomes of CPR - impact ALE inference accuracy. To do so, we performed two additional simulations: one in which gene family originations occur at random on the tree (which results in homogeneous simulated gene contents and genome sizes), and one in which gene originations were constrained to occur at the same points as they do in the real data. This latter simulation results in data that recapitulate the variation in gene content and genome size observed in the empirical data. We then compared inference accuracy on the two datasets (Fig. S8). The results suggest that genome size heterogeneity does not substantially affect inference accuracy, with all errors centred on 0 in both datasets (mean errors without heterogeneity: D ~ 0.00342, T ~ 0.0044, L ~ 0.0238; with heterogeneity: D ~ 0.00349, T ~ -0.0007, L ~ -0.0268).

Finally, we investigated the impact of lineage extinction on the accuracy of DTL estimates. In principle, ALE estimates ought to be robust to lineage extinction because gene acquisitions from extinct (or unsampled) lineages are accounted for in the method (*85*). To investigate, we used Zombi (*78*) to simulate 1000 species trees with 30 extant (sampled) taxa, with the speciation rate equal to the extinction rate. We performed simulations on species trees with relatively small numbers of tips because, since Zombi is a forward simulator, simulating trees with 265-341 tips in the context of high extinction rates is computationally intractable (that is, only a very small proportion of simulated trees will grow to have 100s of extant tips when the extinction rate is equal to or larger than the speciation rate). 100 gene families were simulated on each species tree; of these, 69921 had at least 4 surviving gene copies in the extant tip genomes and could be used for comparison of DTL inference accuracy. For each family, we calculated inference accuracy as (inferred number of events - simulated number of events)/family size, as above, and evaluated the relationship between the proportion of extinct lineages on the species tree and inference accuracy (fig. S2). We detected a statistically significant but quantitatively small impact of extinction on accuracy, with a higher proportion of extinct lineages corresponding to a slight increase in error (Correlation coefficients between accuracy and proportion of extinct lineages: -0.024 (D); 0.041 (T); -0.093 (L); fig. S2); errors remain centred on 0 even when almost all lineages had gone extinct.

Testing the robustness of the inferred root region

Simulations (see above) are the most direct way to evaluate the performance of ALE, because they provide a controlled situation in which we know the truth with certainty. However, real data are heterogeneous in ways that are difficult to recapitulate in simulations, particularly in terms of variation and biases in the rates of evolutionary processes (DTL, speciation, extinction and substitution rates) across the tree. We therefore performed a range of sensitivity analyses to evaluate the robustness of the inferred root region.

Distributions of DTL rates, rate ratios, and the impact of excluding gene families from the root calculation based on these and other criteria.

Firstly, we ranked gene families by inferred duplication, transfer and loss rates and rate ratios (figs S10-11), and performed a gene-filtering analysis (fig. S12) in which families at either end of the distribution were progressively removed and root likelihoods re-evaluated. This approach is analogous to fast site removal, in that families with very high or low rates may be difficult to model and so mislead inference.
Based on these rankings, we performed gene filtering analyses in which the highest ranked families were progressively removed and the difference in likelihoods between roots (ΔLL) re-evaluated (fig. S12A-S). The first of these plots (fig. S12B-C) illustrates the effect of progressively filtering out the most widely-distributed families (the metric is the number of species with at least one gene in the family). For each pair of plots, the bottom panel shows the threshold value corresponding to the percentile removed. For example, removing 5% of the families represented in the most species corresponds to a threshold of being represented in 150 species or more for MCL families. Note that these threshold plots can also be interpreted as the cumulative distribution of the ranking criteria, i.e. the above example implies that 5% of MCL families are represented on 150 or more genomes. The left hand side of each plot indicates the summed likelihood of each candidate root position (Fig. S12A) on all of the data; moving to the right along the x-axis illustrates how the summed likelihood for each root changes as families at the top end of the distribution are filtered out. Filtering out broadly-distributed families has a similar effect in both the MCL and Clusters of Orthologous Genes (COG) datasets: ΔLL starts to diminish. That is, ALE starts to lose the ability to distinguish between different root hypotheses. Conversely, when ranking families by the opposite criterion (the number of species absent, fig. S12D-E), the difference in likelihood between the roots remains unchanged until a very large fraction of families is excluded.

We next evaluated whether families with different estimated D, T or L rates agree on the optimal root region (that is, whether the root signal from families with high, moderate and low rates is consistent). To evaluate the signal in a root-independent manner, we calculated the mean per family D, T and L rates weighted according to the likelihood of that family for different roots. We can see in fig 12F-K that removing even a substantial (10-20%) fraction of families with the highest D, T or L rates does not change the order of likelihoods for different roots, and it is only after nearly half the families are discarded that we start to see a loss of resolution.

To filter families with potentially problematic rates, we also calculated each of the 6 possible rate ratios (D/T, T/D, D/L, L/D, T/L and L/T) and ranked families according to the maximum of these six ratios (fig. S12L-M). While our results remain unchanged as long as fewer than 10% of families with the most extreme rate ratios are removed, removing between 10 and 30% of the families with the most extreme rate ratios changed the order of the most likely roots while still retaining resolution (see for colors). This indicates that the Fusobacteriota root (dark green) derives its support, in part, from families that are outliers in terms of DTL rate ratios.

Performing further analogous threshold analyses (fig. S12N-S) in terms of mean copy number and verticality, we find that the rooting analysis is robust to filtering based on these criteria.

Effect of gene family clustering

We next evaluated the impact of gene family clustering on our analysis by repeating the entire analysis using COG (*39*) families; the same root region was recovered with the addition of one adjacent branch, with the reduced resolution likely due to the smaller size of the COG family dataset (3,723 vs 11,272 mcl families; see table S5).

Effect of taxon sampling

Taxon sampling is known to be an important factor in phylogenetic analyses (*22*), so we next evaluated whether our method of sampling taxa representatively from GTDB impacted our results. To do so, we repeated the entire analysis using a different approach, selecting representative taxa from major bacterial clades previously described in the literature *(11,15)*. We based our sampling on the analysis of (*11*), in which CPR comprise 40% of bacterial diversity. To do so, we inferred a tree of the bacterial portion of the published (*11*) concatenate under the LG+G4+F model in IQ-TREE. We divided the tree into 7 major bacterial clades based on a literature search (table S12) and additional environmental lineages with branch length diversity comparable to the known groups. For each group defined in this way, we manually subsampled taxa so as to maintain genetic diversity, while avoiding the longest and shortest branches. We sampled 341 species, comprising 200 ‘classic’ bacteria, 124 CPR bacteria and one bacterial genome respectively from each of the 17 new phyla described by (*14*); see table S9. We used the same marker gene set as in the focal analysis. A species tree was inferred in IQ-TREE using the LG+C20+G4 model with PMSF *(86)*. Additional trees were inferred in PhyloBayes under the CAT+GTR+G4 model using a recoded alignment using the four-category scheme of Susko and Roger *(87)*, and under the multispecies coalescent model in ASTRAL (*88*). To infer homologous gene families, we used the same pipeline as that used in the focal analysis. This resulted in 11,781 gene families with 4 or more sequences, which were analysed as in the focal analysis. The unrooted species trees are congruent with each other, except in the placement of a small group of phyla comprising Fusobacteriota, Aquificota, Synergistota, Spirochaetota and Thermtogota (“FASST”). These taxa are resolved in different positions in each of the three unrooted topologies, and are found to be monophyletic in the tree inferred from the recoded alignment. Similarly to the focal analysis, the ALE analyses yielded two

root positions that could not be rejected (AU test, $p > 0.05$), summarised in fig. S13. Both of these rooted phylogenies are congruent with that of the focal (GTDB) analysis, with Terrabacteria and Gracilicutes on either side of the root, with the only differences being in the placement of the FASST taxa. In the focal analysis, as well as the LG+PMSF+G4 and ASTRAL trees inferred as part of the GTDB-independent analysis, FASST were not recovered as a monophyletic group.

Inference of relative divergence times of bacterial clades

We applied the pipeline documented and implemented in the Online Data Supplement (*80*, RelativeDating.zip) to obtain the relative dated trees. We performed the analysis 5 times: 3 times in the focal dataset, and 2 times in the secondary dataset (using all the roots that could not be rejected). The pipeline consists of parsing the transfers inferred by ALEml_undated (using the MCL families) and discarding those with posterior probability < 0.05. We used bootstrapping to estimate constraint support in the following way: for each of the three branches in the root region, we sampled the gene families 100 times with replacement and, for each replicate, converted detected transfers to constraints and performed a Maximum Time Consistency (MaxTiC) analysis (*42*, *43)*. We then selected the phyla that were represented by 5 genomes or more in both datasets. We pruned the tree of the focal dataset to maintain only those phyla. Finally, we selected the constraints that were supported by both datasets and use those constraints to generate 1000 time orders compatible with those constraints the script order_explorer.py *(80)*). We then ranked all interior nodes on the tree, with the root node having rank 0 and the most recent speciation node having rank 11.

Quantifying vertical and horizontal signals in bacterial genome evolution

In the context of our analyses, “verticality” is the proportion of inferred evolutionary events on a branch of the rooted species tree that reflect vertical descent, estimated using gene tree-species tree reconciliation. We defined branch-wise verticality as V/(V+O+T), where V is the inferred number of vertical transmissions of a gene from the ancestral to descendant ends of the branch; O is the number of new gene originations on the branch; and T is the number of gene transfers into the branch. We defined transfer propensity as T/(V+T), where V and T refer to inferred numbers of events within the history of a gene family (table S11). The numbers reported in the main text have been averaged over the reconciliations obtained using the three possible roots of the focal analysis.

Ancestral gene content and metabolic reconstruction

Protein and protein family functional annotation

Protein sequences from all genomes used for phylogenetic analyses in this study were annotated using a variety of databases. Functional annotations were obtained using hmmsearch v3.1b2 (settings: -E 1e-5) (*89,90*) against KEGG Orthology (KO) annotations from the KEGG Automatic Annotation Server (KAAS; downloaded April 2019) (*91*). Additionally, all proteins were scanned for protein domains using InterProScan (v5.31-70.0; settings: --iprlookup --goterms) (*92*).

Multiple hits corresponding to the individual domains of a protein are reported using a custom script (parse_IPRdomains_vs2_GO_2.py). For the functional annotation of the 4256 COG families investigated in our ancestral reconstructions, we assigned KOs using a majority rule, i.e. we assigned the KO reported in the majority of sequences comprising each of the COG families yielding a COG-to-KO mapping file. Subsequently, we mapped COG descriptions, COG Process/Class, Category description, kegg id, kegg description, and kegg pathway to the COG-to-KO mapping file. COG descriptions were collected from the root annotations (1_annotations.tsv) downloaded at EggNOG (v5.0.0) (*93*). COG functional category and Process/Class descriptions were derived from eggNOG (v4.0) (*94*). KO pathways were manually curated based on an in-house KO-to-pathway mapping file, and were subsequently mapped to the respective KO. The scripts for annotation and mapping are included in the Online Data Supplement *(80)*.

COG gene families for ancestral gene content reconstruction

We built a set of gene families based on the COG (*95*) database for ancestral functional inference. To do so, we annotated each genome in the dataset using eggNOG-mapper v2(*79*), then clustered proteins into families based on their COG annotations. For proteins annotated with more than one COG category (8% of proteins), we included the protein in both COG families. This resulted in 4256 COG families, of which 3723 had 4 or more sequences. COG families are ideal for ancestral reconstruction because they comprise all of the sequences on extant genomes that can be annotated with a given unambiguous function from the COG ontology. In addition, the hierarchical nature of the COG classification (comprising gene family annotations nested within 23 broader functional categories) enabled us to explicitly model the different evolutionary ages of gene functional classes as part of the analysis, by using category-specific root origination priors (see below).

Our COG families are useful for functional reconstruction, but are perhaps less well suited for investigating other aspects of bacterial evolution because they are constructed only from proteins that could be annotated with eggNOG-mapper. By contrast, MCL families represent --- within the limitations of the clustering approach, as discussed above --- an unbiased view of gene family diversity for the set of genomes we analyzed. We therefore base analyses other than those regarding the functional annotation of LBCA on the MCL families. However, since gene clustering methods are not consummate and each has strengths and weaknesses, we also investigated the root signal from the COG families. This analysis identified a similar root region of four adjacent branches, comprising the root region from the focal analysis (3 branches) plus one additional root, in which Spirochaetota branched on the Terrabacteria side of the root (table S5).

Root gene mapping approach

To estimate root presence posterior probabilities (PPs) for each gene family for each of the three supported roots, we first estimated the probability of origination at the root (O_R) by maximum likelihood, finding the O_R value that maximises the total reconciliation likelihood summed over all gene families (table S11). We then used the global ML O_R value to calculate the root presence posterior probabilities

for each family; that is, the probability that one or more copies of a given gene family were present at the root, given the ML O_R value. We estimated root origination rates independently for each of the 23 COG functional categories, and used these rates to estimate the posterior probability of presence at the root node for each gene family. Note that, for all nodes of the tree (including the root nodes), we additionally estimated PPs directly from the sampled reconciliations. Python code implementing this procedure is provided in (*80*) at Code/O_R_Optimization.py.

Initial gene content and metabolic inferences at a particular node were based on gene families with a posterior presence probability (PP) of >0.95 at that node. This approach is conservative and could miss the presence of certain pathways which may be represented by proteins with a range of PP values. Therefore, we manually investigated the PPs of pathways discussed in this manuscript and inferred the presence of specific pathways or functional modules if the majority of the components were found with PP >0.50, as described in the main text, Fig. 4 and fig. S17.

Impact of root branch on LBCA gene content

The credible set of root branches from the ALE analysis comprised three adjacent branches at the centre of the tree (Fig. 1b). The difference between these three root positions relates to the placement of Fusobacteria, either as the root branch or as the most basal split on either the Gracilicutes or Terrabacteria+DST “sides” of the rooted tree. We therefore estimated root PPs for COG families on all three branches; root PPs under all three roots are provided in the Online Data Supplement *(80)*.

Metabolic comparisons

Results from the PP analysis were used as the framework for metabolic comparisons and reconstruction of the proteome of LBCA. First, the occurrence of an individual COG family across each taxon was counted in R (v3.6.3) (table S4). This binary presence/absence matrix was combined with the PP values for Nodes corresponding to the CPR, Chloroflexota+CPR, Chloroflexota, Terrabacteria, DST+Terrabacteria, Gracilicutes-Spirochaetota, Gracilicutes+Spirochaetota, Root 1, Root 2, and Root 3, filtered with a cutoff of PP>0.50. The combined count table was summarized using the ddply function of the plyr package (v1.8.4), which was used to summarize the counts across each phylogenetic cluster, node, and root. Data is visualized in a heatmap generated using the ggplot function with geom_tile and facet_grid of the ggplot2 package (v3.2.0). Heatmap categories for pathways were scaled based on the number of COG families, results were plotted using the grid.draw function of the grid package (v3.6.3). Heatmaps were manually merged with a species tree in Adobe Illustrator (v22.0.1).

Evaluating the robustness of the LBCA reconstruction to taxon sampling based on the secondary dataset

To evaluate the impact of taxon sampling, we analyzed both the primary and secondary datasets with the same ancestral reconstruction pipeline. Across all

protein families, agreement between the datasets was highly significant, though lower than among the root regions for each dataset (table S13). The PPs between the focal and secondary analysis were significantly correlated independent of rooting (Pearson's correlation 0.48-0.53 $p<10^{-16}$, see table S13). PPs between root positions in the context of the same analysis were very strongly correlated, with the highest correlation between root 1 and root 2 of the secondary analysis (0.96) and roots 1 and 2 of the primary analysis (0.94). Overall, root 3 of the focal analysis (corresponding to the Fusobacteriota root) correlated the least with other roots. The focal analysis compared to the secondary analysis considered fewer taxa (265 vs 341) and correspondingly fewer COGs (3782 vs 4220 with four or more genes). The focal analysis also recovered fewer COGs with high confidence (PP>0.9) at the root (see gray diagonal fields in table S13). Of the COGs recovered with high confidence at the root in the focal analysis, the majority (63%-78%) were also recovered with high confidence at the root in the secondary analysis (see gray off-diagonal fields in table S13). Considering COGs recovered with high confidence at the root, root 3 of the focal analysis is again the least congruent with the other gene sets, and at the same time is the one with the smallest number of COGs recovered at the root. Importantly, all of the ancestral pathways discussed in the main text were recovered at the root with moderate to high PP support in both datasets, as indicated (table S7).

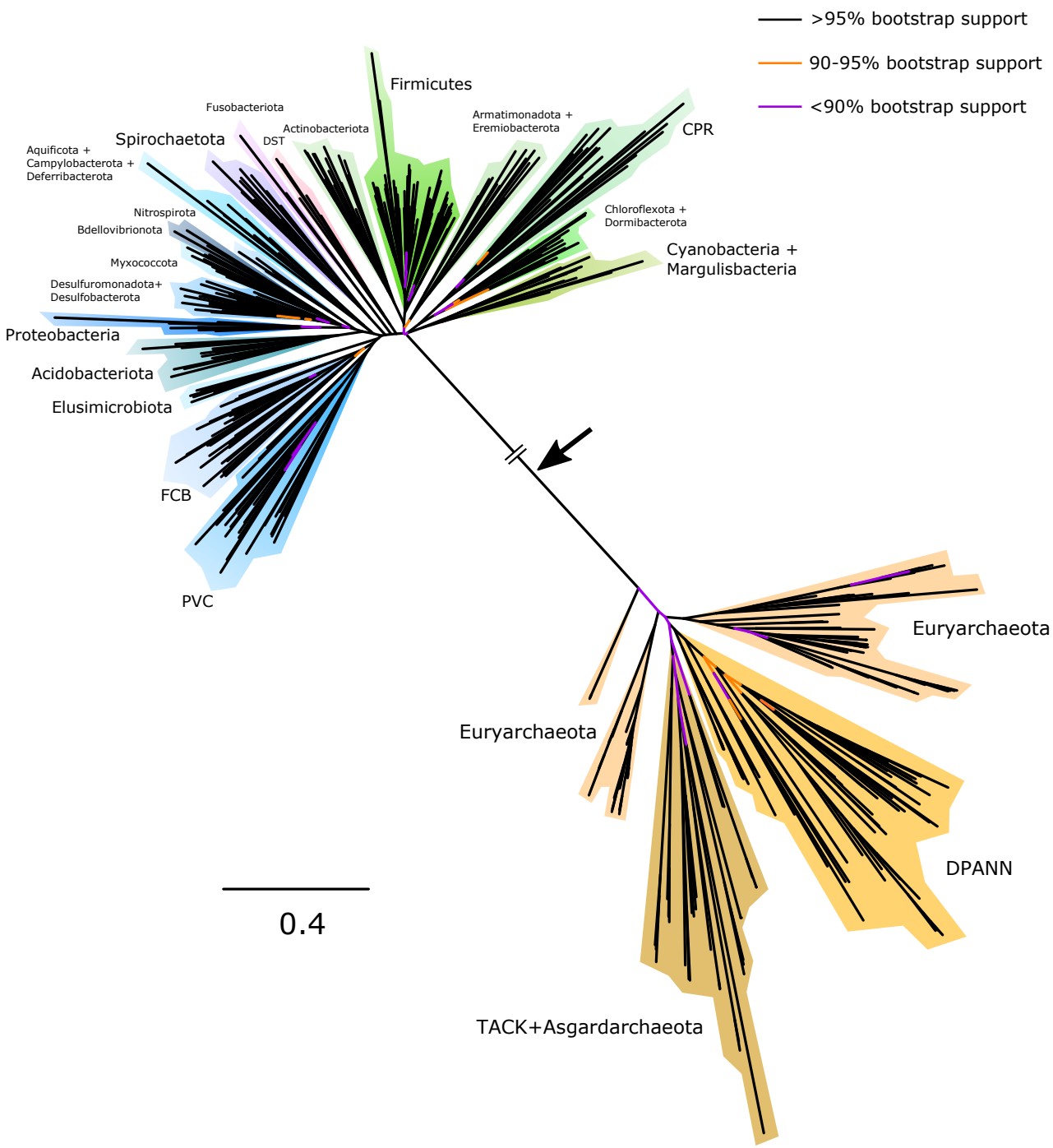


Fig. S1: Maximum likelihood outgroup-rooted bacterial phylogeny. The maximum likelihood phylogeny obtained under the best-fitting LG+C60+R8+F model on a concatenation of 30 marker genes shared between Bacteria and Archaea. The bacterial root (marked by a black arrow) separates CPR, Cyanobacteria +Margulisbacteria, and Chloroflexota+Dormibacterota from the rest of the bacterial tree, but this position has poor bootstrap support and a range of alternative hypotheses could not be rejected statistically; note also that a basal position for DPANN within Archaea (*1*, *81*) could not be rejected using an Approximately Unbiased (AU) test (table S2). FCB are the Fibrobacterota, Chlorobia, Bacteroidota and related lineages; PVC are the Planctomycetota, Verrucomicrobiota, Chlamydiia and related lineages; DST are the Deinococcota, Synergistota, and Thermotogota; ACD are Aquificota, Campylobacterota, and Deferribacterota; FA are Firmicutes and Actinobacteriota. Branch supports are ultrafast bootstraps, as indicated by the colour key. Branch lengths are proportional to the expected number of substitutions per site.

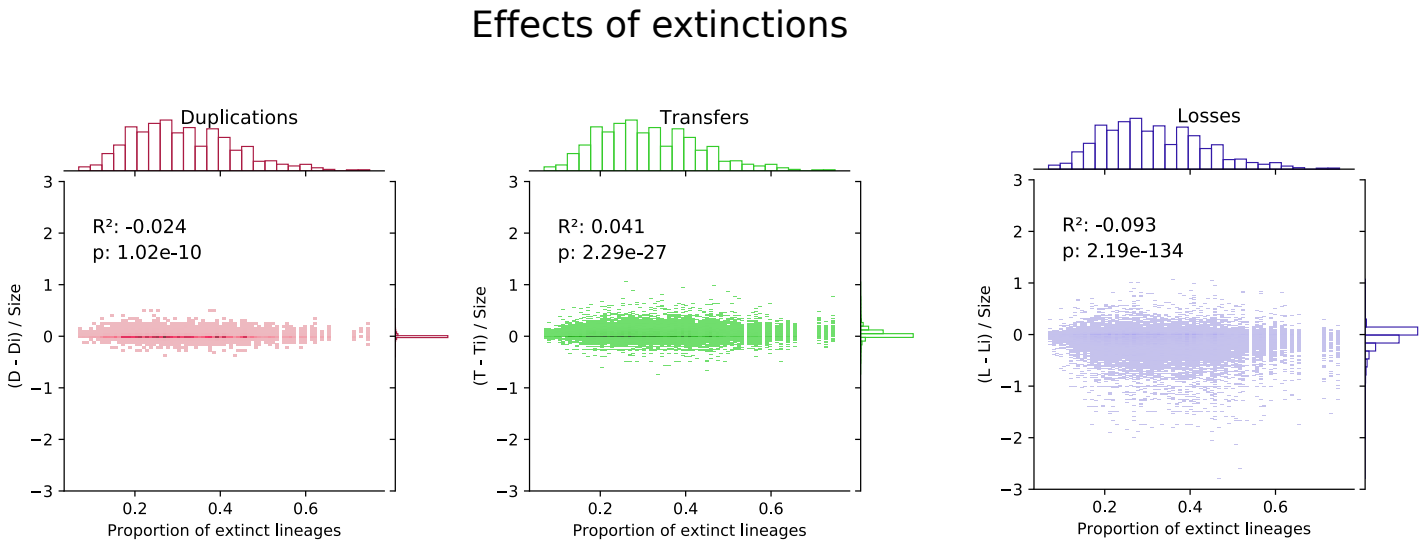


Fig. S2: Lineage extinction has a negligible impact on the accuracy of DTL estimates using ALE. We simulated species trees and gene families in the context of varying levels of lineage extinction, then evaluated the correlation between inference error (inferred - simulated numbers of events / family size). There is a statistically significant but quantitatively small impact of extinction on inference accuracy, with higher proportions of extinct lineages resulting in increased inference error (Pearson's r; values and p-values indicated in figure). Inference errors remain centred on 0 even when most lineages are extinct.

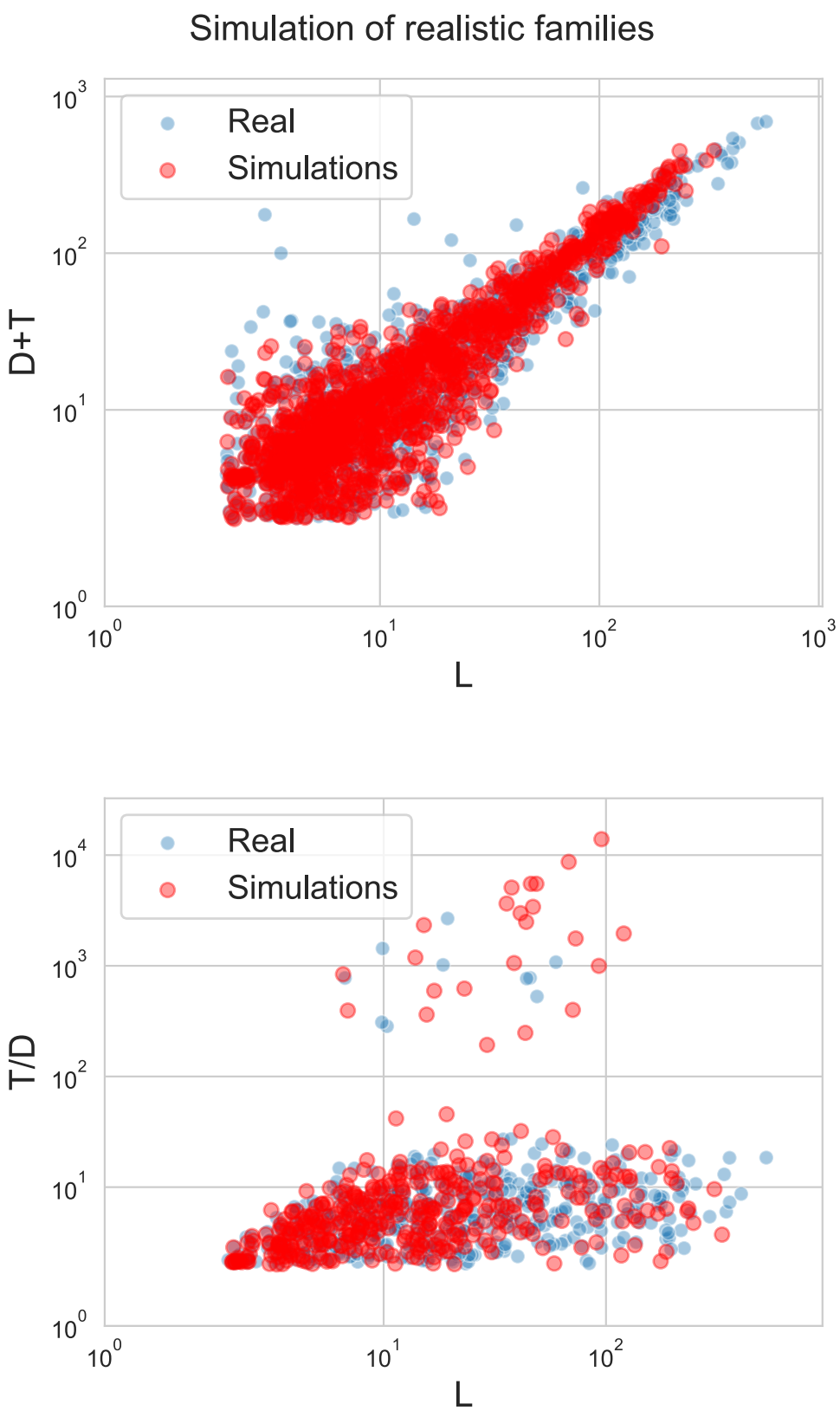


Fig. S3: Simulating gene families with numbers of DTL events similar to the real data. To align our simulations with empirical reality, we simulated a large number of gene families and then subsampled families for analysis by matching them against a random subsample of the empirical families (see Supplementary Materials and Methods). The result was a simulated dataset with distributions of DTL events close to the real data. On top, the gain events (y axis) against the loss events for simulated and real families. On the bottom, the ratio of T/D against losses. The simulated families resemble the real families in terms of DTL rates and rate ratios.

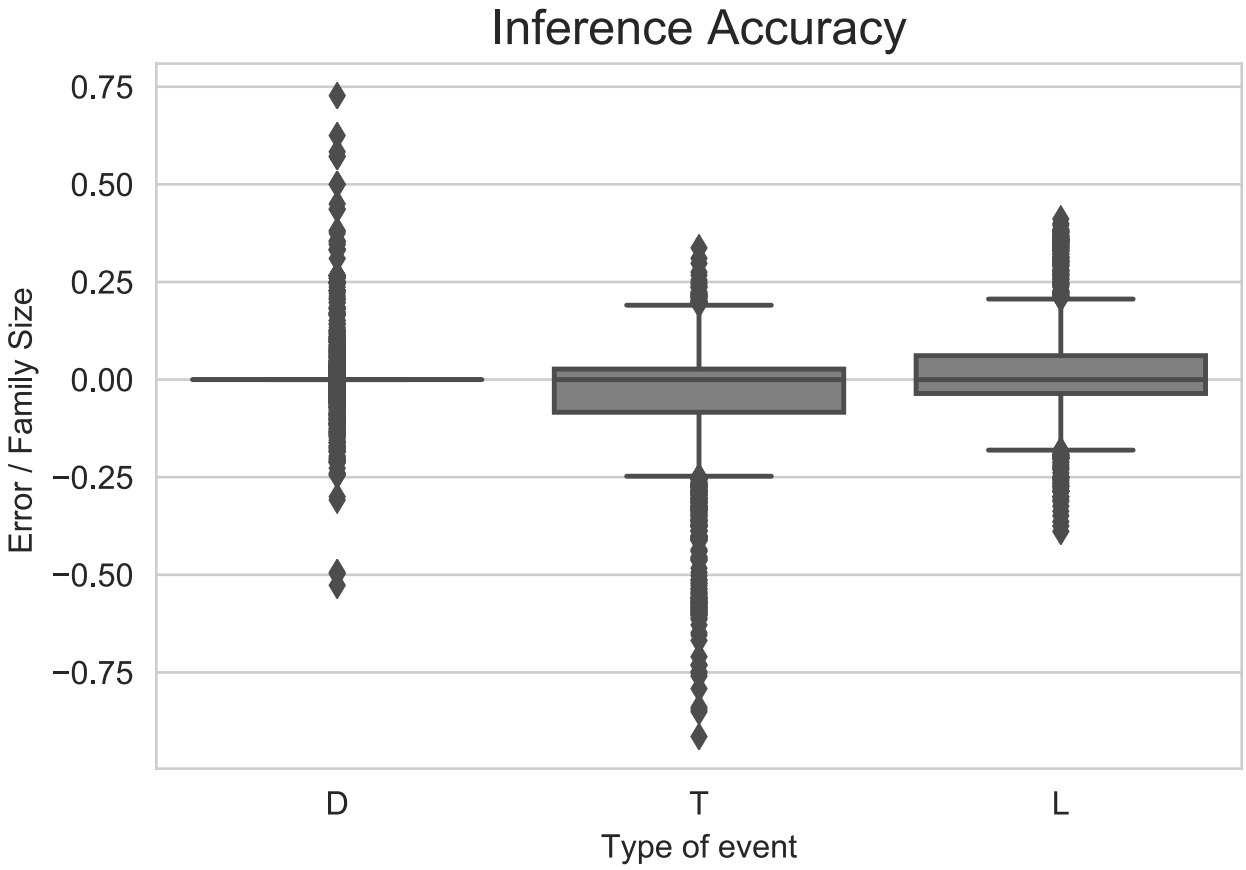


Fig. S4: Accuracy of ALE inference of duplication (D), transfer (T) and loss (L) events. The plot shows the accuracy ((inferred number of events - real number of events) / family size) of ALE inferences on a set of 2000 simulated families with observed numbers of DTL events similar to the real data (see fig. S3). ALE accurately estimates the correct number of DTL events for simulated families, with mean errors of all three types of events close to zero (D ~ -0.002, T ~-0.0210, L ~ 0.048).

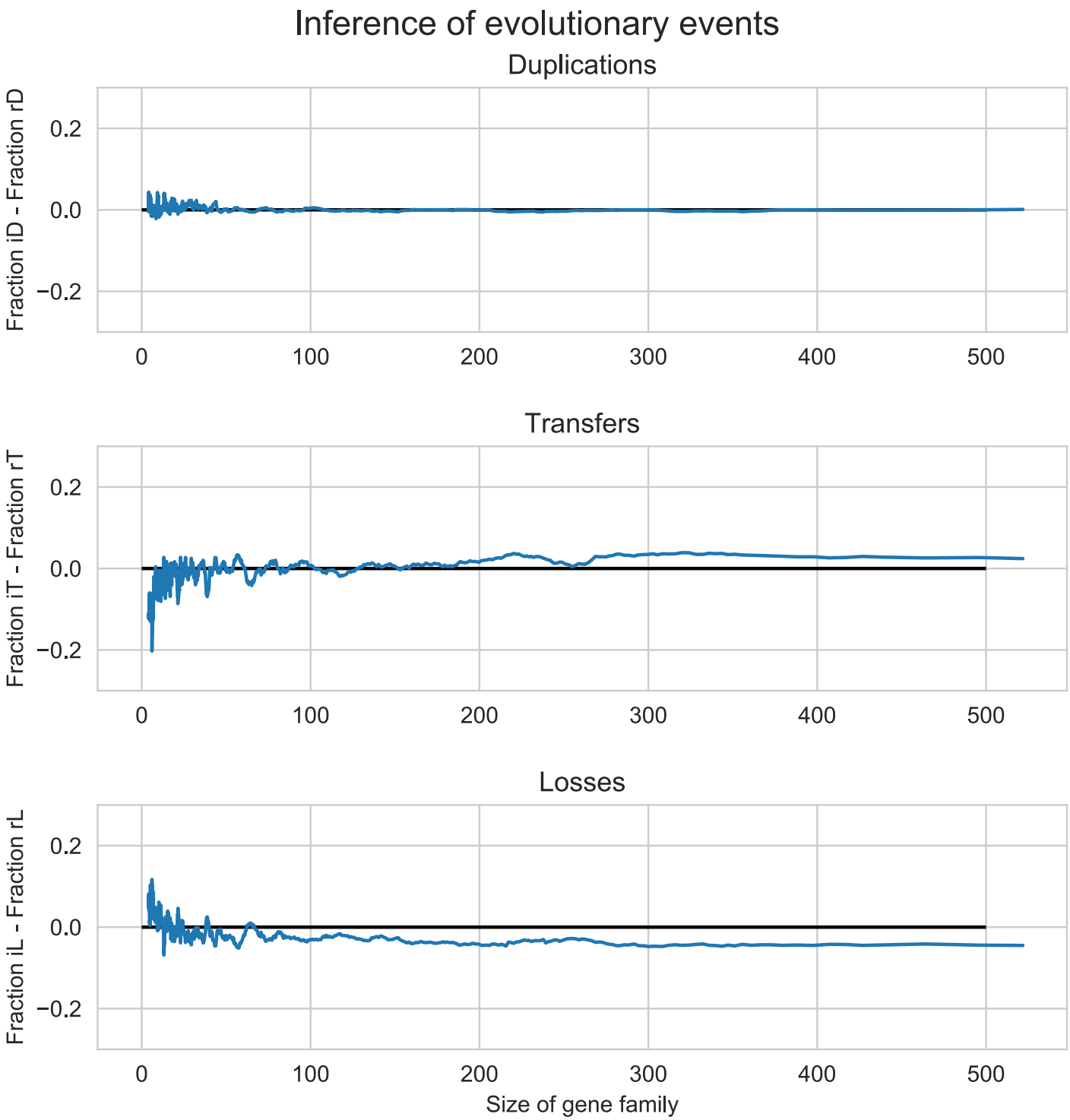


Fig. S5: Inference errors are most common in small gene families. We ordered the families by number of genes and computed the rolling mean of the difference between the real (simulated) fraction of events and the fraction inferred by ALE. Most inference errors occur in small families.

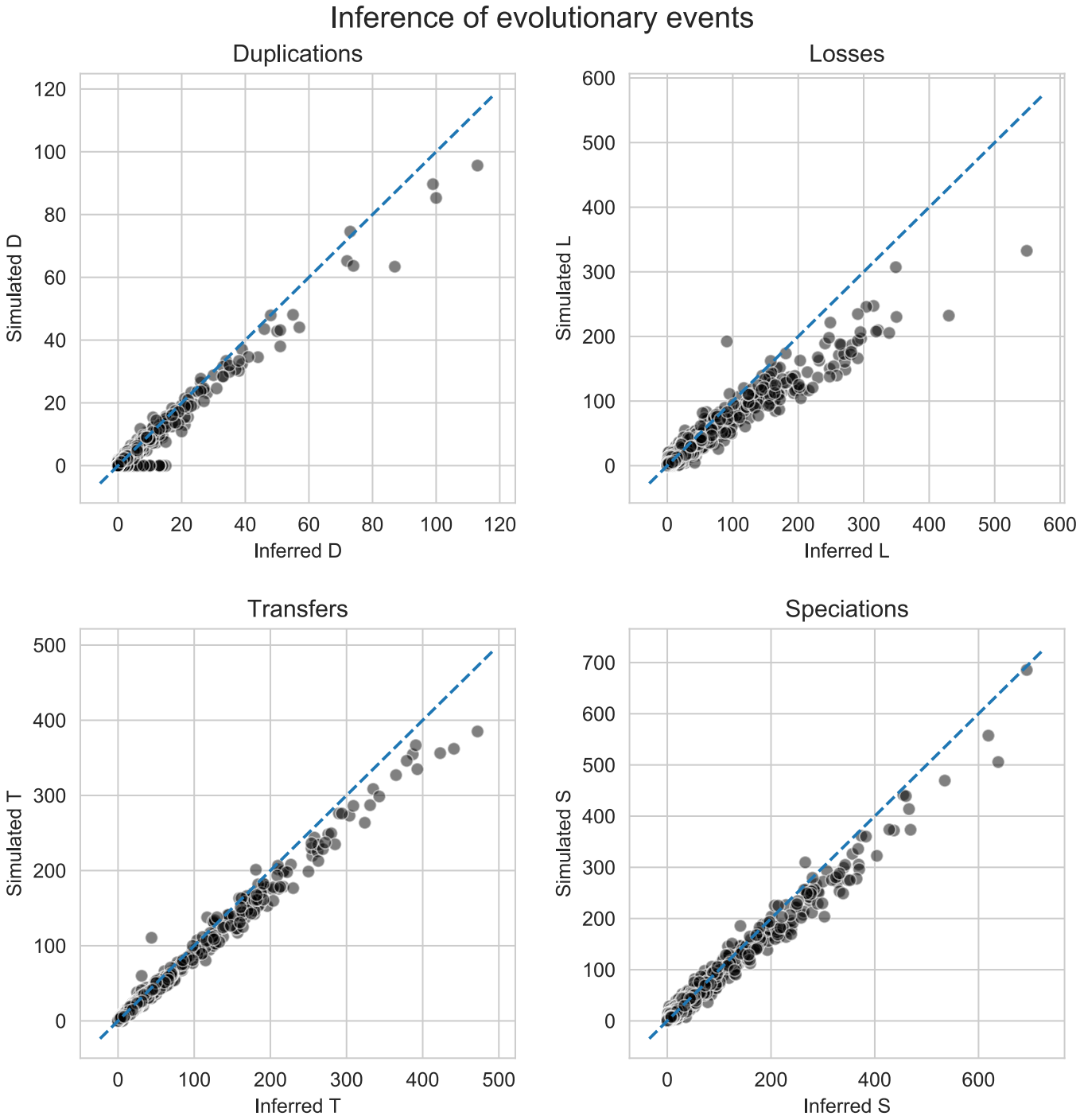


Fig. S6: Performance of ALEml_undated on simulated datasets. These scatterplots visualise the relationship between inferred and inferable numbers of duplications, transfers, losses and speciations on simulated gene families. Inferable events are events that leave traces in gene trees; for example, gene losses in extinct lineages cannot be inferred from gene trees. “Speciations” refer to the number of speciation events that a gene lineage passes through. The dotted lines correspond to perfect accuracy. ALE accurately recovers all types of events over a broad range of rates, although the inferred number of events is underestimated when the true number of events is high. This is most evident for losses, which may be because multiple recent losses can be misinterpreted as a single loss in the common ancestor of the affected branches.

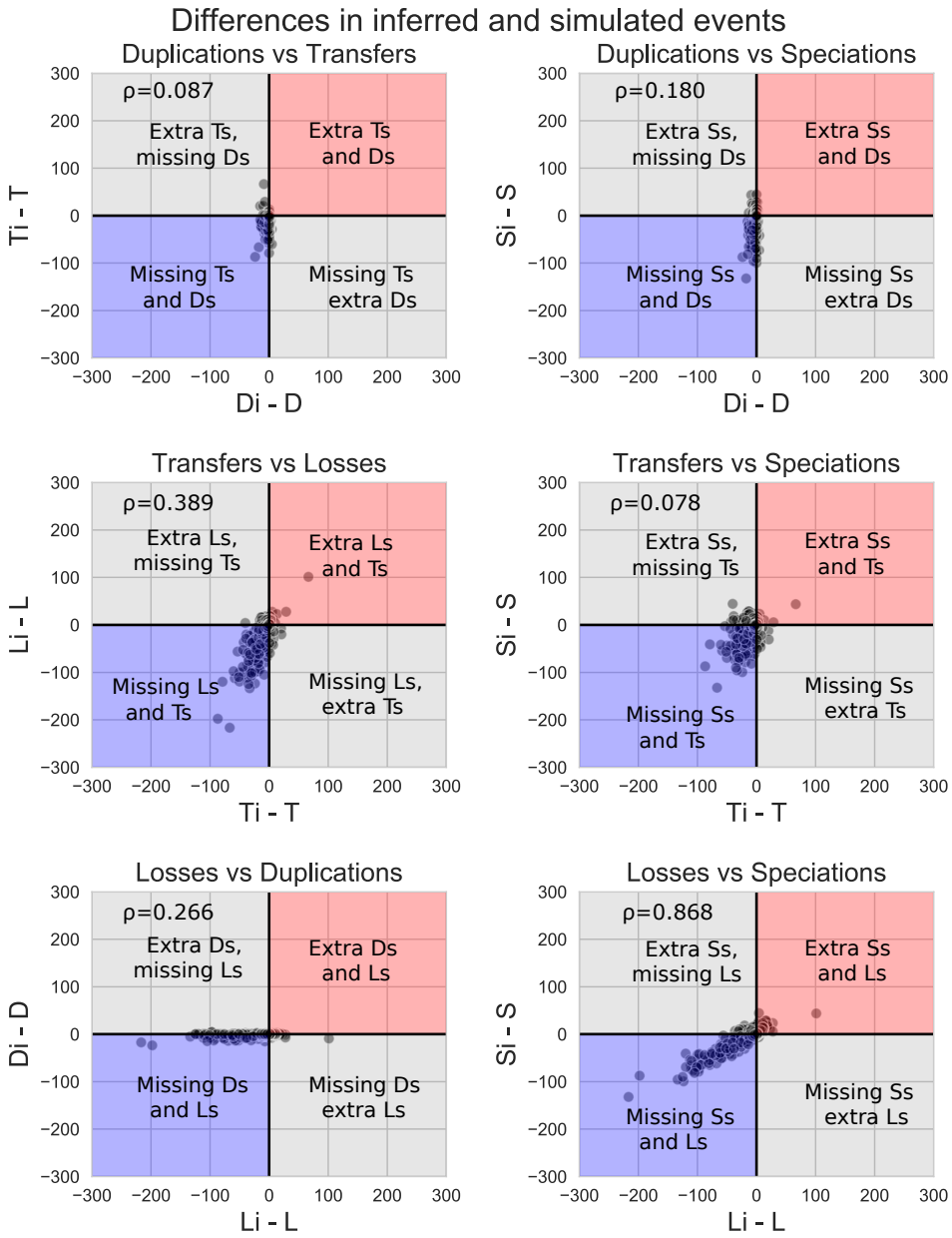


Fig. S7: Comparison of errors in inferred DTL events from simulated data suggests that ALE can distinguish duplications, transfers and losses from their impact on gene tree topologies. We investigated the relationship between inference errors for duplications (D), transfers (T), losses (L) and speciations (S), with error calculated as the difference between the inferred and true (simulated) number of events per gene family. A negative correlation between errors (for example, over-estimation of the number of transfers coupled with under-estimation of duplications or losses) would indicate that ALE tends to mistake one kind of event for another. There is no evidence of negative correlation between any types of events (all correlations were weakly to moderately positive, indicating that co-underestimation is the most common type of error, as observed for the independent events in fig. S6 above), suggesting ALE can distinguish events based on the patterns they leave in gene trees. As expected, errors in speciations and losses are positively correlated: if a gene is inferred to have originated more recently in the species tree than is in fact the case (that is, if the number of speciations it has experienced are underestimated), then the number of losses required to explain the gene tree will also be underestimated.

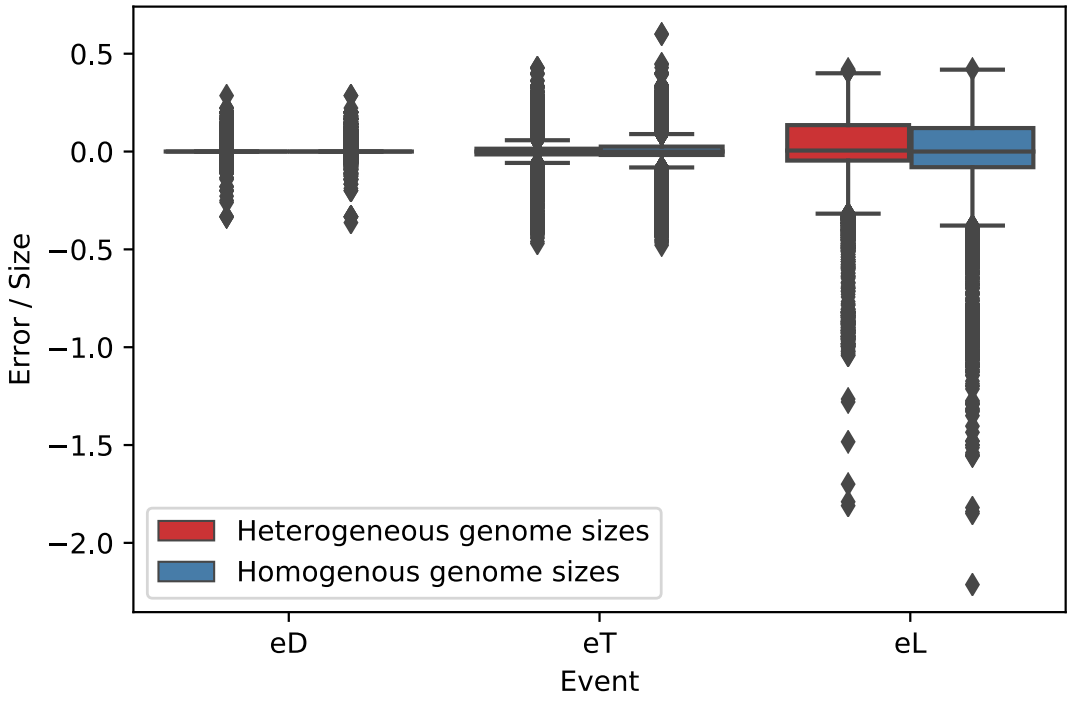


Fig. S8: Impact of genome size heterogeneity (variation in DTL rates across the tree) on ALE inference accuracy. To investigate the effect of lineage-specific biases in DTL rates on ALE inference accuracy, we performed a simulation in which gene family originations were constrained to occur at the same nodes as in the real data. This procedure results in a set of simulated gene families that recapitulate the variation in genome size and gene content observed in the real data (for example, with a clade of small-genome CPR). We then compared inference accuracy (simulated - inferred events / family size) for D, T and L rates to a control simulation in which originations occurred at random, resulting in homogeneous genome sizes and gene contents across the tree. The impact of genome size heterogeneity is negligible, and all errors in both settings were centred on 0.

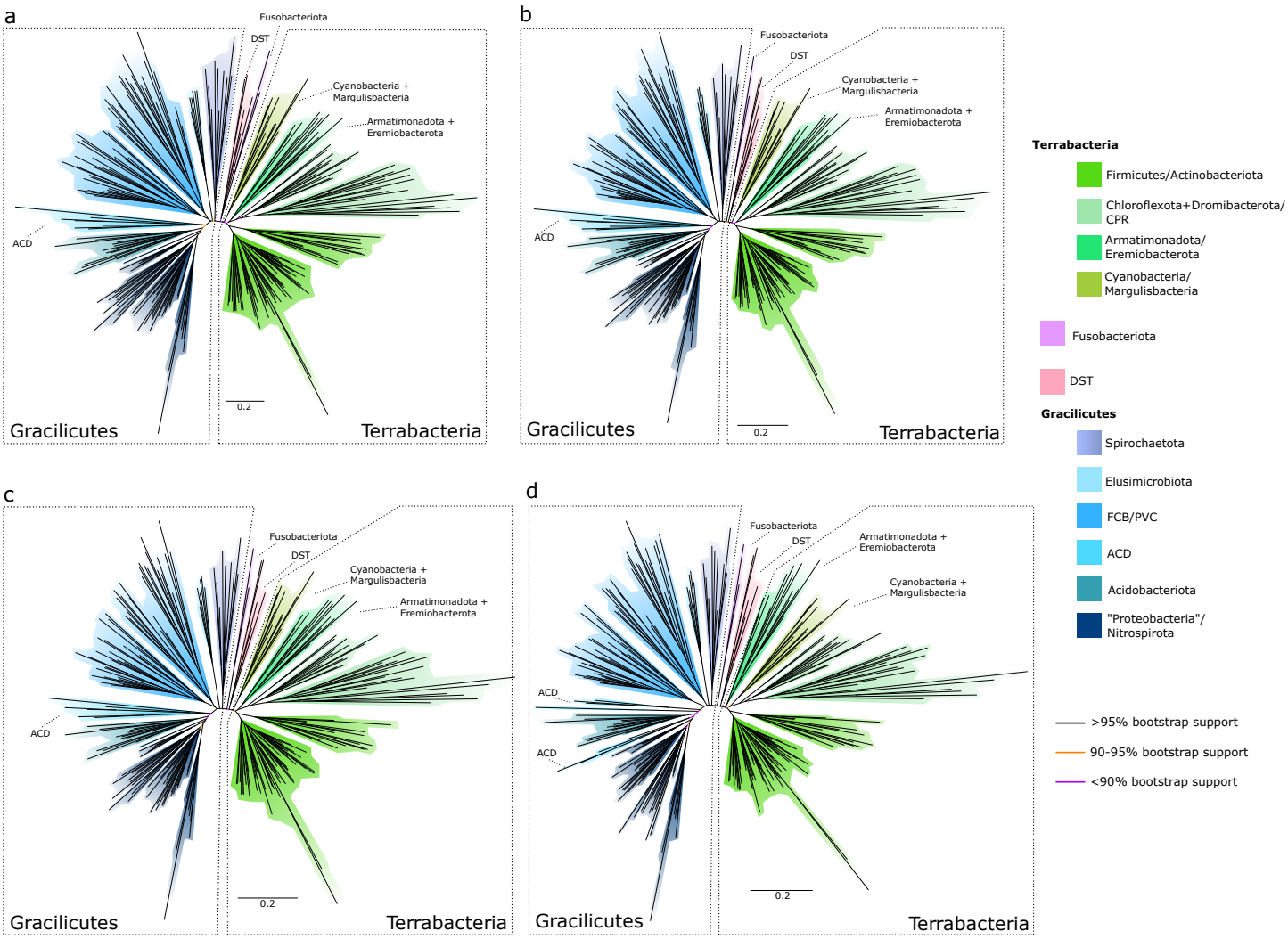


Fig. S9: Maximum likelihood unrooted bacterial phylogeny under the best-fitting substitution model (LG+C60+R8+F) following removal of the 20%-80% most compositionally heterogeneous sites. While the trees were similar overall, the position of Fusobacteriota was unstable, either branching as in the focal tree (40%, 60%, 80% of most compositionally heterogeneous site removed) or with (DST (20% of sites removed). Sites were identified and removed using Alignment Pruner. (a) 20% most compositionally heterogeneous removed, with 14580/18234 sites remaining following site stripping; (b) 40% most compositionally heterogeneous removed, with 10941/18234 sites remaining following site stripping; (c) 60% most compositionally heterogeneous removed, with 7294/18234 sites remaining following site stripping; (d) 80% most compositionally heterogeneous removed, with 3647/18234 sites remaining following site stripping; Branch supports are ultrafast bootstraps, branch lengths are proportional to the expected number of substitutions per site.

a)

b)

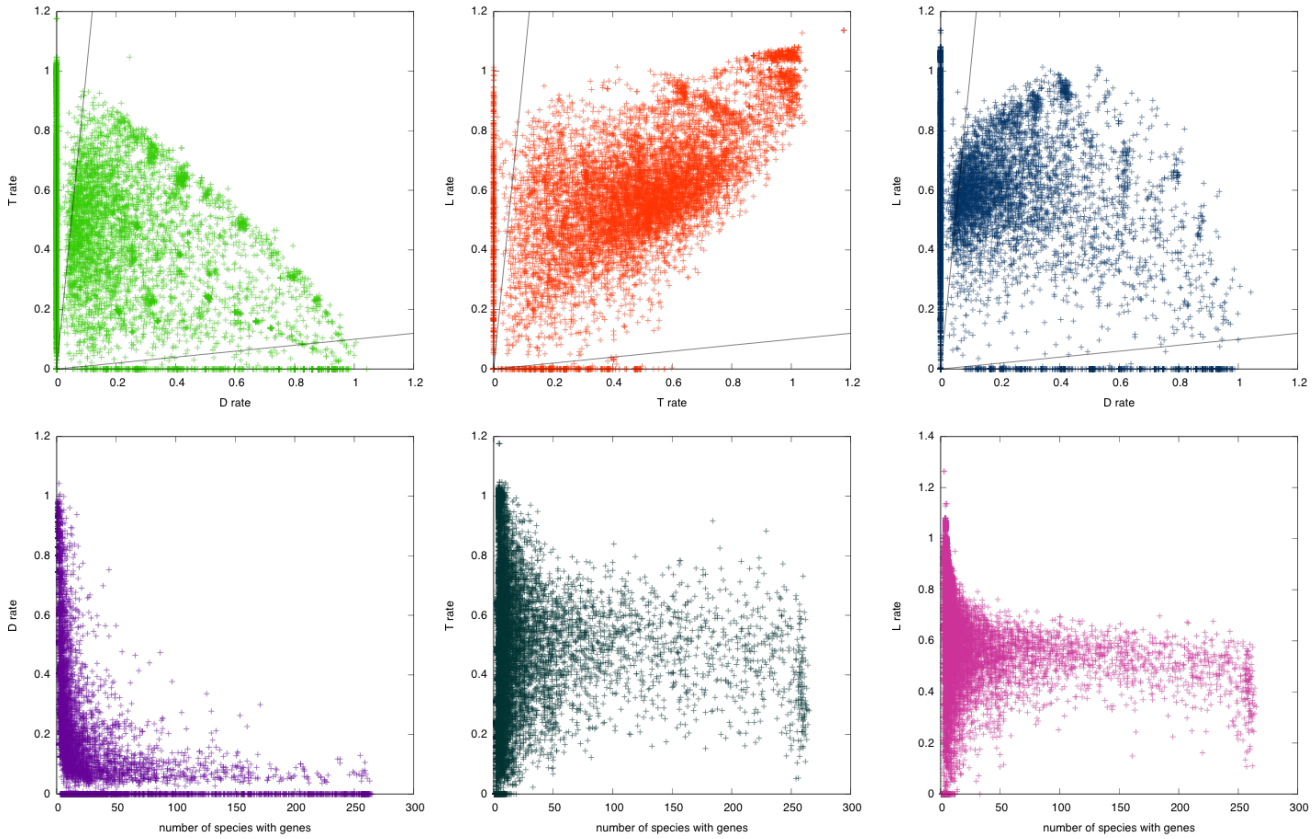


Fig. S10. Duplication (D), transfer (T) and loss (L) rates for MCL families. a) scatter plots of T vs D, L vs T and L vs D rates against each other, with ratios above ten falling outside (below and to the left) of the continuous black lines. The plots show that extreme rate ratios typically result from very low rates. b) plotting per-family DTL rates versus the number of species with genes shows that, with the exception of the D rate, very small rates are only inferred for gene families with very limited species representation. This result implies that there exist some broadly-distributed gene families that do not fix duplicates, whereas broadly-distributed families typically undergo at least some transfers and losses.

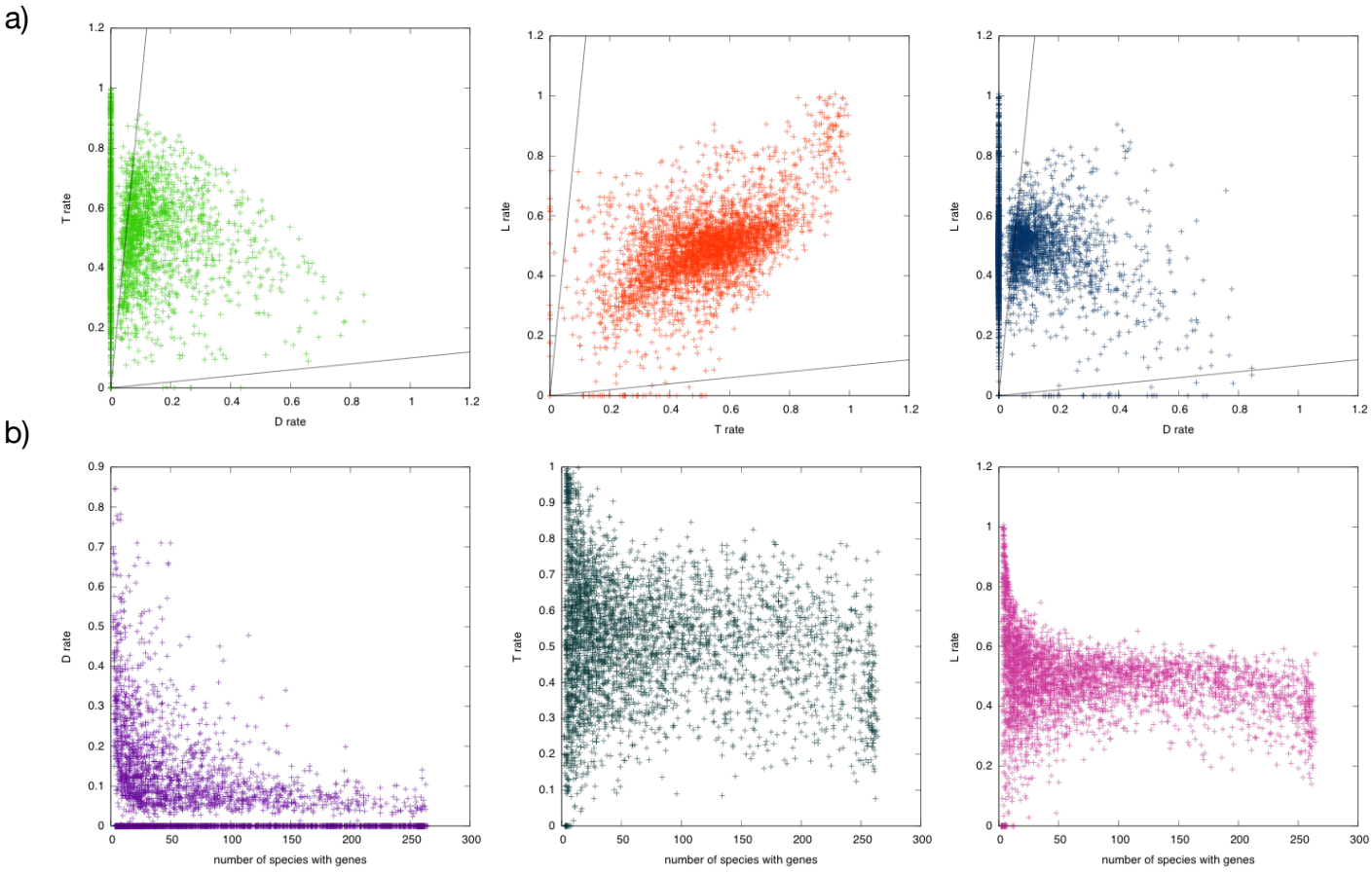


Fig. S11. Duplication (D), transfer (T) and loss (L) rates for COG families. a) Scatter plots of T vs D, L vs T and L vs D rates against each other, with ratios above ten falling outside (below and to the left) of the continuous black lines. The plots show that extreme rate ratios typically result from very low rates. b) Plotting per-family DTL rates versus the number of species with genes shows that, with the exception of the D rate, very small rates are only inferred for gene families with very limited species representation.

Figure S12. Impact of gene family filtering on ALE root likelihoods. We ranked gene families according to various criteria, then performed analyses to evaluate the impact on root likelihoods of excluding (filtering out) an increasing proportion of the top-ranked gene families for each criterion. The procedure represents a phylogenomic equivalent of site filtering in traditional phylogenetics, and is a useful way to explore the kinds of gene families that support each root position. Each analysis was performed on both the MCL and COG gene family datasets. For each analysis, the top panel plots ΔLL, the difference in log likelihood between a given root and the maximum likelihood root, as a function of filtered gene families. Positive values of ΔLL indicate that a given root has a better likelihood than the overall ML root at that percentile. The second panel shows the distribution of that test quantity, and the value of the quantity at each percentile of gene families filtered. Panel (a) maps the colours used in subsequent panels to the position of that root on a schematic unrooted tree; for example, the root between Gracilicutes and the rest of Bacteria corresponds to the light green line in panels (b)-(s). Removal of gene families by (b-c) species representation (families present in the most species filtered first); (d-e) species representation (families present in the fewest species filtered first); (f-g) highest copy number; (h-i) highest duplication (D) rate; (j-k) highest transfer (T) rate; (l-m) highest loss (L) rate; (n-o) highest rate ratio; (p-q) highest verticality; (r-s) lowest verticality; (t-u) highest transfer/loss (T/L) rate ratio.

Fig. S12(a). Mapping of root likelihoods to branches on the bacterial tree.

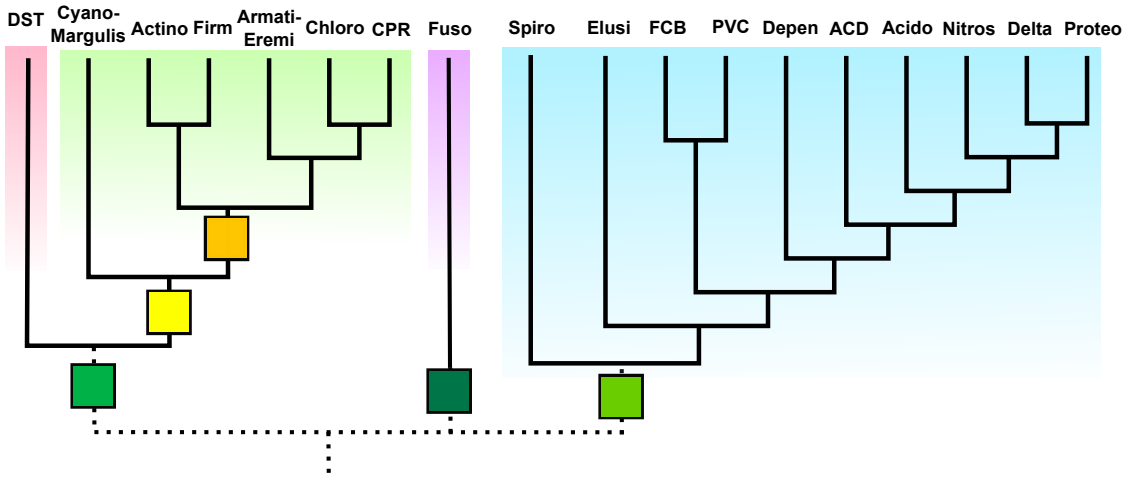


Fig. S12(b). Filtering MCL families ranked by representation in largest number of species

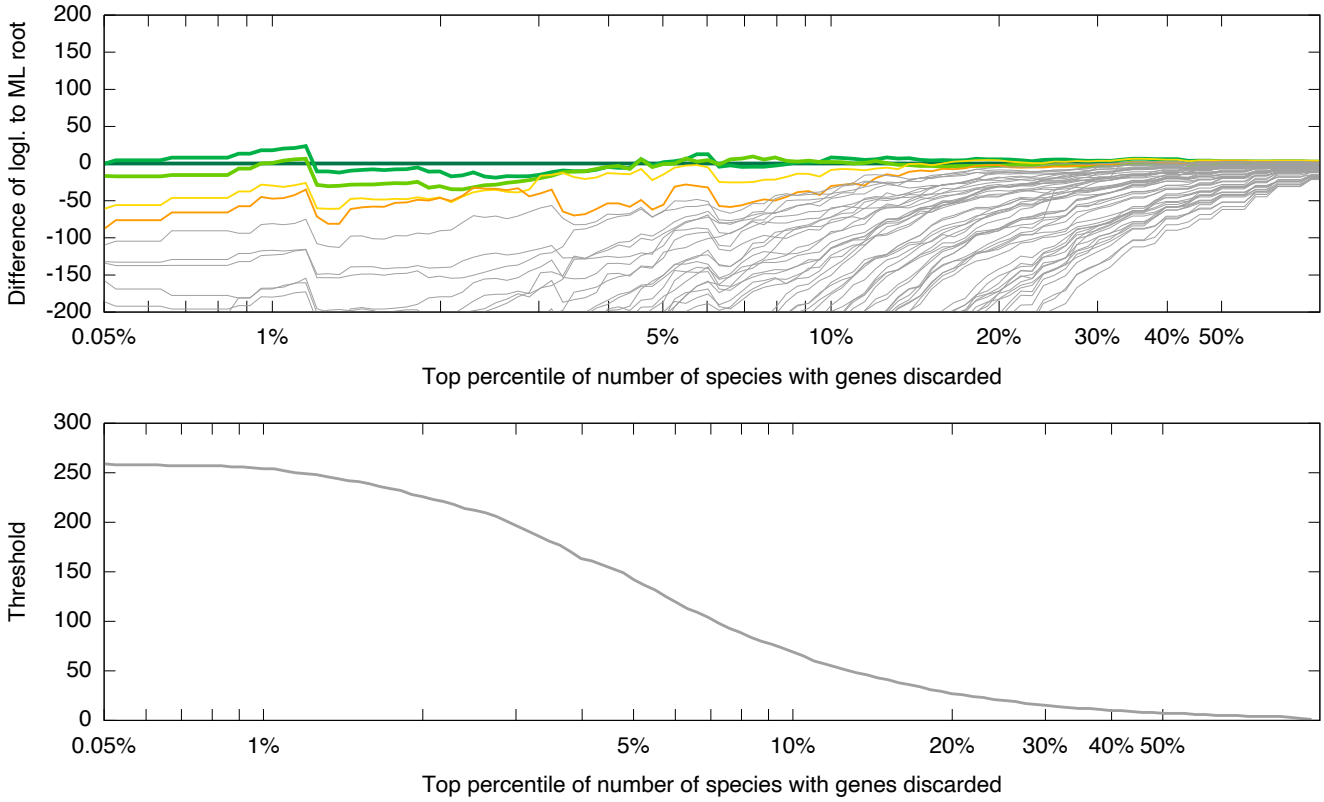


Fig. S12(c). Filtering COG families ranked by representation in largest number of species

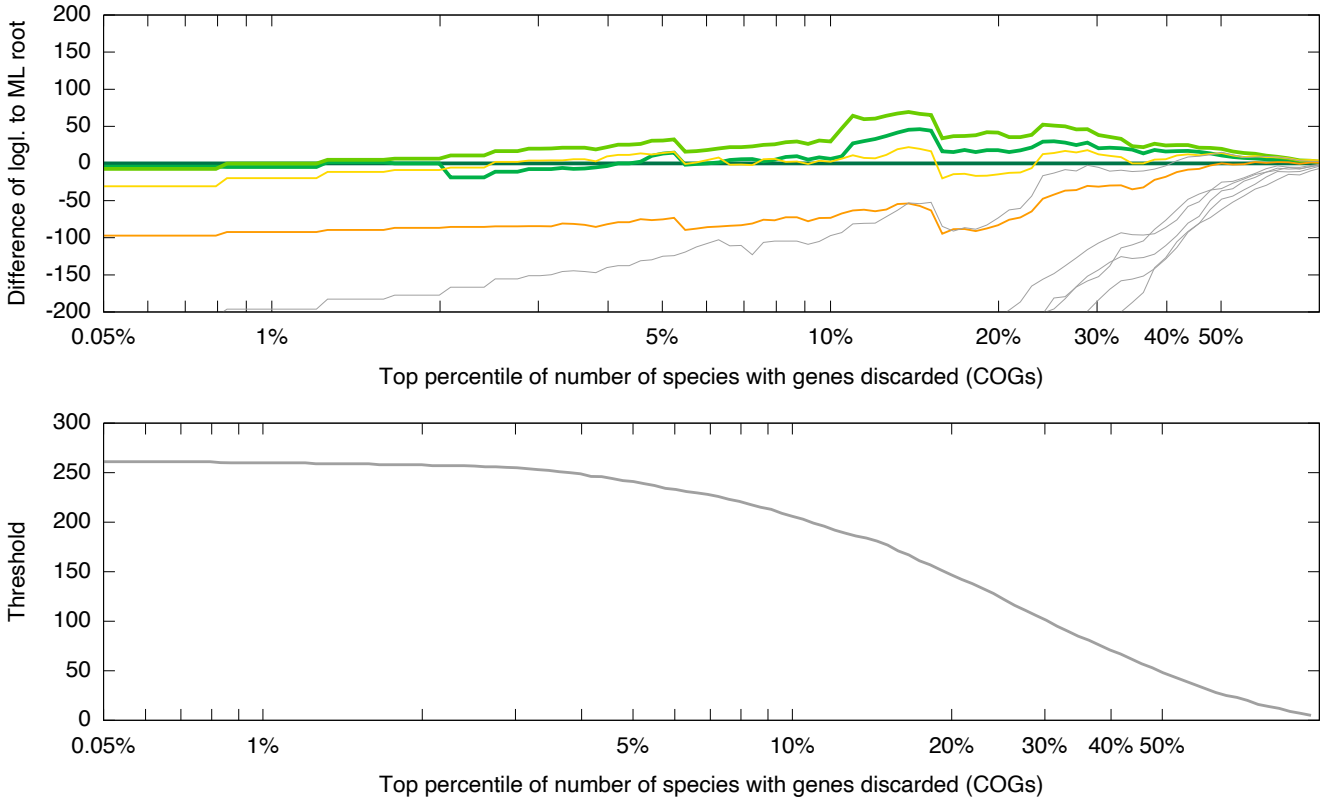


Fig. S12(d). Filtering MCL families ranked by representation in smallest number of species

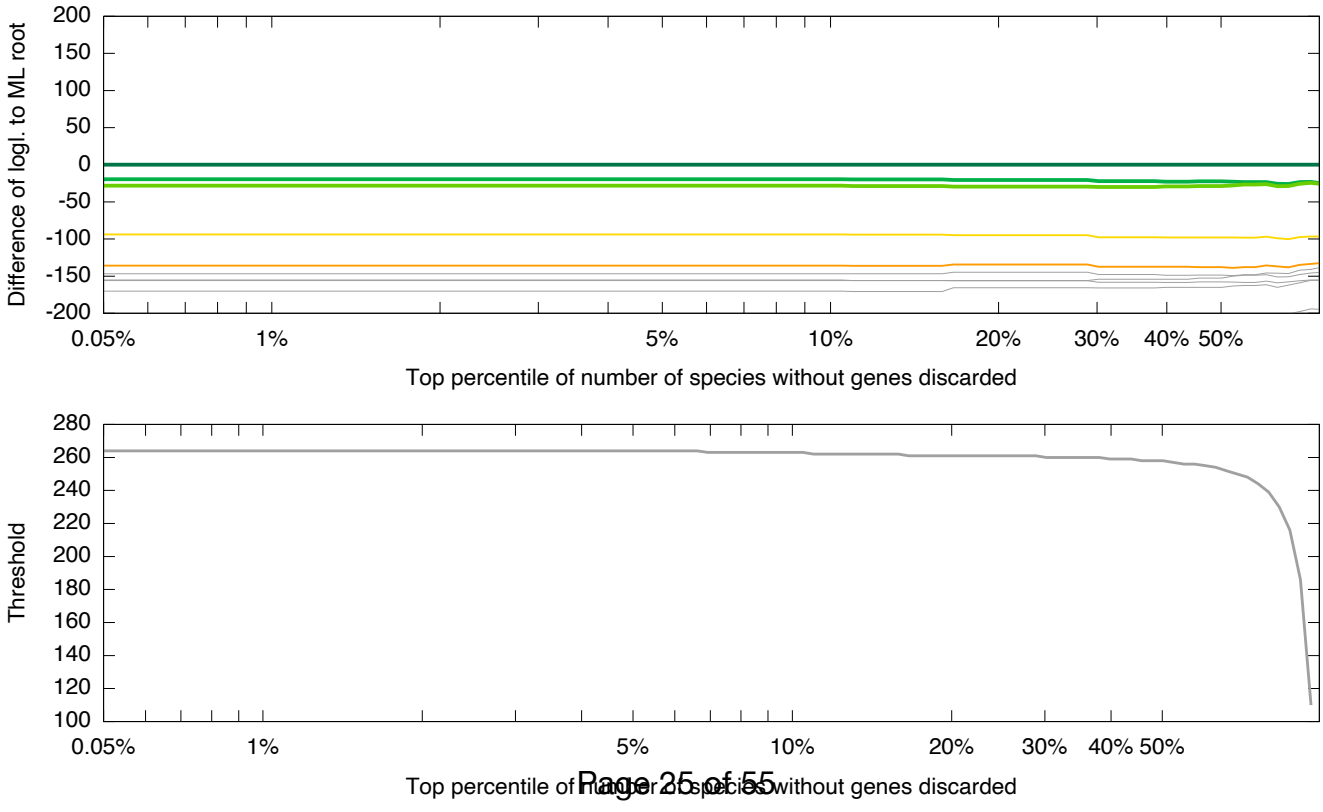


Fig. S12(e). Filtering COG families ranked by representation in smallest number of species

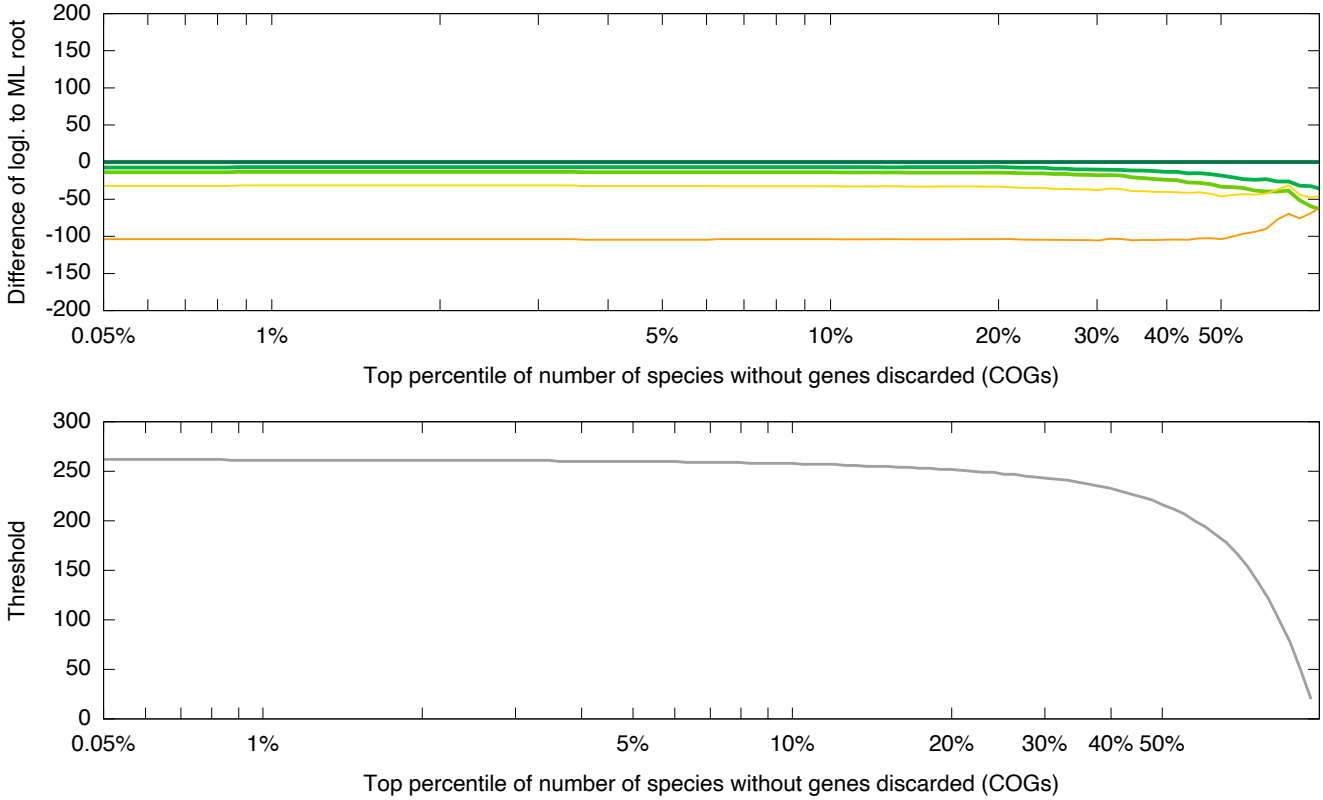


Fig. S12(f). Filtering MCL families ranked by highest mean copy number

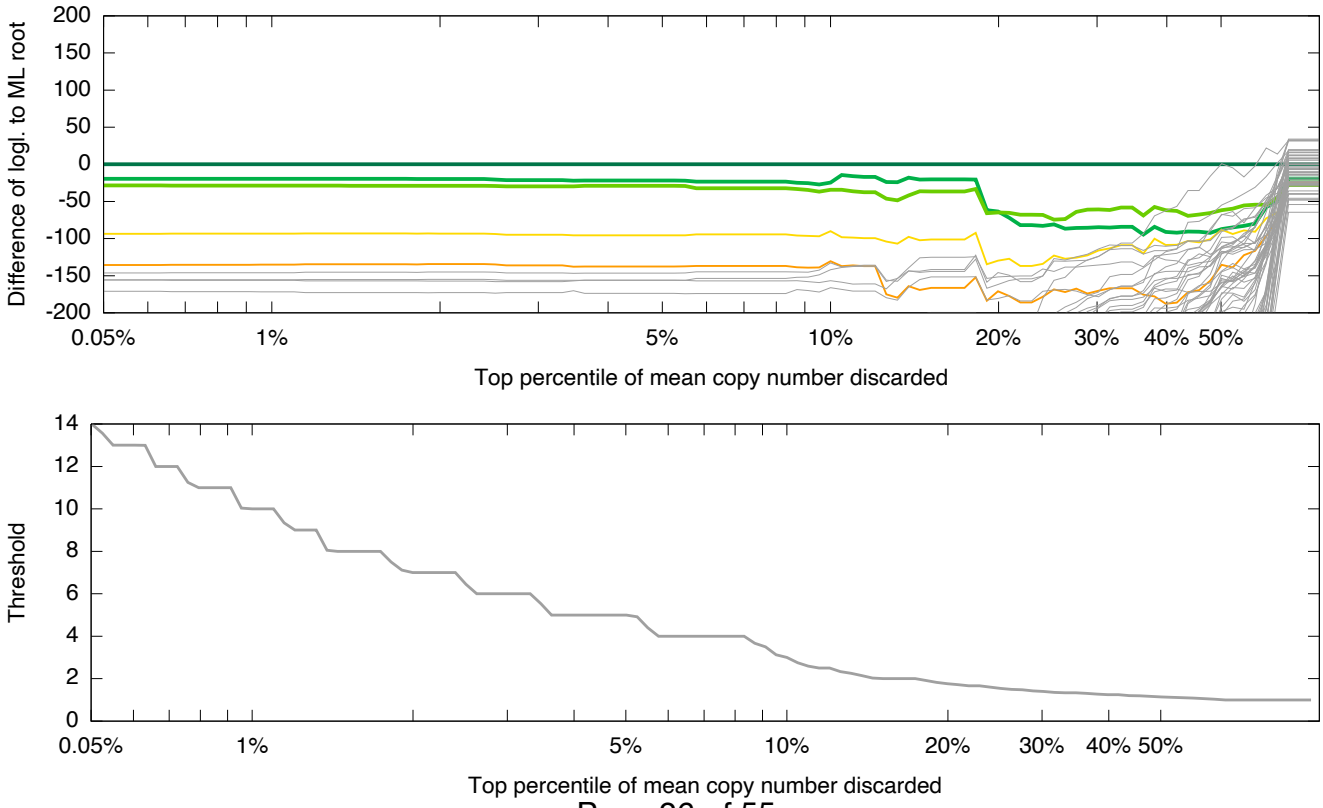


Fig. S12(g). Filtering COG families ranked by highest mean copy number

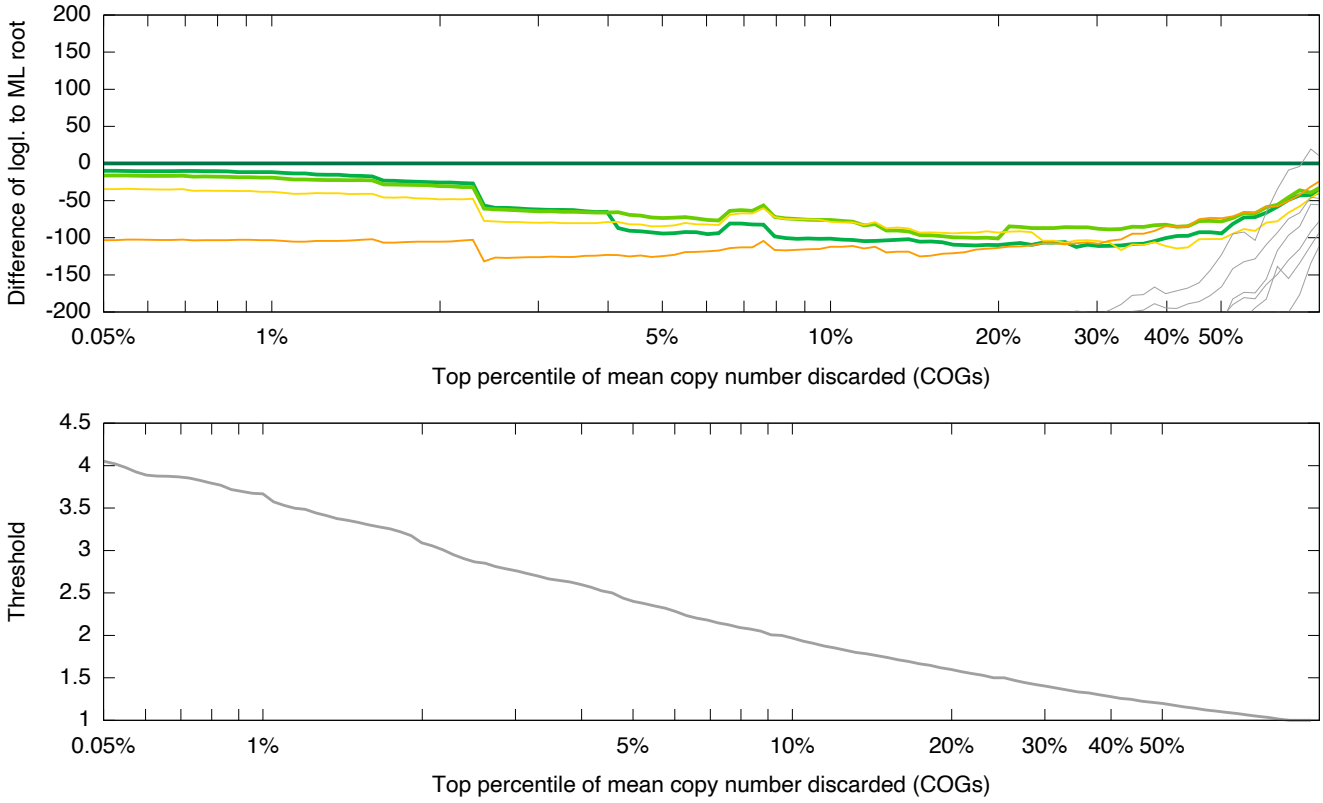


Fig. S12(h). Filtering MCL families ranked by D rate, high to low

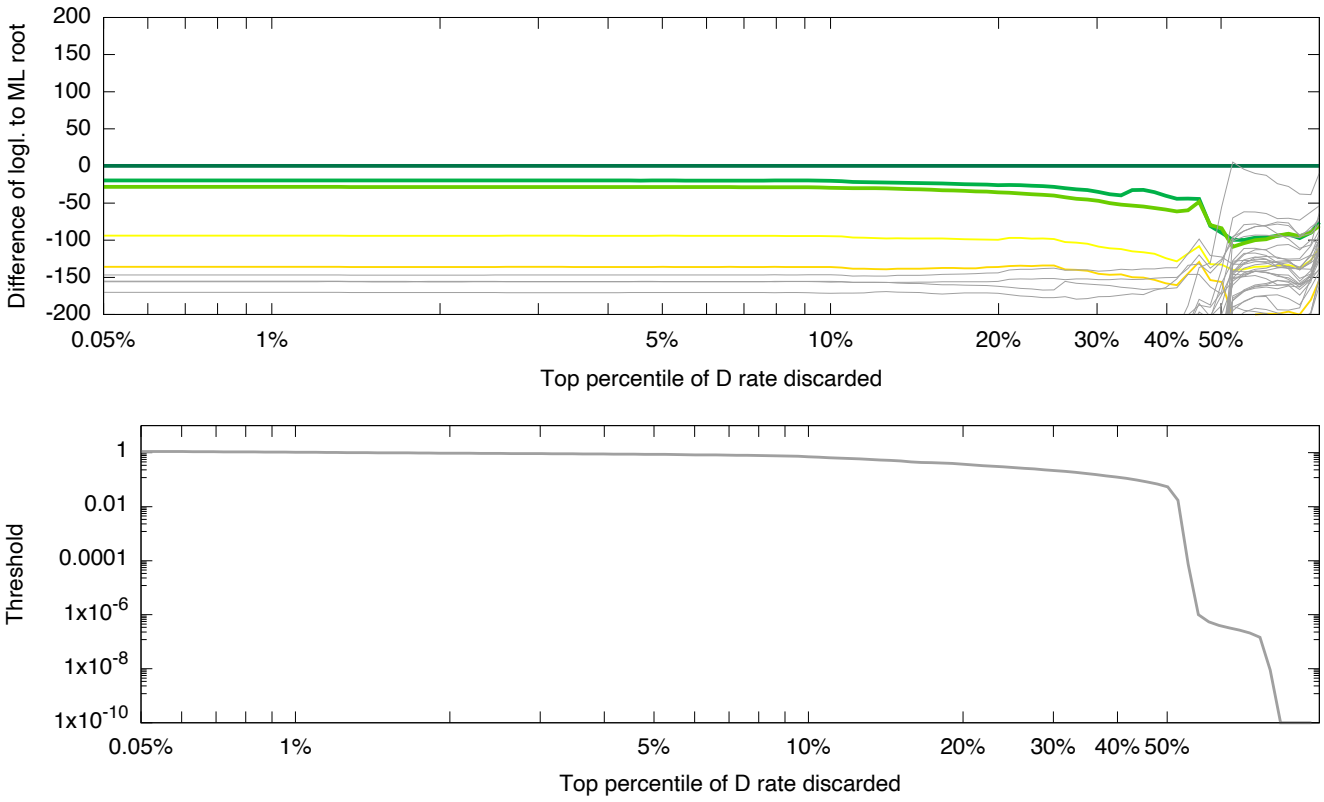


Fig. S12(i). Filtering COG families ranked by D rate, high to low

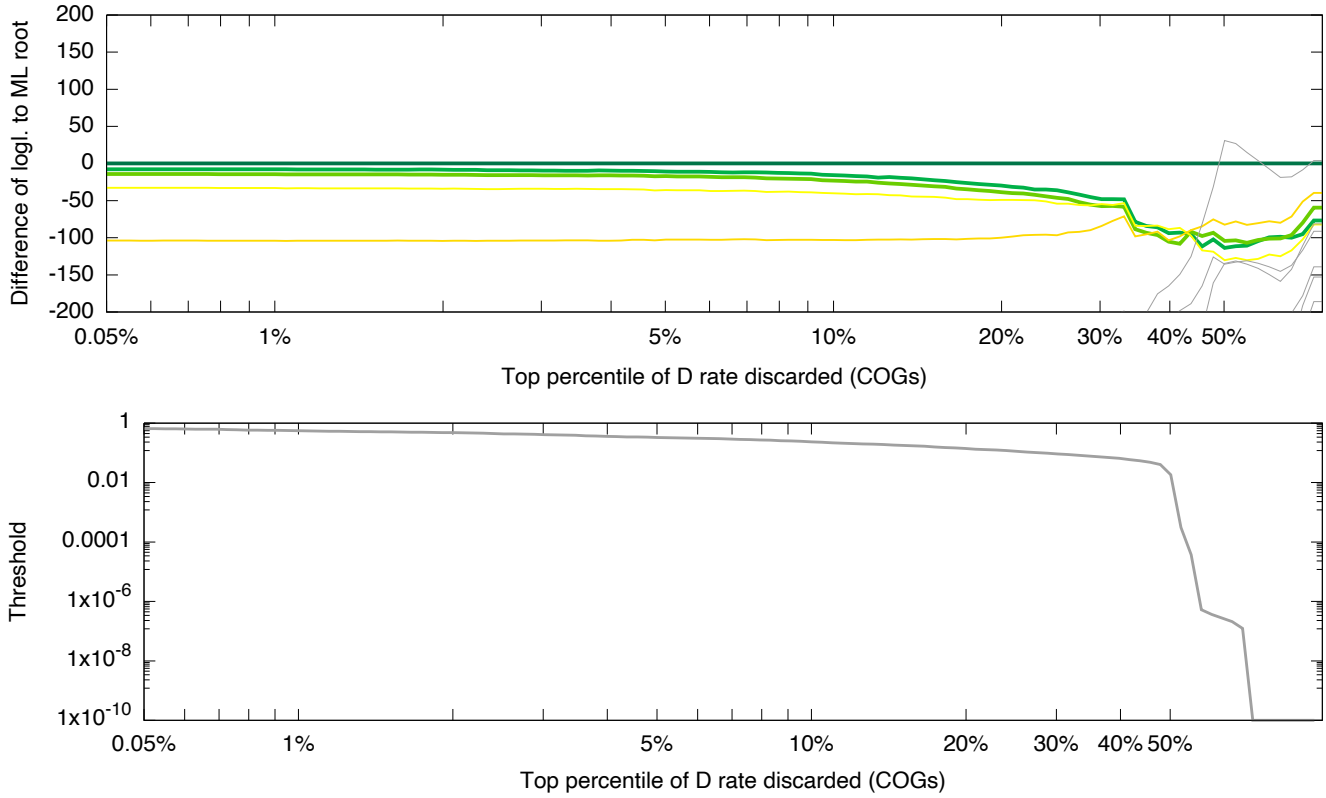


Fig. S12(j). Filtering MCL families ranked by T rate, high to low

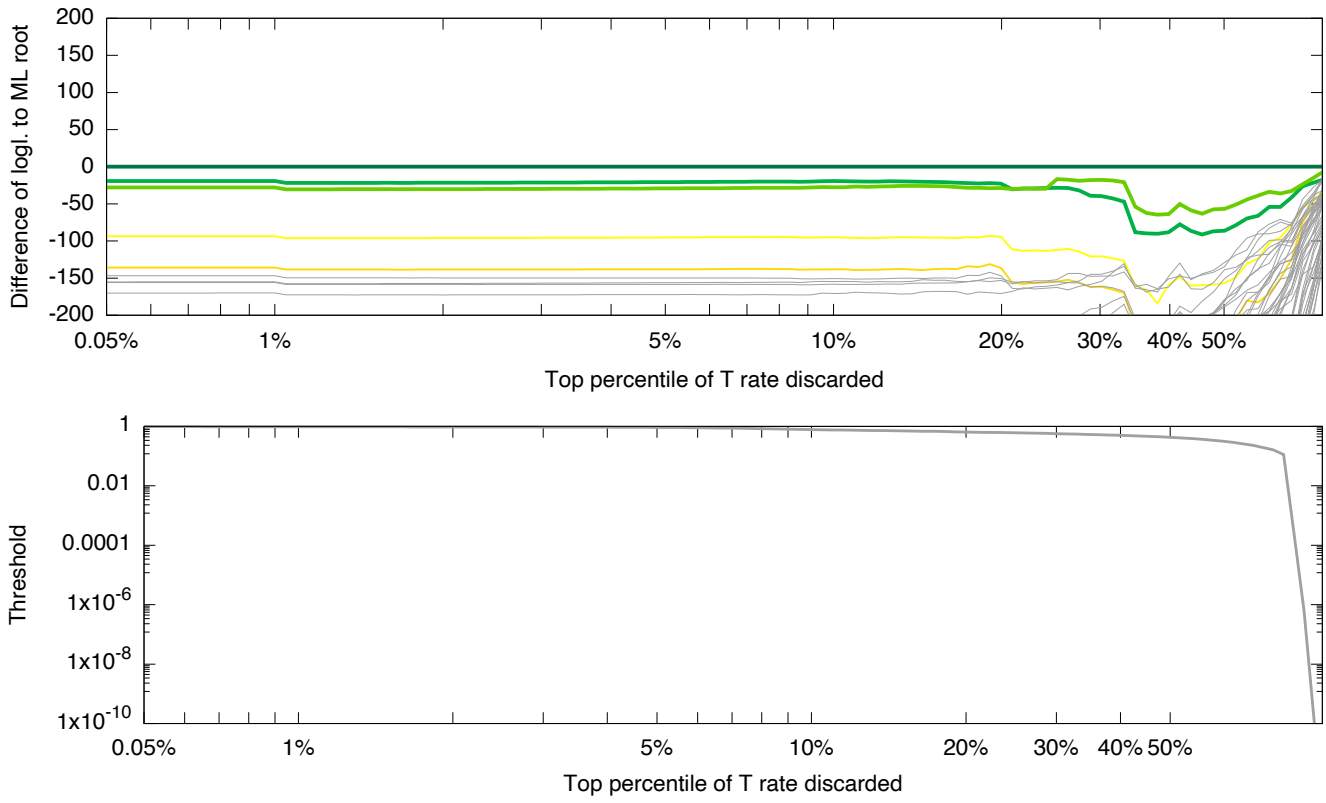


Fig. S12(k). Filtering COG families ranked by T rate, high to low

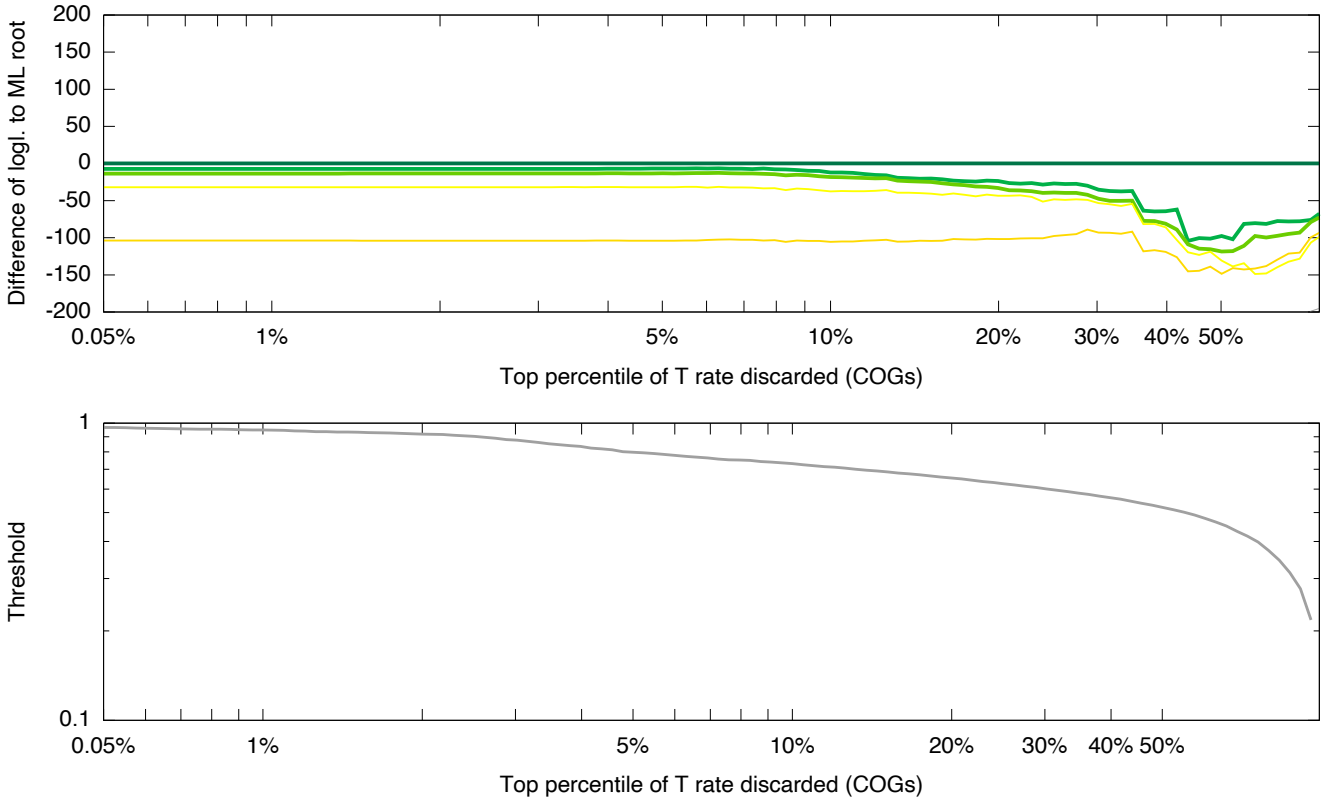


Fig. S12(l). Filtering MCL families ranked by L rate, high to low

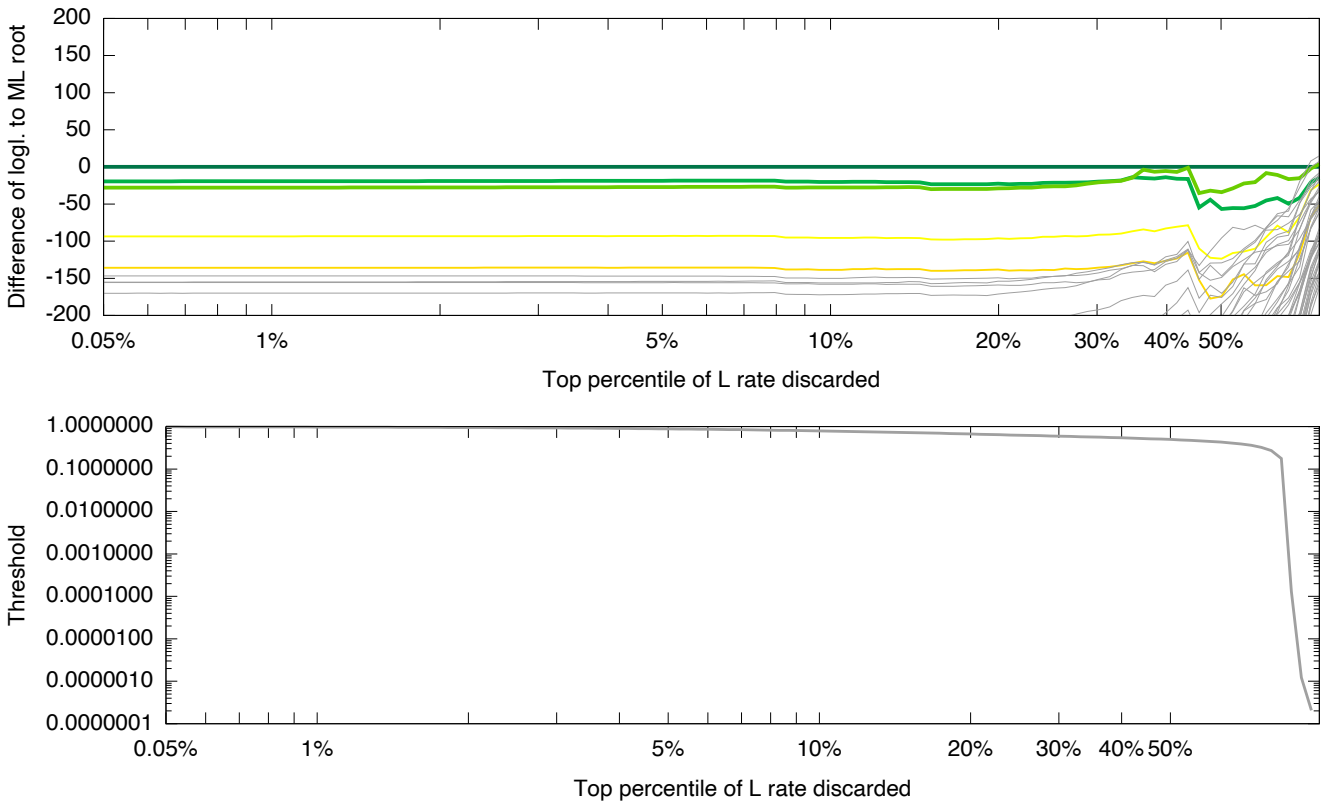


Fig. S12(m). Filtering COG families ranked by L rate, high to low

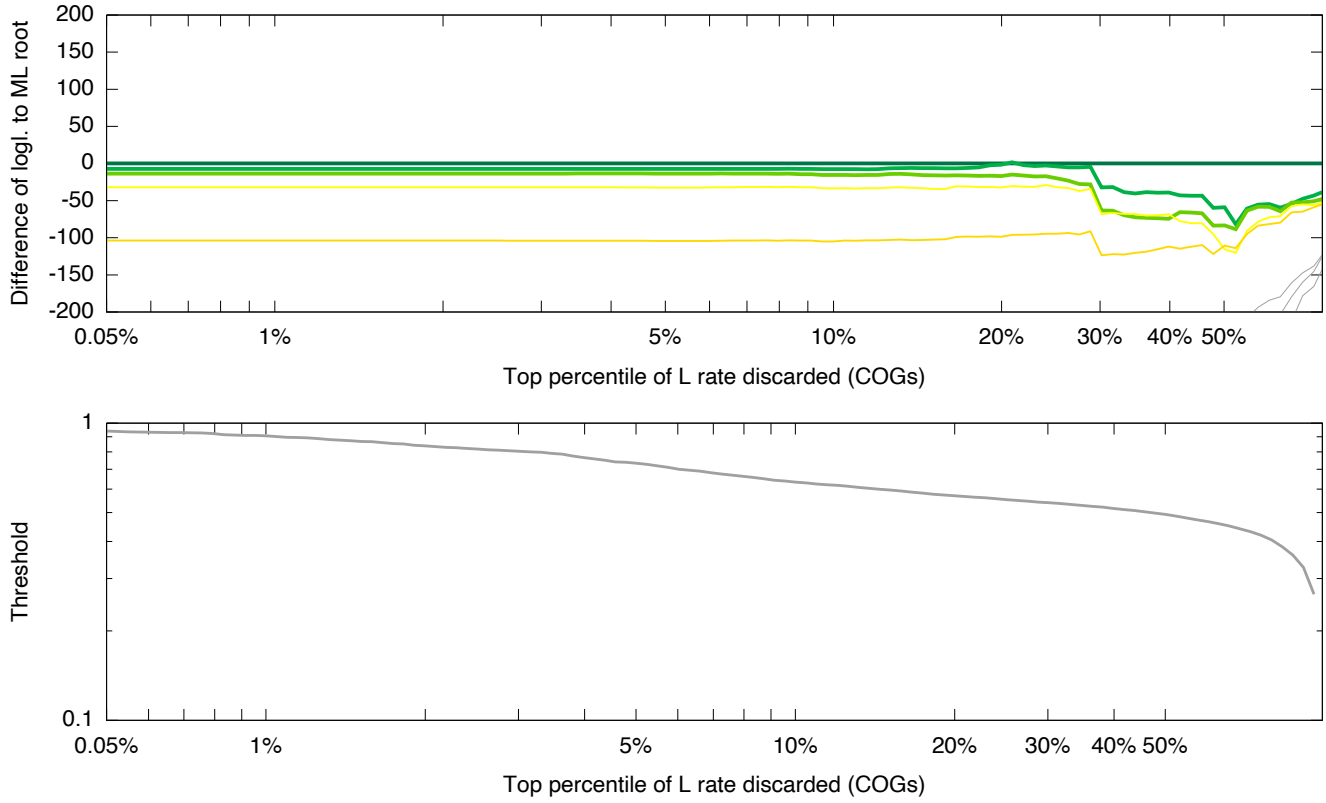


Fig. S12(n). Filtering MCL families ranked by largest DTL rate ratio

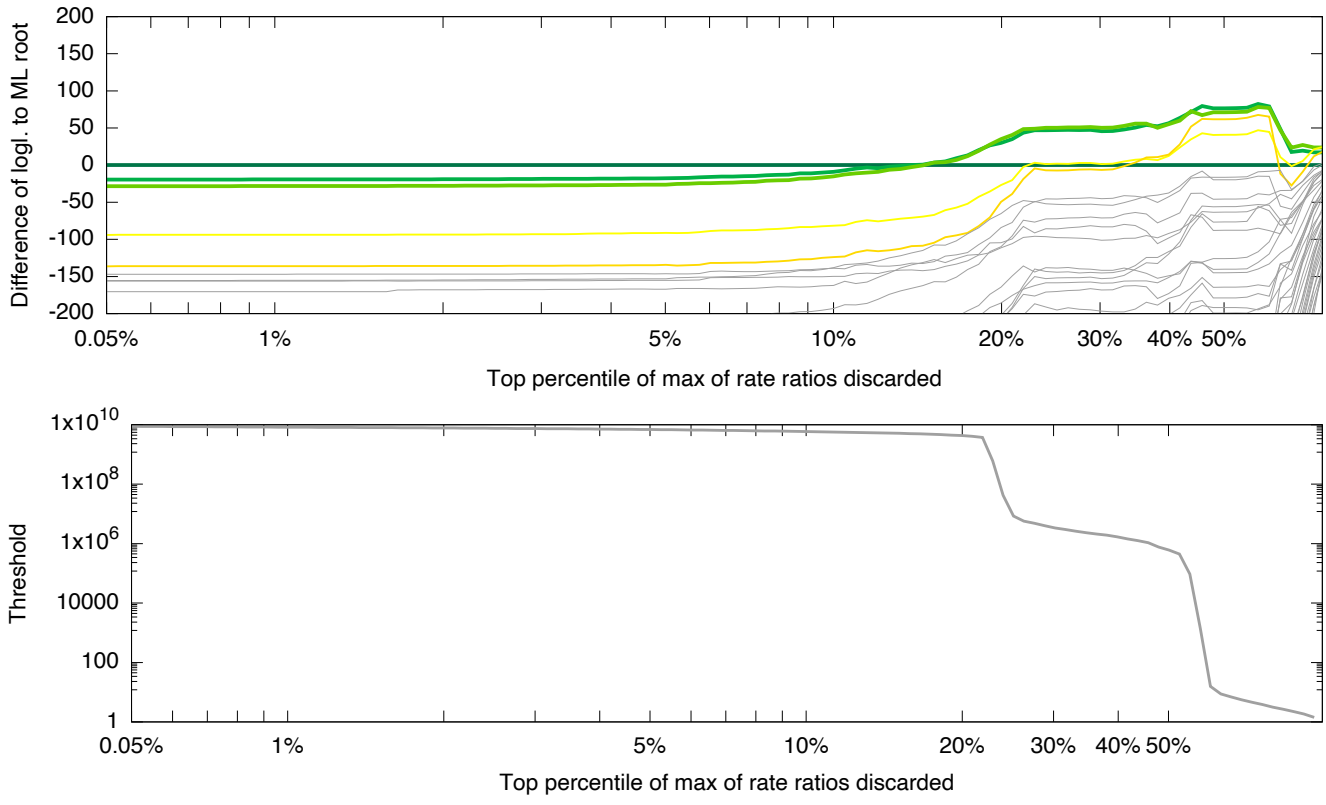


Fig. S12(o). Filtering COG families ranked by largest DTL rate ratio

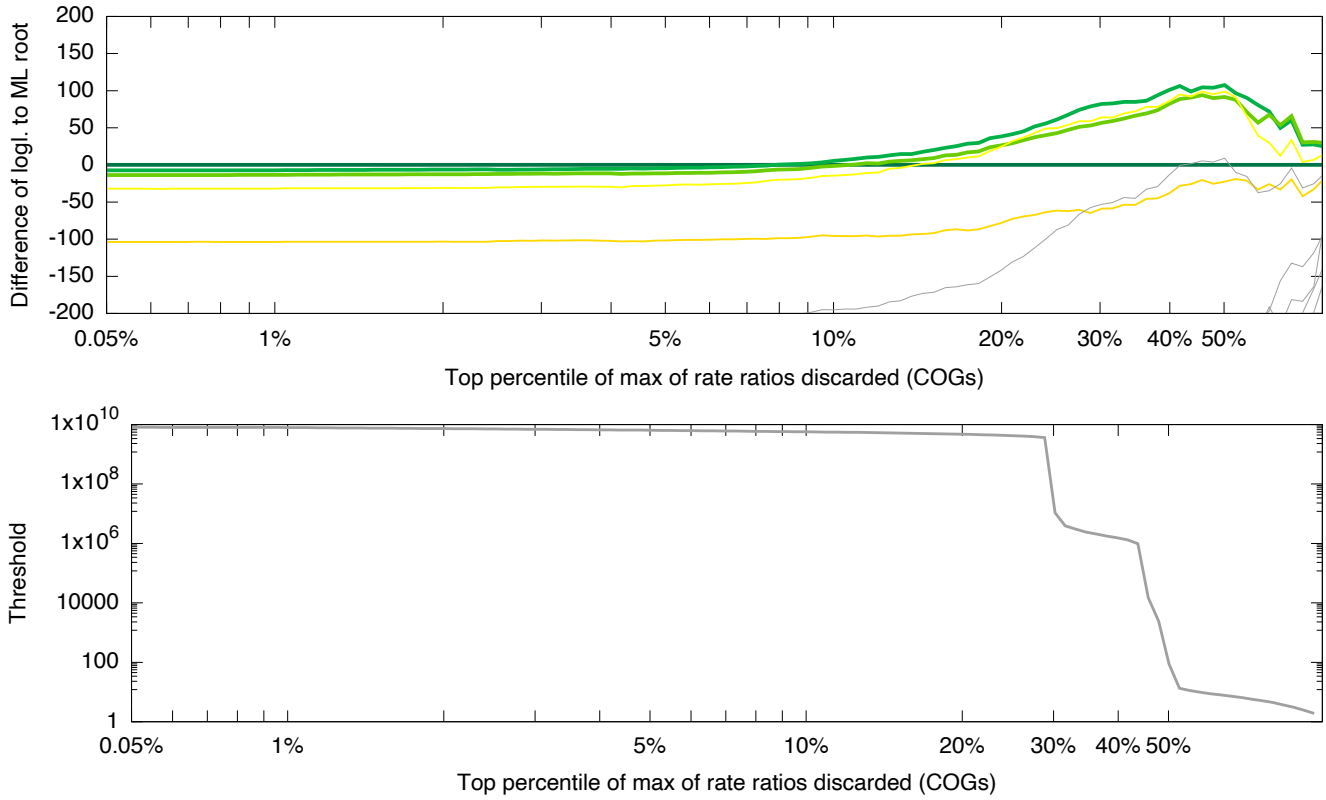


Fig. S12(p). Filtering MCL families ranked by verticality, high to low

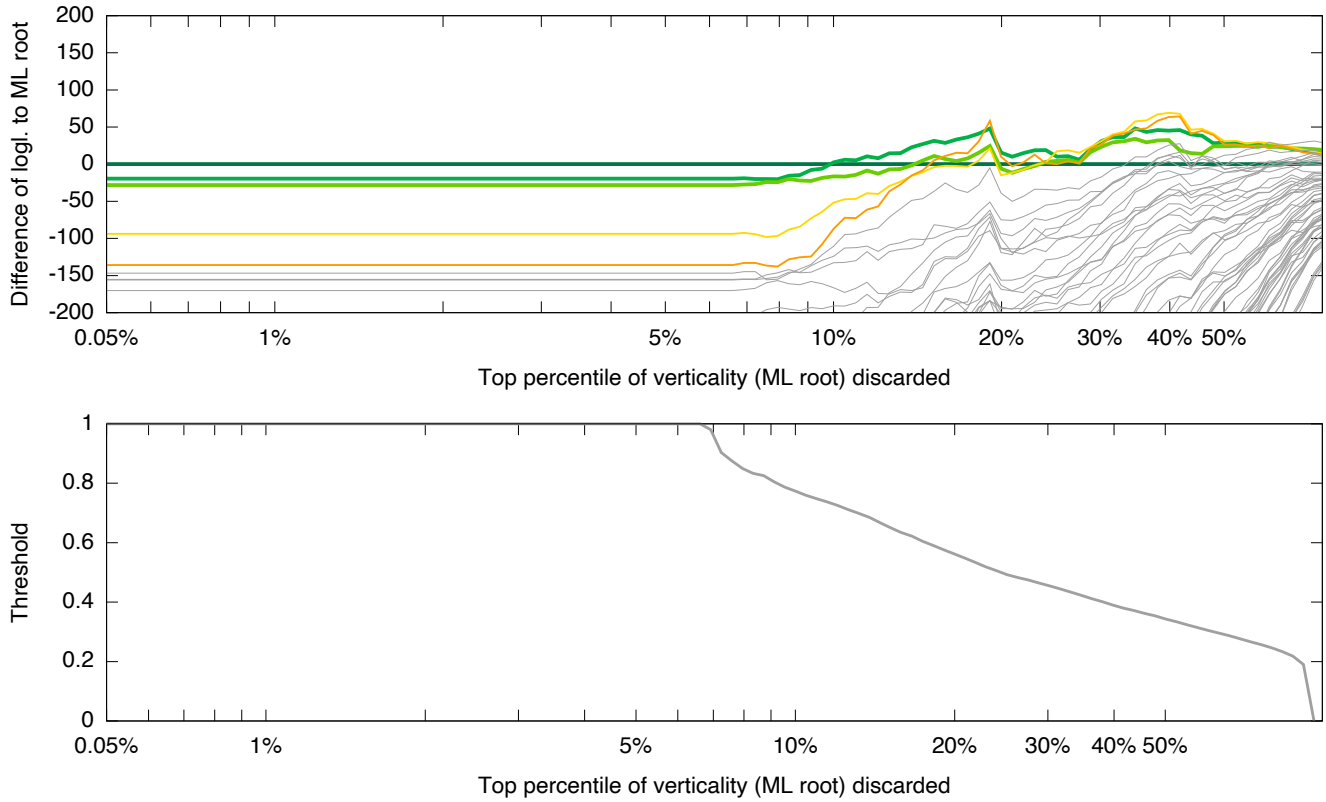


Fig. S12(q). Filtering COG families ranked by verticality, high to low

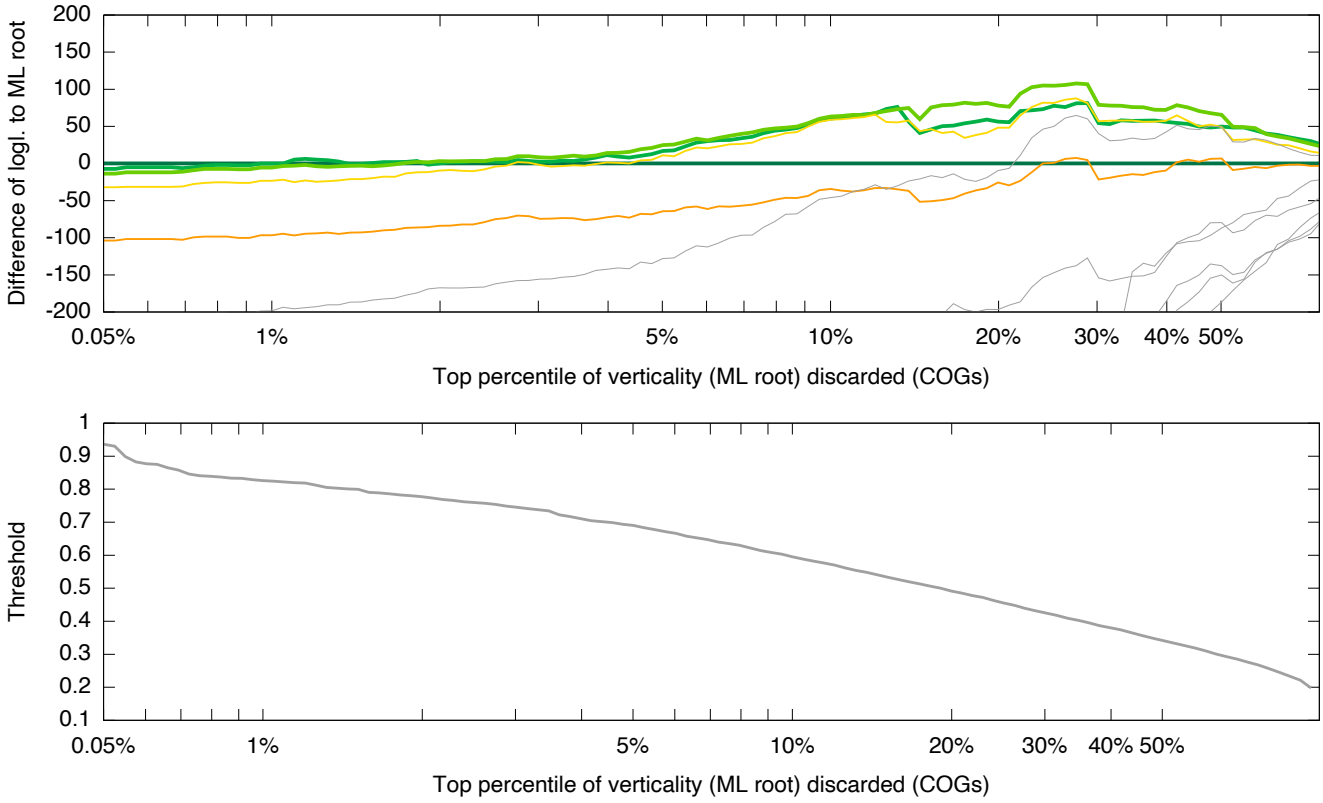


Fig. S12(r). Filtering MCL families ranked by verticality, low to high

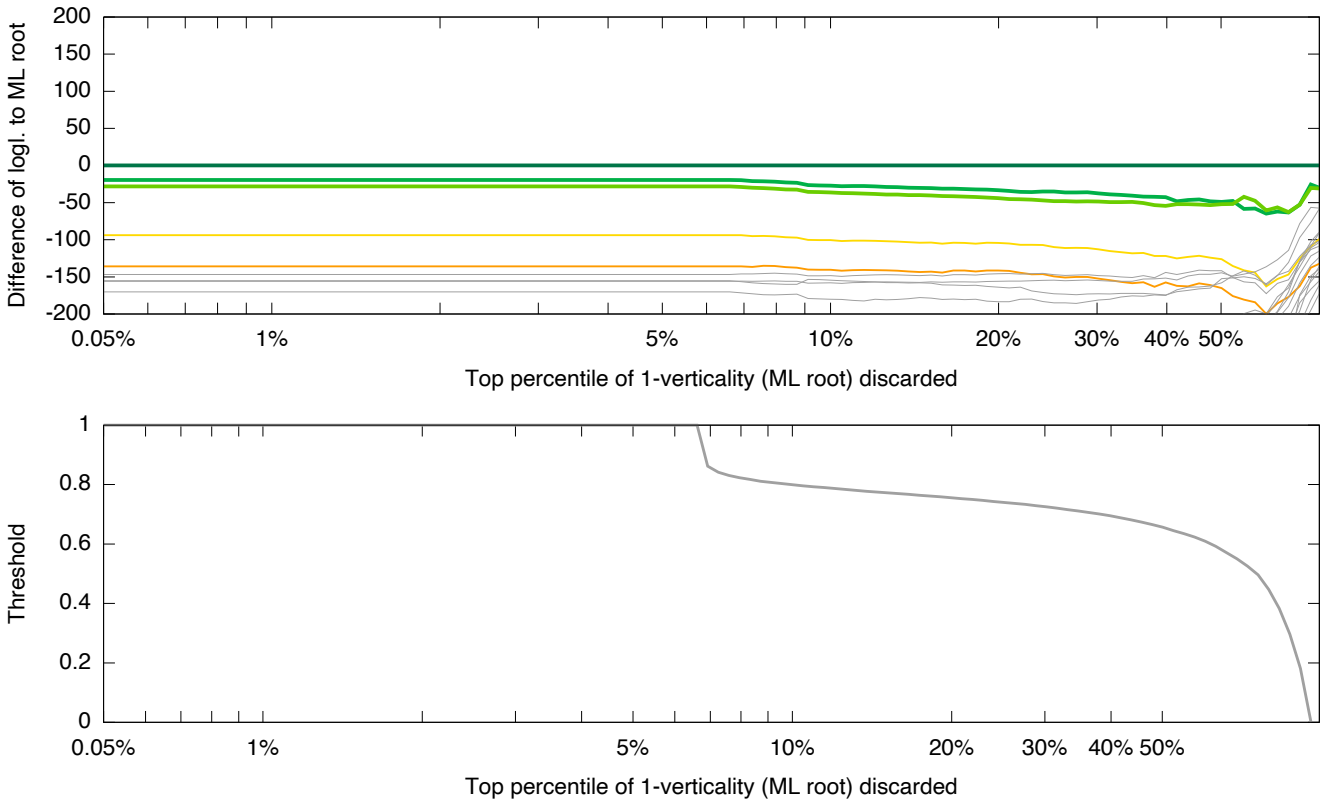
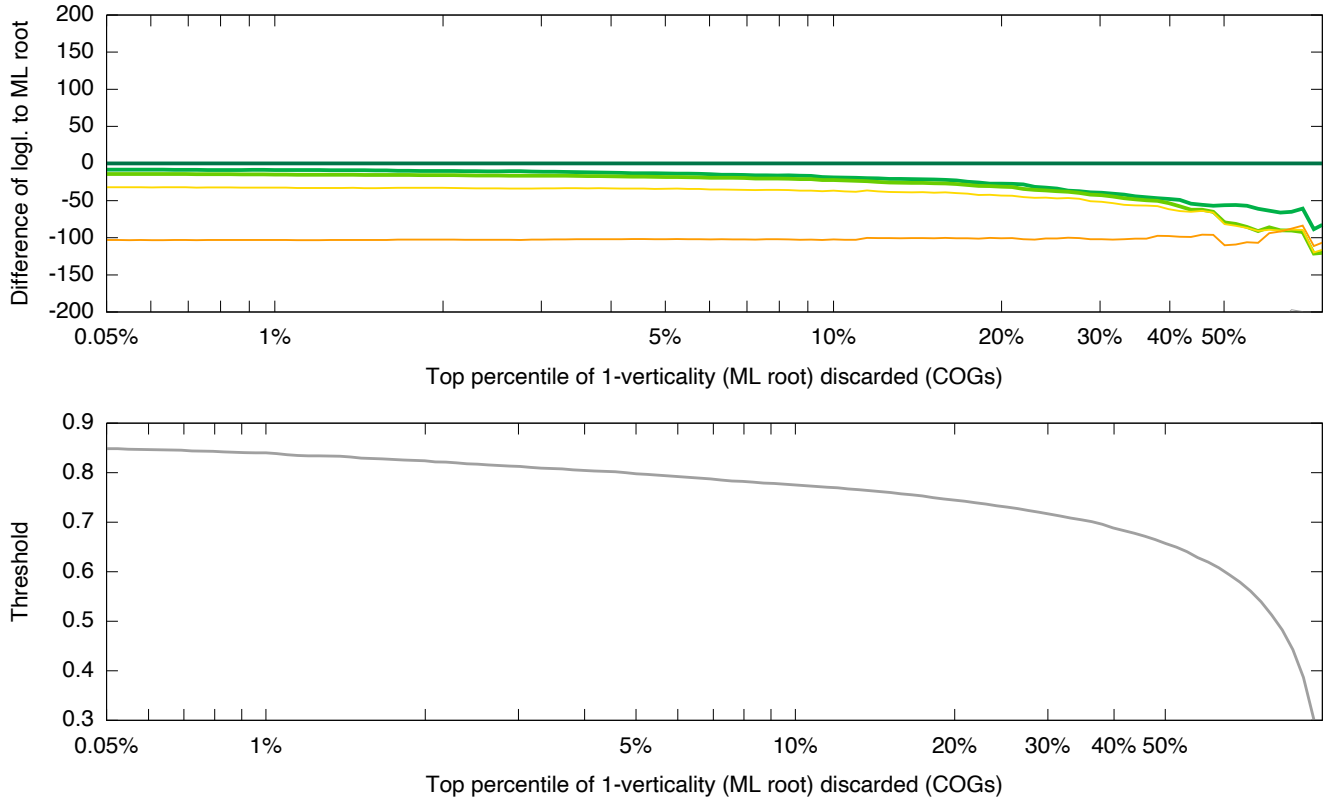


Fig. S12(s). Filtering COG families ranked by verticality, low to high



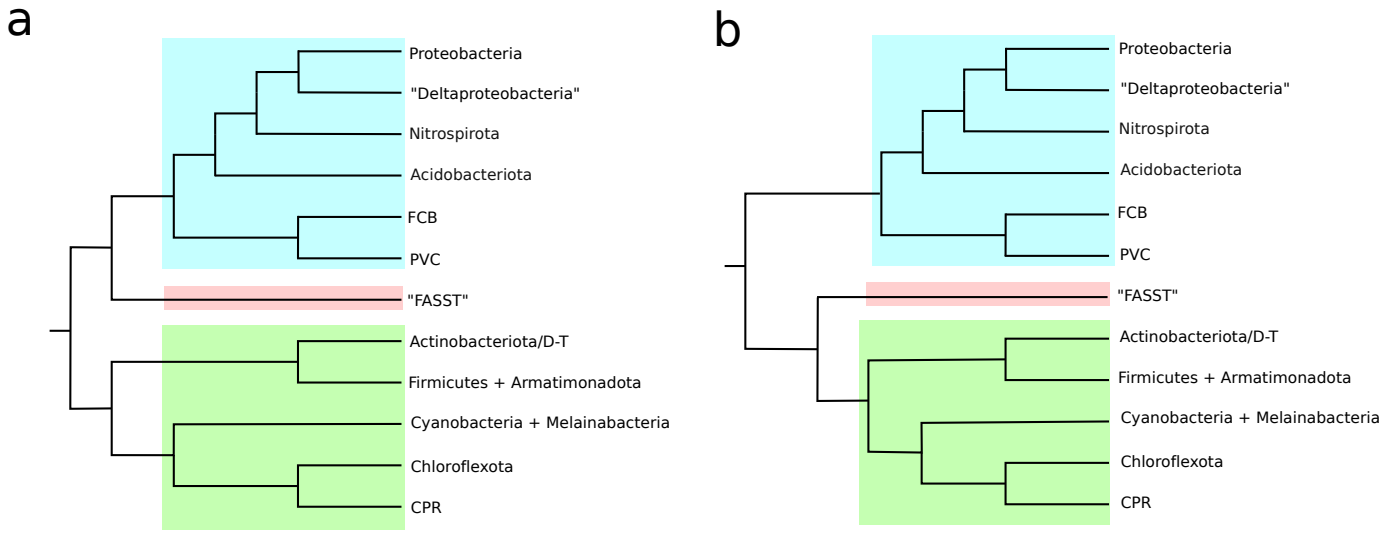


Fig. S13: Two rooted topologies from the secondary, GTDB-independent analysis that could not be rejected by the AU test, from ALE analysis incorporating genome completeness. AU p-values are 0.973 for tree (a) and 0.064 for tree (b). Both trees are in agreement with each other and with the focal analysis in placing the root between Terrabacteria and Gracilicutes, but disagree in the placement of the “FASST" taxa comprising Fusobacteriota, Aquificota, Synergistota, Spirochaetota and Thermotogota. D-T stands for Deinococcota-Thermus; “Deltaproteobacteria'' is Desulfuromonadota, Desulfobacterota, Bdellovibrionota, and Myxococcota.

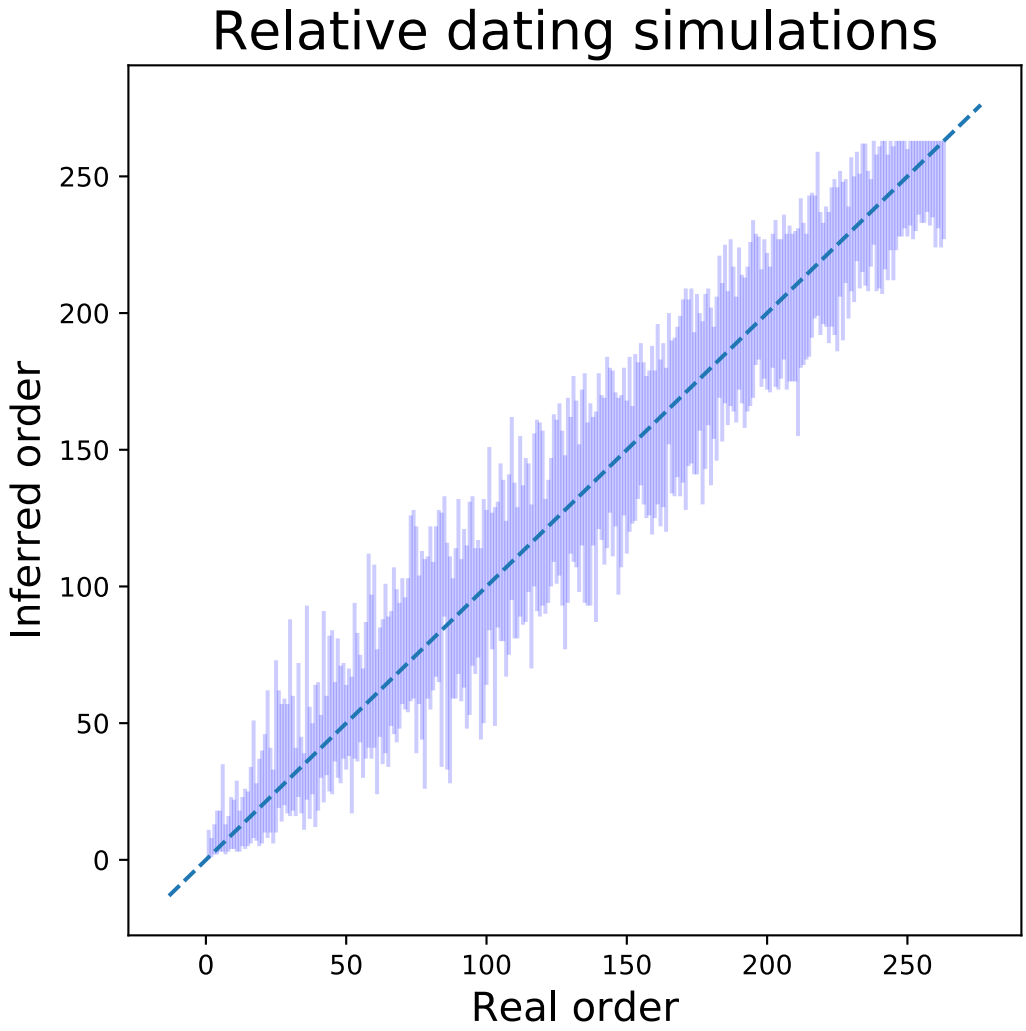


Fig. S14: Performance of relative dating approach on simulated data. We performed a relative dating analysis on a simulated dataset using the same number of gene families and rates as in the real dataset (11272), on a tree with 265 species. We applied the same pipeline used on the real dataset to obtain the relative dates. In the simulated dataset we obtain 16,943 highly supported constraints (found at least in 95 of the bootstrap replicates, see (*36*)), roughly twice as many as in the real dataset. The vertical blue lines represent the uncertainty in the ranking of the different speciation nodes. This approach correctly recovers the ranking of 98.4% of the nodes of the species tree, with only 4 being outside the distribution.

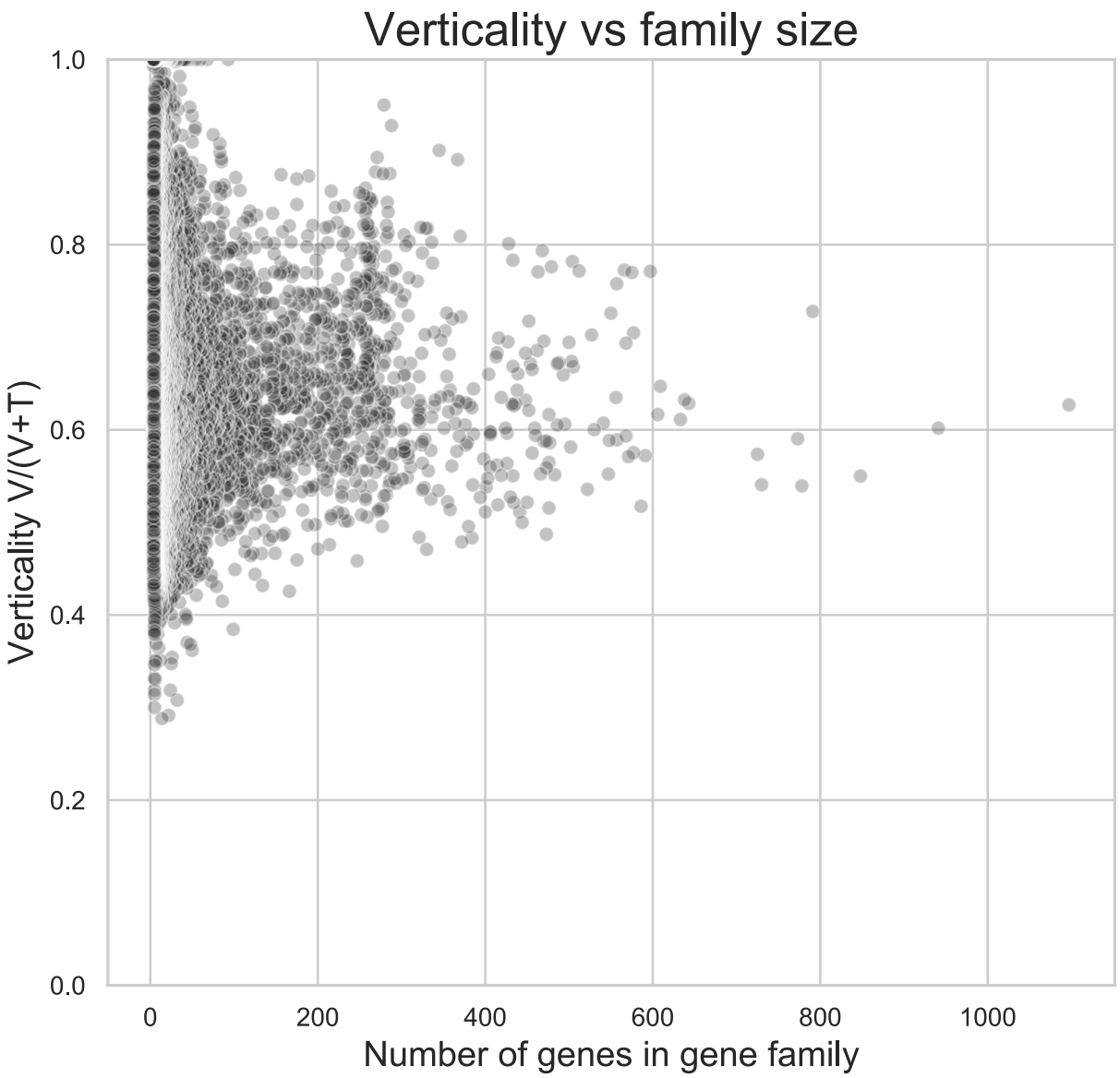
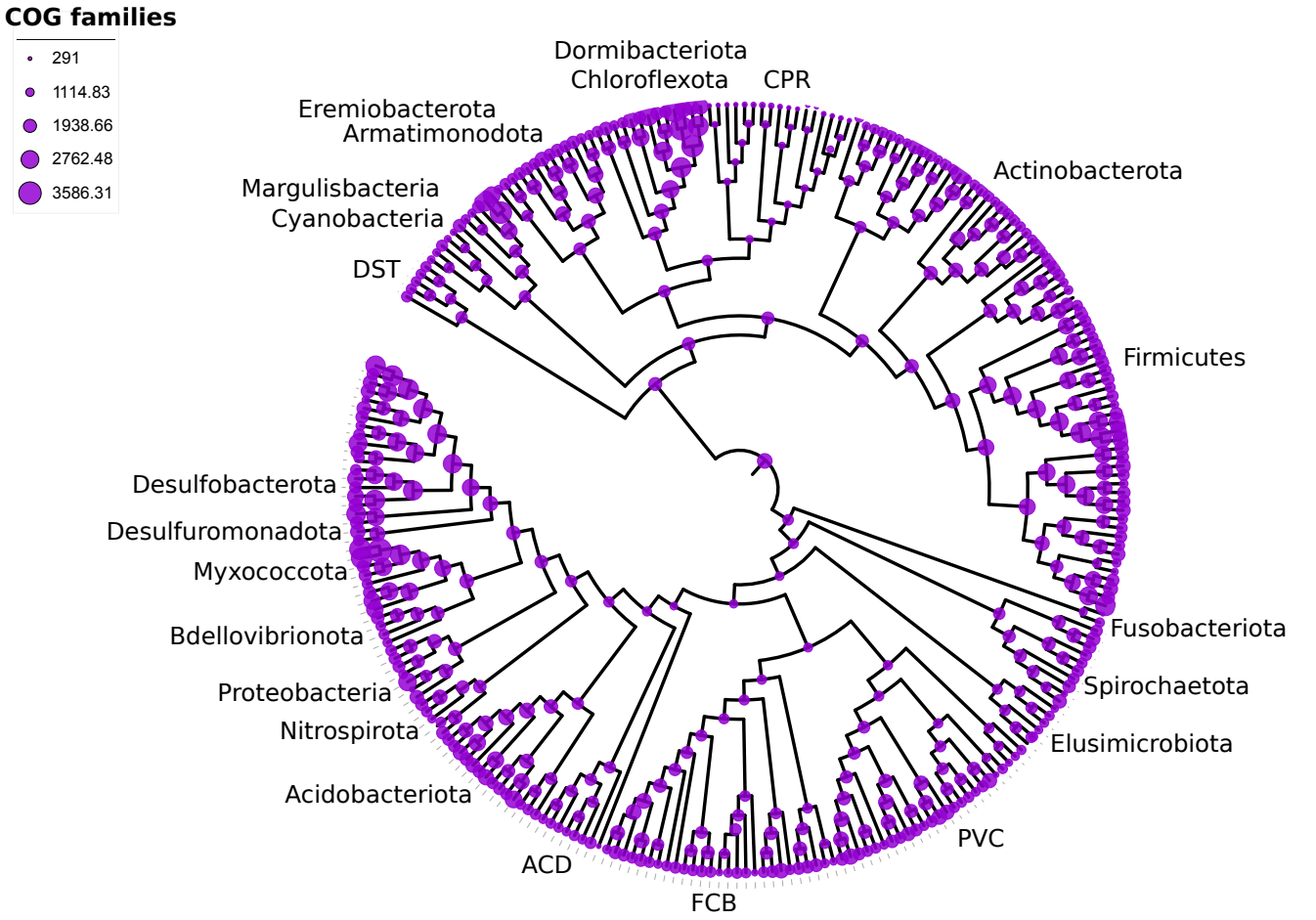


Fig. S15: The relationship between verticality and gene family size. Most gene families have experienced many transfers. Verticality varies with gene functional class, but families with very low transfer rates are small; these might represent young families that have not yet had enough time to experience gene transfer.


COG families
291
1114.83
1938.66
2762.48
3586.31
Dormibacteriota
Chloroflexota
CPR
Eremiobacterota
Armatimonodota
Actinobacterota
Margulisbacteria
Cyanobacteria
DST
Firmicutes
Desulfobacterota
Desulfuromonadota
Myxococcota
Fusobacteriota
Bdellovibrionota
Spirochaetota
Proteobacteria
Elusimicrobiota
Nitrospirota
Acidobacteriota
PVC
ACD
FCB

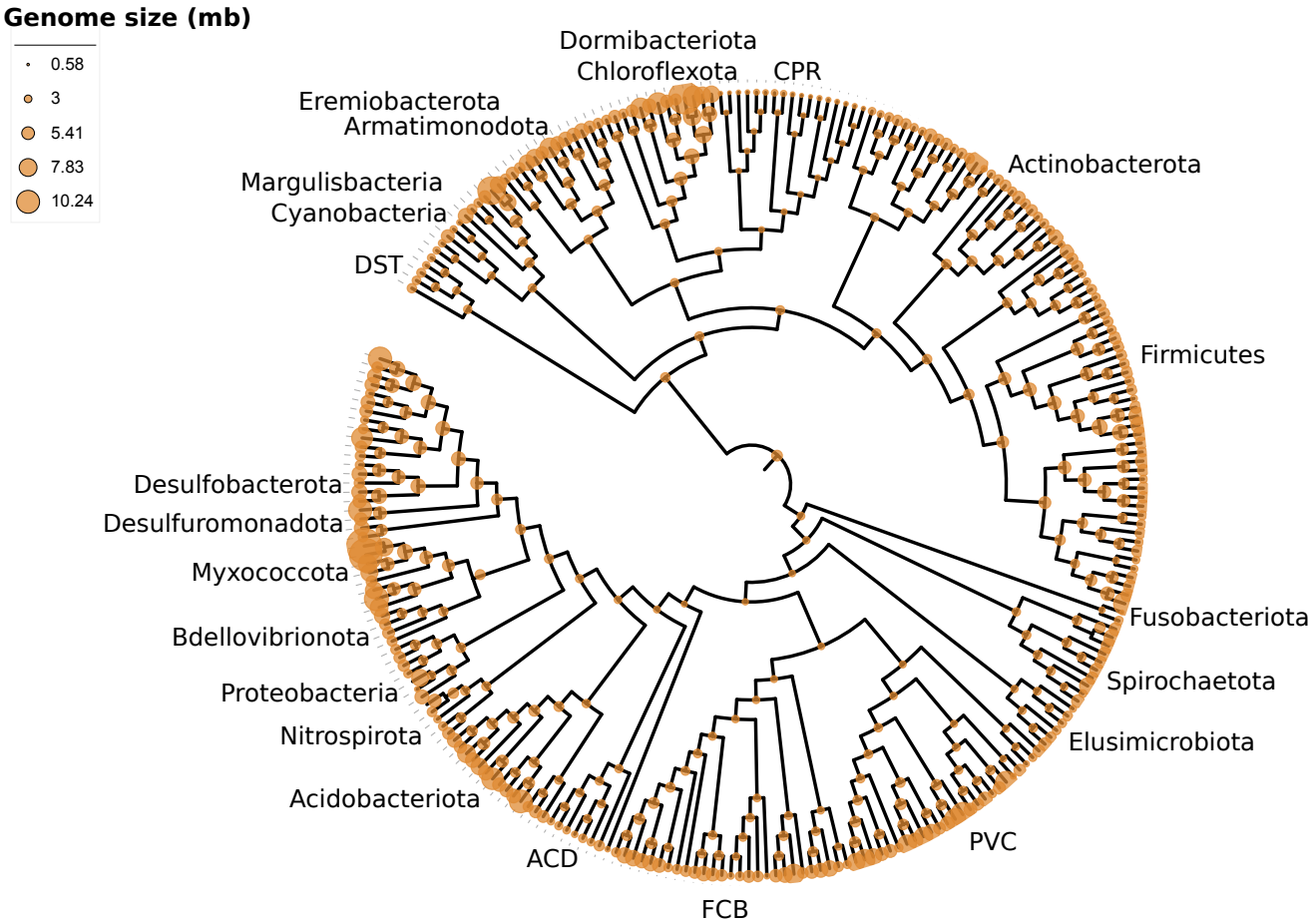
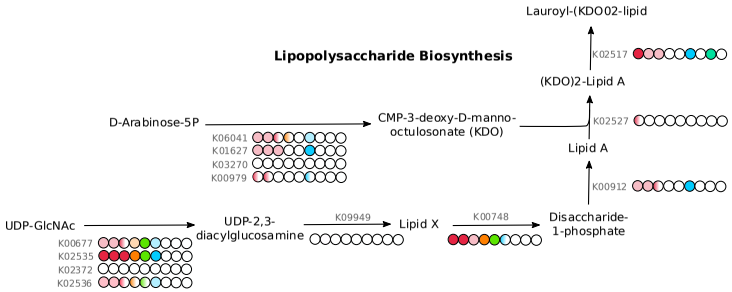
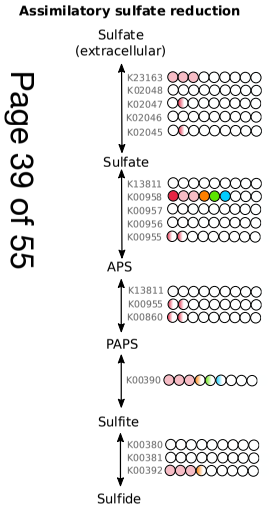
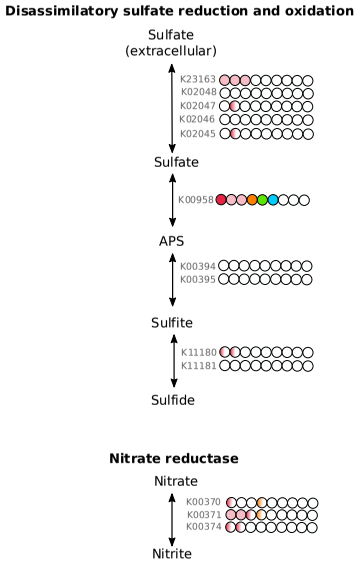

Genome size (mb)
0.58
3
5.41
7.83
10.24
Dormibacteriota
Chloroflexota
CPR
Eremiobacterota
Armatimonodota
Actinobacterota
Margulisbacteria
Cyanobacteria
DST
Firmicutes
Desulfobacterota
Desulfuromonadota
Myxococcota
Fusobacteriota
Bdellovibrionota
Spirochaetota
Proteobacteria
Elusimicrobiota
Nitrospirota
Acidobacteriota
PVC
ACD
FCB

Fig. S16: Evolution of COG family repertoires and inferred genome size over the bacterial tree. (a) The inferred number of COG family members and (b) inferred genome size at each internal node of the tree. Genome sizes were predicted from the relationship between COG family members and genome size among extant Bacteria (LOESS regression). Circle diameter is proportional to family number or genome size. FCB are the Fibrobacterota, Chlorobia, Bacteroidota, and related lineages; PVC are the Planctomycetota, Verrucomicrobiota, Chlamydiia, and related lineages; DST are the Deinococcota, Synergistota, and Thermotogota; ACD are Aquificota, Campylobacterota, and Deferribacterota; FA are Firmicutes and Actinobacteriota. The figure depicts inferences for root 1 (as shown in Fig. 1B); the data for all three roots are provided in GenomeSizeTable.tsv in the Online Data Supplement *(80)*.


Lipopolysaccharide Biosynthesis
Lauroyl-(KDO02-lipid
(KDO)2-Lipid A
Lipid A
D-Arabinose-5P
CMP-3-deoxy-D-manno-octulosonate (KDO)
UDP-GlcNAc
UDP-2,3-diacylglucosamine
Lipid X
Disaccharide-1-phosphate
K06041
K01627
K03270
K00979
K02517
K02527
K00912
K09949
K00748
K00677
K02535
K02372
K02536


Assimilatory sulfate reduction
Sulfate (extracellular)
Sulfate
APS
PAPS
Sulfite
Sulfide
K23163
K02048
K02047
K02046
K02045
K13811
K00958
K00957
K00956
K00955
K00955
K00860
K00390
K00380
K00381
K00392


Disassimilatory sulfate reduction and oxidation
Sulfate (extracellular)
Sulfate
APS
Sulfite
Sulfide
K23163
K02048
K02047
K02046
K02045
K00958
K00394
K00395
K11180
K11181
Nitrate reductase
Nitrate
Nitrite
K00370
K00371
K00374

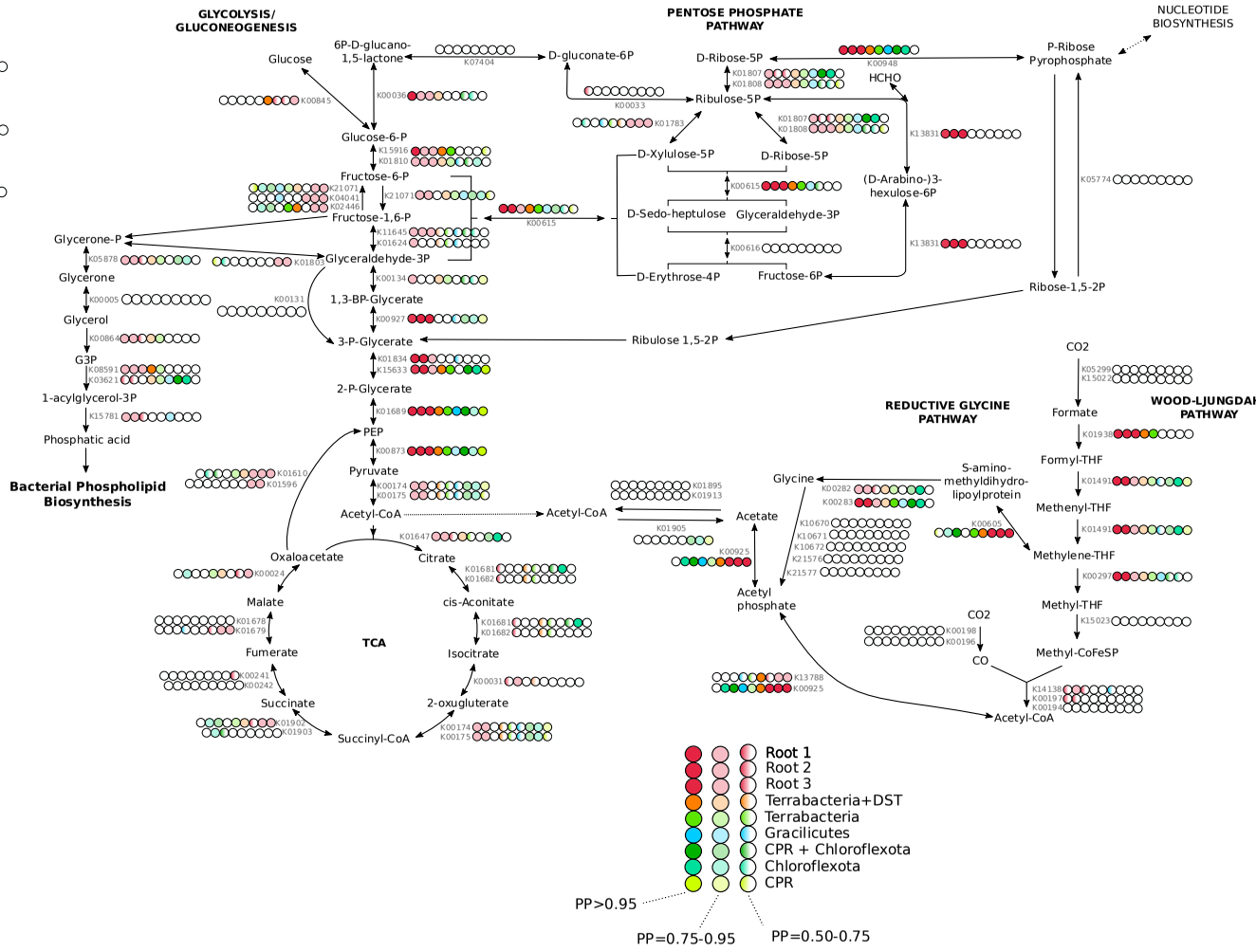

GLYCOLYSIS/ GLUCONEOGENESIS
PENTOSE PHOSPHATE PATHWAY
NUCLEOTIDE BIOSYNTHESIS
REDUCTIVE GLYCINE PATHWAY
WOOD-LJUNGDAHL PATHWAY
TCA
Bacterial Phospholipid Biosynthesis
Glucose
6P-D-glucano-1,5-lactone
D-gluconate-6P
Glucose-6-P
Fructose-6-P
Fructose-1,6-P
Glyceraldehyde-3P
Glycerone-P
Glycerone
Glycerol
G3P
1-acylglycerol-3P
Phosphatic acid
1,3-BP-Glycerate
3-P-Glycerate
2-P-Glycerate
PEP
Pyruvate
Acetyl-CoA
Oxaloacetate
Citrate
cis-Aconitate
Isocitrate
2-oxugluterate
Succinyl-CoA
Succinate
Fumerate
Malate
D-Ribose-5P
Ribulose-5P
D-Xylulose-5P
D-Ribose-5P
D-Sedo-heptulose
Glyceraldehyde-3P
D-Erythrose-4P
Fructose-6P
(D-Arabino-)3-hexulose-6P
HCHO
P-Ribose Pyrophosphate
Ribose-1,5-2P
Ribulose 1,5-2P
CO2
Formate
Formyl-THF
Methenyl-THF
Methylene-THF
Methyl-THF
Methyl-CoFeSP
CO
Acetyl-CoA
Acetate
Acetyl phosphate
Glycine
S-amino-methyldihydro-lipoylprotein
Root 1
Root 2
Root 3
Terrabacteria+DST
Terrabacteria
Gracilicutes
CPR + Chloroflexota
Chloroflexota
CPR
PP>0.95
PP=0.75-0.95
PP=0.50-0.75

Fig. S17: Metabolic map of the central metabolic pathways inferred in the last bacterial common ancestor (LBCA) and a selection of subsequent nodes. The reconstruction is based on genes that could be mapped to a given node with PP >0.5. The presence of a gene within a pathway is indicated as shown in the key. Annotations and PP values for KOs in this figure can be found in table S8. Annotations and PP values for all KOs can be found in table S7.

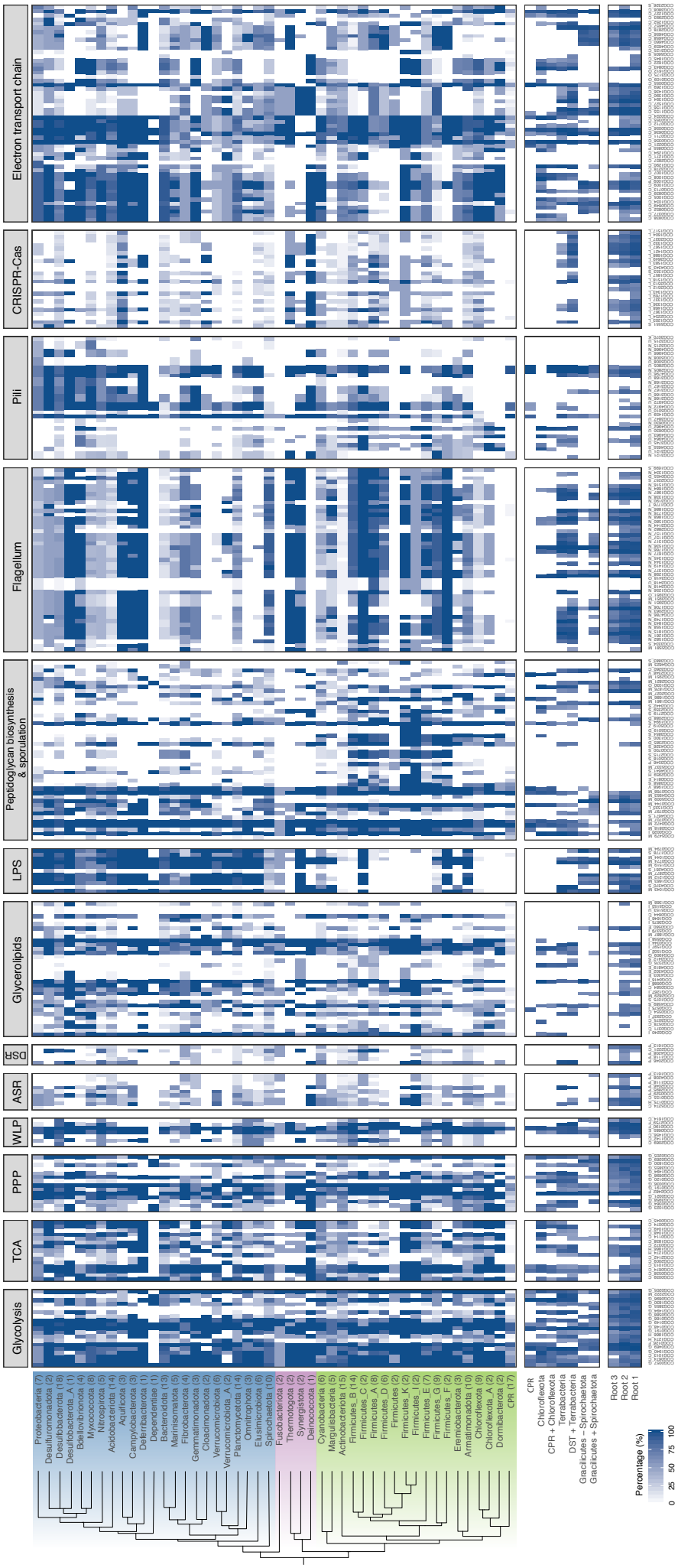

Glycolysis
TCA
PPP
WLP
ASR
DSR
Glycerolipids
LPS
Peptidoglycan biosynthesis & sporulation
Flagellum
Pili
CRISPR-Cas
Electron transport chain
Proteobacteria (7)
Desulfuromonadota (2)
Desulfobacterota (16)
Desulfobacterota_A (1)
Bdellovibrionota (4)
Myxococcota (8)
Nitrospirota (5)
Acidobacteriota (14)
Aquificota (3)
Campylobacterota (3)
Deferribacterota (1)
Dependentiae (1)
Bacteroidota (13)
Marinisomatota (5)
Fibrobacterota (4)
Gemmatimonadota (3)
Cloacimonadota (2)
Verrucomicrobiota (6)
Verrucomicrobiota_A (2)
Planctomycetota (14)
Omnitrophota (3)
Elusimicrobiota (6)
Spirochaetota (10)
Fusobacteriota (2)
Thermotogota (2)
Synergistota (2)
Deinococcota (1)
Cyanobacteria (6)
Margulisbacteria (5)
Actinobacteriota (15)
Firmicutes_B (14)
Firmicutes_C (2)
Firmicutes_A (8)
Firmicutes_D (6)
Firmicutes (2)
Firmicutes_K (2)
Firmicutes_I (2)
Firmicutes_G (7)
Firmicutes_E (9)
Firmicutes_F (2)
Elusimicrobiota (3)
Armatimonadota (10)
Chloroflexota (9)
Chloroflexota_A (2)
Dormibacterota (2)
CPR (17)
CPR
Chloroflexota
CPR + Chloroflexota
Terrabacteria
DST + Terrabacteria
Gracilicutes – Spirochaetota
Gracilicutes + Spirochaetota
Root 3
Root 2
Root 1
Percentage (%)
0
25
50
75
100

Fig. S18: Distribution of COG families from key metabolic pathways inferred to the last bacterial common ancestor (LBCA). The occurrence of COG families in the taxa sampled in this study are represented as percentage presence across phylogenetic clusters (phylum) based on a presence/absence table. COG families inferred to the given nodes and the tree possible root positions (see Methods) are represented by corresponding PP values (PP>0.5). TCA=Tricarboxylic Acid Cycle, PPP=Pentose phosphate pathway, ASR=Assimilatory sulfate reduction, DSR=Dissimilatory sulfate reduction, LPS=Lipopolysaccharide. Table S7 lists the PP values for all COG occurrences across roots, nodes, and tips and the metabolic genes featured in this plot can be found in table S8. A full heat map for all COGs, and additional heat maps by COGs category, can be found in (*80*), in heatmaps.zip.

Supplementary Tables

KO number	Gene name	Annotation	Used in outgroup tree?
K03046	rpoC	DNA-directed RNA polymerase subunit beta'	**y**
K03043	rpoB	DNA-directed RNA polymerase subunit beta	
K02337	dnaE	DNA polymerase III subunit alpha	**y**
K03070	secA	Protein translocase subunit SecA	
K01873	VARS, valS	Valine--tRNA ligase	
K02335	polA	DNA polymerase I	**y**
K01872	AARS, alaS	Alanine tRNA ligase	
K02469	gyrA	DNA gyrase subunit A	
K00962	pnp, PNPT1	Polyribonucleotide nucleotidyltransferase	**y**
K02355	fusA, GFM, EFG	Translation elongation factor G	**y**
K01972	E6.5.1.2, ligA, ligB	DNA ligase NAD	
K03702	uvrB	Excinuclease ABC subunit B	
K02470	gyrB	DNA gyrase subunit B	
K04077	groEL, HSPD1	Molecular chaperone GroEL	
K02313	dnaA	Chromosomal replication initiator protein	
K02314	dnaB	Replicative DNA helicase	
K02433	gatA, QRSL1	aspartyl-tRNA(Asn)/glutamyl-tRNA(Gln) amidotransferase subunit A	
K03076	secY	Protein translocase subunit SecY	**y**
K04485	radA, sms	DNA repair protein RadA/Sms	
K02112	ATPF1B, atpD	F-type H+/Na+-transporting ATPase subunit beta	**y**

K03590	ftsA	Cell division protein FtsA	
K02358	tuf, TUFM	Elongation factor Tu	**y**
K06942	ychF	Redox Regulated ATPase YchF	
K00927	PGK, pgk	Phosphoglycerate kinase	
K01889	FARSA, pheS	Phenylalanine--tRNA ligase alpha subunit	
K03551	ruvB	Holliday junction branch migration DNA helicase RuvB	**y**
K04485	radA, sms	DNA recombination repair protein RecA	**y**
K02835	prfA, MTRF1, MRF1	Peptide chain release factor 1	
K02886	RP-L2, MRPL2, rplB	50S ribosomal protein L2	**y**
K01803	TPI, tpiA	Triose phosphate isomerase	**y**
K03438	mraW, rsmH	16S rRNA (cytosine1402-N4)-methyltransferase	
K00554	trmD	tRNA (guanine37-N1)-methyltransferase	
K02863	RP-L1, MRPL1, rplA	50S ribosomal protein L1	**y**
K03685	rnc, DROSHA, RNT1	Ribonuclease III	
K02967	RP-S2, MRPS2, rpsB	30S ribosomal protein S2	**y**
K02982	RP-S3, rpsC	30S ribosomal protein S3	
K02906	RP-L3, MRPL3, rplC	50S ribosomal protein L3	**y**
K03470	rnhB	Ribonuclease HII	
K01358	clpP, CLPP	ATP dependent Clp protease proteolytic subunit	
K06187	recR	Recombination protein RecR	
K15034	yaeJ	Aminoacyl tRNA hydrolase, ribosome-associated protein	
K02931	RP-L5, MRPL5, rplE	50S ribosomal protein L5	**y**
K02933	RP-L6, MRPL6, rplF	50S ribosomal protein L6	**y**

K02601	nusG	Transcription termination antitermination protein NusG	
K02988	RP-S5, MRPS5, rpsE	30S ribosomal protein S5	**y**
K02992	RP-S7, MRPS7, rpsG	30S ribosomal protein S7	**y**
K03664	smpB	SsrA binding protein	
K02838	frr, MRRF, RRF	Ribosome recycling factor	
K02867	RP-L11, MRPL11, rplK	50S ribosomal protein L11	
K02878	RP-L16, MRPL16, rplP	50S ribosomal protein L16	**y**
K02871	RP-L13, MRPL13, rplM	50S ribosomal protein L13	**y**
K02994	RP-S8, rpsH	30S ribosomal protein S8	**y**
K02948	RP-S11, MRPS11, rpsK	30S ribosomal protein S11	**y**
K02952	RP-S13, rpsM	30S ribosomal protein S13	**y**
K02935	RP-L7, MRPL12, rplL	50S ribosomal protein L7/12	
K02996	RP-S9, MRPS9, rpsI	30S ribosomal protein S9	
K02874	RP-L14, MRPL14, rplN	50S ribosomal protein L14	**y**
K02887	RP-L20, MRPL20, rplT	50S ribosomal protein L20	
K02946	RP-S10, MRPS10, rpsJ	30S ribosomal protein S10	**y**
K02965	RP-S19, rpsS	30S ribosomal protein S19	**y**
K02956	RP-S15, MRPS15, rpsO	30S ribosomal protein S15	
K02518	infA	Translation initiation factor IF 1	

Table S1: 62 orthologous genes used to infer the species tree, with those used in the outgroup rooting analysis indicated.

Root hypothesis	log-likelihood difference to ML	*p*-value	Study
Observed outgroup root (Fig. S1)	0	0.58	This study (ML tree)
Between Firmicutes and Actinobacteriota	-3.3	0.50	(*26*)
Deinococcota, Synergistota and Thermotogota basal*	-3.7	0.52	
Planctomycetota basal	-4.9	0.49	(*8*)
Chloroflexota basal	-6.2	0.52	(*9*)
CPR basal	-16.7	0.37	(*11*, *16*)
DPANN basal within archaeal outgroup	-19.4	0.37	(*1*, *20*)
Fusobacteriota basal	-24.2	0.33	
Between Gracilicutes and Terrabacteria	-24.5	0.33	This study (ALE root, see below)

Table S2: Support for published hypotheses using outgroup rooting. *Our unrooted topology was incompatible with some published hypotheses, including a clade of Thermotogales and Aquificales at the root (*6*, *7*).

Root	*p*-value	Study
CPR basal	2e-04	*(11,16)*
Chloroflexota (Chloroflexi) basal	1e-41	*(9)*
Between Firmicutes and Actinobacteriota	9e-05	*(26)*
Thermotogota/ Synergistota/ Deinococcota basal*	0.004	
Planctomycetota (Planctomycetes) basal	2e-26	*(8)*

Table S3: **Support for published rooting hypotheses from our outgroup-free analyses.** *Our unrooted topology was incompatible with some published hypotheses, including a clade of Thermotogota and Aquificota at the root (*6*, *7*).

Table S4: AU *p*-values for all tested roots in ALE analysis (Excel-formatted spreadsheet)

Root name	LLs	AU
Fusobacteria root (398)	-7.3	0.589
Fusobacteriota on Terrabacteria side (527)	7.3	0.519
Fusobacteriota on Gracilicutes side (528)	13.6	0.432
Fusobacteriota and Spirochaetota on Terrabacteria side (520)	32.1	0.251
DST root (464)	103.7	0.008
Cyanobacteria on Gracilicutes side (517)	215.8	1e-09
Dormi/Chloroflexi+CPR (510)	372.2	7e-06
CPR root (496)	425.2	2e-67
Dormibacterota/Chloroflexota (505)	630.7	5e-05
Omnitrophota/Verrucomicrobiota/ Planctomycetota (492)	674.4	2e-04
Fibrobacterota/Bacteroidota/ Marinisomatota (511)	1000.7	3e-71

Table S5: AU-test results for an ALE root analysis on the focal dataset using 3595 COG families.

Root branch	1	2	3
Median singleton support per branch	999.09	998.47	999.21
Singleton support for branches subtending the root	98.259, 140.45	91.56, 151.09	95.25,117.16
Mean verticality	0.641	0.641	0.642

Table S6: Singleton support (the number of genes that evolve vertically from one end of a branch to the other) on the credible set of rooted trees. Root numbers correspond to the three root branches depicted in Fig. 1(b).

Table S7: Protein family annotations (COG and KO) and root presence posterior probabilities (PPs) for key pathways used in ancestral reconstruction (Excel-formatted spreadsheet).

Table S8: COG families lost on the CPR stem (Excel-formatted spreadsheet).

Table S9: Species names and accessions of the genomes used in this study (Excel-formatted spreadsheet).

COG	Annotation	Number of families	Median transfer propensity	Mean transfer propensity
V	Defense mechanisms	32	0.4720	0.4361331747
T	transduction	88	0.4201	0.3955678194
G	Carbohydrate	186	0.4086	0.3984357156
Q	Secondary metabolites	73	0.4036	0.3926403801
L	Replication	179	0.4012	0.3849087174
P	Inorganic ion	199	0.4011	0.3884050415
O	Post-translational modification	123	0.3986	0.3868587153
K	Transcription	131	0.3922	0.3659200418
I	Lipid	77	0.3875	0.3783924085
C	Energy	239	0.3854	0.3724628343
M	Cell wall/membrane	141	0.3845	0.3746415978
H	Coenzyme	165	0.3766	0.3704811363
E	Amino acid	226	0.3765	0.3684339442
F	Nucleotide	102	0.3577	0.3629977725
N	Cell motility	70	0.3174	0.3198036954
D	Cell cycle	45	0.3150	0.3094970449
U	Intracellular trafficking	83	0.3146	0.3119251964
J	Translation	177	0.3093	0.312946966

Table S10: Median and mean transfer propensity T/(V+T) by COG functional category.

COG category	Number at root (flat prior)	Number at root (ML prior, PP >= 0.95)	Number at root (ML prior, PP >= 0.8)	Number of families in category	ML O_R
J	3	125	157	177	6739.434
F	1	57	89	102	5294.183
L	4	42	100	179	1468.033
H	3	26	85	165	1393.54
-	0	0	1	2	1290.933
N	1	18	38	70	1270.758
E	2	45	116	226	1146.195
D	2	7	16	45	850.0477
P	4	16	64	199	650.1995
M	1	15	49	141	625.6504
C	4	22	57	239	581.1914
G	3	11	43	186	513.1411
I	0	1	12	77	457.9165
U	1	3	12	83	364.7495
K	1	3	10	131	260.9843
O	0	1	9	123	256.3592
S	6	9	69	1374	163.0099
B	0	0	1	5	126.1288
T	0	1	3	88	125.0658
V	0	0	0	32	76.72498
Q	0	0	0	73	1.196214
A	0	0	0	4	1
Z	0	0	0	2	1

Table S11: Estimated root origination rates and root presences by COG functional category.

Clade	No. of taxa sampled
Firmicutes	25
Actinobacteriota+Cyanobacteria+Chloroflexota	35
CPR	125
FCB+PVC+Elusimicrobiota	35
Proteobacteria	50
“Deltaproteobacteria” +Nitrospirota+Acidobacterota+Aquificota	30
FASST+environmental lineages	25
New Phyla (Parks et 2018)	17

Table S12: Number of taxa sampled from each clade in the GTDB-independent analysis.

	Focal root 1	Focal root 2	Focal root 3	Secondary root 1	Secondary root 2
Focal root 1	475/3782	0.94	0.89	0.53	0.50
Focal root 2	403/3782	502/3782	0.9	0.53	0.51
Focal root 3	271/3782	265/3782	272/3782	0.49	0.48
Secondary root 1	364/3677	375/3677	210/3677	1383/4220	0.96
Secondary root 2	360/3677	378/3677	213/3677	1294/4220	1468/4220

Table S13. Comparison of ancestral gene contents between the focal and secondary datasets for different roots. Values with a white background correspond to correlation coefficients between PPs for COGs shared between two candidate roots (Pearson's correlation, all values highly significant with $p<10^{-16}$); values with gray background correspond to the number of shared COGs with PP>0.9 / total number of shared COGs, with the diagonal fields giving corresponding values for each root on its own.

References and Notes

1. T. A. Williams, G. J. Szöllősi, A. Spang, P. G. Foster, S. E. Heaps, B. Boussau, T. J. G. Ettema, T. M. Embley, Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4602–E4611 (2017). doi:10.1073/pnas.1618463114 Medline

2. J. R. Brown, W. F. Doolittle, Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 2441–2445 (1995). doi:10.1073/pnas.92.7.2441 Medline

3. J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, M. Yoshida, Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6661–6665 (1989). doi:10.1073/pnas.86.17.6661 Medline

4. N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9355–9359 (1989). doi:10.1073/pnas.86.23.9355 Medline

5. O. Zhaxybayeva, P. Lapierre, J. P. Gogarten, Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227**, 53–64 (2005). doi:10.1007/s00709-005-0135-1 Medline

6. F. U. Battistuzzi, S. B. Hedges, A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26**, 335–343 (2009). doi:10.1093/molbev/msn247 Medline

7. M. Bocchetta, S. Gribaldo, A. Sanangelantoni, P. Cammarano, Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J. Mol. Evol.* **50**, 366–380 (2000). doi:10.1007/s002399910040 Medline

8. C. Brochier, H. Philippe, A non-hyperthermophilic ancestor for Bacteria. *Nature* **417**, 244 (2002). doi:10.1038/417244a Medline

9. T. Cavalier-Smith, Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 19 (2006). doi:10.1186/1745-6150-1-19 Medline

10. J. A. Lake, Evidence for an early prokaryotic endosymbiosis. *Nature* **460**, 967–971 (2009). doi:10.1038/nature08183 Medline

11. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016). doi:10.1038/nmicrobiol.2016.48 Medline

12. S. Mukherjee, R. Seshadri, N. J. Varghese, E. A. Eloe-Fadrosh, J. P. Meier-Kolthoff, M. Göker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H.-P. Klenk, N. C. Kyrpides, 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017). doi:10.1038/nbt.3886 Medline

13. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018). doi:10.1038/nbt.4229 Medline

14. D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017). doi:10.1038/s41564-017-0012-7 Medline

15. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, J. F. Banfield, Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015). doi:10.1038/nature14486 Medline

16. Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab, R. Knight, Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019). doi:10.1038/s41467-019-13443-4 Medline

17. C. J. Castelle, J. F. Banfield, Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018). doi:10.1016/j.cell.2018.02.016 Medline

18. C. J. Castelle, C. T. Brown, K. Anantharaman, A. J. Probst, R. H. Huang, J. F. Banfield, Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018). doi:10.1038/s41579-018-0076-2 Medline

19. J. P. Beam, E. D. Becraft, J. M. Brown, F. Schulz, J. K. Jarett, O. Bezuidt, N. J. Poulton, K. Clark, P. F. Dunfield, N. V. Ravin, J. R. Spear, B. P. Hedlund, K. A. Kormas, S. M. Sievert, M. S. Elshahed, H. A. Barton, M. B. Stott, J. A. Eisen, D. P. Moser, T. C. Onstott, T. Woyke, R. Stepanauskas, Ancestral absence of electron transport chains in Patescibacteria and DPANN. *Front. Microbiol.* **11**, 1848 (2020). doi:10.3389/fmicb.2020.01848 Medline

20. C. J. Castelle, K. C. Wrighton, B. C. Thomas, L. A. Hug, C. T. Brown, M. J. Wilkins, K. R. Frischkorn, S. G. Tringe, A. Singh, L. M. Markillie, R. C. Taylor, K. H. Williams, J. F. Banfield, Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015). doi:10.1016/j.cub.2015.01.014 Medline

21. R. Méheust, D. Burstein, C. J. Castelle, J. F. Banfield, The distinction of CPR bacteria from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019). doi:10.1038/s41467-019-12171-z Medline

22. A. Graybeal, Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17 (1998). doi:10.1080/106351598260996 Medline

23. S. M. Hedtke, T. M. Townsend, D. M. Hillis, Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* **55**, 522–529 (2006). doi:10.1080/10635150600697358 Medline

24. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005). doi:10.1111/j.1096-0031.2005.00059.x

25. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: The phylogenetic jury is still out. *Philos. Trans. R. Soc. London Ser. B* **370**, 20140329 (2015). doi:10.1098/rstb.2014.0329 Medline

26. J. A. Lake, R. G. Skophammer, C. W. Herbold, J. A. Servin, Genome beginnings: Rooting the tree of life. *Philos. Trans. R. Soc. London Ser. B* **364**, 2177–2185 (2009). doi:10.1098/rstb.2009.0035 Medline

27. R. G. Skophammer, J. A. Servin, C. W. Herbold, J. A. Lake, Evidence for a Gram-positive, eubacterial root of the tree of life. *Mol. Biol. Evol.* **24**, 1761–1768 (2007). doi:10.1093/molbev/msm096 Medline

28. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002). doi:10.1080/10635150290069913 Medline

29. G. J. Szöllosi, B. Boussau, S. S. Abby, E. Tannier, V. Daubin, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17513–17518 (2012). doi:10.1073/pnas.1202997109 Medline

30. L. A. David, E. J. Alm, Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011). doi:10.1038/nature09649 Medline

31. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013). doi:10.1093/sysbio/syt054 Medline

32. L. A. Katz, J. R. Grant, L. W. Parfrey, J. G. Burleigh, Turning the crown upside down: Gene tree parsimony roots the eukaryotic tree of life. *Syst. Biol.* **61**, 653–660 (2012). doi:10.1093/sysbio/sys026 Medline

33. D. M. Emms, S. Kelly, STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017). doi:10.1093/molbev/msx259 Medline

34. A. Zwaenepoel, Y. Van de Peer, Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol. Biol. Evol.* **36**, 1384–1404 (2019). doi:10.1093/molbev/msz088 Medline

35. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008). doi:10.1093/molbev/msn067 Medline

36. See supplementary materials.

37. D. Posada, T. R. Buckley, Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808 (2004). doi:10.1080/10635150490522304 Medline

38. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002). doi:10.1093/nar/30.7.1575 Medline

39. M. Y. Galperin, D. M. Kristensen, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Microbial genome analysis: The COG approach. *Brief. Bioinform.* **20**, 1063–1070 (2019). doi:10.1093/bib/bbx117 Medline

40. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6670–6675 (2015). doi:10.1073/pnas.1420858112 Medline

41. P. S. Adam, G. Borrel, S. Gribaldo, Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E1166–E1173 (2018). doi:10.1073/pnas.1716667115 Medline

42. A. A. Davín, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, G. J. Szöllősi, Gene transfers can date the tree of life. *Nat. Ecol. Evol.* **2**, 904–909 (2018). doi:10.1038/s41559-018-0525-3 Medline

43. C. Chauve, A. Rafiey, A. A. Davín, C. Scornavacca, P. Veber, B. Boussau, G. J. Szöllősi, V. Daubin, E. Tannier, MaxTiC: fast ranking of a phylogenetic tree by maximum time consistency with lateral gene transfers. bioRxiv 127548 [Preprint]. 6 October 2017. https://doi.org/10.1101/127548.

44. T. Dagan, W. Martin, Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 870–875 (2007). doi:10.1073/pnas.0606318104 Medline

45. W. F. Doolittle, Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128 (1999). doi:10.1126/science.284.5423.2124 Medline

46. W. F. Doolittle, E. Bapteste, Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2043–2049 (2007). doi:10.1073/pnas.0610699104 Medline

47. T. Dagan, W. Martin, The tree of one percent. *Genome Biol.* **7**, 118 (2006). doi:10.1186/gb-2006-7-10-118 Medline

48. D. Alvarez-Ponce, P. Lopez, E. Bapteste, J. O. McInerney, Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E1594–E1603 (2013). doi:10.1073/pnas.1211371110 Medline

49. R. Jain, M. C. Rivera, J. A. Lake, Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801–3806 (1999). doi:10.1073/pnas.96.7.3801 Medline

50. P. Puigbò, Y. I. Wolf, E. V. Koonin, The tree and net components of prokaryote evolution. *Genome Biol. Evol.* **2**, 745–756 (2010). doi:10.1093/gbe/evq062 Medline

51. R. Liu, H. Ochman, Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7116–7121 (2007). doi:10.1073/pnas.0700266104 Medline

52. F. El Baidouri, C. Venditti, S. Suzuki, A. Meade, S. Humphries, Phenotypic reconstruction of the last universal common ancestor reveals a complex cell. bioRxiv 2020.08.20.260398 [Preprint]. 21 August 2020. https://doi.org/10.1101/2020.08.20.26039.

53. I. Sela, Y. I. Wolf, E. V. Koonin, Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11399–11407 (2016). doi:10.1073/pnas.1614083113 Medline

54. L. C. Antunes, D. Poppleton, A. Klingl, A. Criscuolo, B. Dupuy, C. Brochier-Armanet, C. Beloin, S. Gribaldo, Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *eLife* **5**, e14589 (2016). doi:10.7554/eLife.14589 Medline

55. D. Megrian, N. Taib, J. Witwinowski, C. Beloin, S. Gribaldo, One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol. Microbiol.* **113**, 659–671 (2020). doi:10.1111/mmi.14469 Medline

56. N. Taib, D. Megrian, J. Witwinowski, P. Adam, D. Poppleton, G. Borrel, C. Beloin, S. Gribaldo, Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat. Ecol. Evol.* **4**, 1661–1672 (2020). doi:10.1038/s41559-020-01299-7 Medline

57. E. I. Tocheva, D. R. Ortega, G. J. Jensen, Sporulation, bacterial cell envelopes and the origin of life. *Nat. Rev. Microbiol.* **14**, 535–542 (2016). doi:10.1038/nrmicro.2016.85 Medline

58. I. Sánchez-Andrea, I. A. Guedes, B. Hornung, S. Boeren, C. E. Lawson, D. Z. Sousa, A. Bar-Even, N. J. Claassens, A. J. M. Stams, The reductive glycine pathway allows autotrophic growth of *Desulfovibrio desulfuricans*. *Nat. Commun.* **11**, 5090 (2020). doi:10.1038/s41467-020-18906-7 Medline

59. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016). doi:10.1038/nmicrobiol.2016.116 Medline

60. F. L. Sousa, W. F. Martin, Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism. *Biochim. Biophys. Acta Bioenerg.* **1837**, 964–981 (2014). doi:10.1016/j.bbabio.2014.02.001 Medline

61. F. L. Sousa, S. Nelson-Sathi, W. F. Martin, One step beyond a ribosome: The ancient anaerobic core. *Biochim. Biophys. Acta Bioenerg.* **1857**, 1027–1038 (2016). doi:10.1016/j.bbabio.2016.04.284 Medline

62. G. Borrel, P. S. Adam, S. Gribaldo, Methanogenesis and the Wood–Ljungdahl pathway: An ancient, versatile, and fragile association. *Genome Biol. Evol.* **8**, 1706–1711 (2016). doi:10.1093/gbe/evw114 Medline

63. G. Fuchs, Alternative pathways of carbon dioxide fixation: Insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011). doi:10.1146/annurev-micro-090110-102801 Medline

64. T. Nunoura, Y. Chikaraishi, R. Izaki, T. Suwa, T. Sato, T. Harada, K. Mori, Y. Kato, M. Miyazaki, S. Shimamura, K. Yanagawa, A. Shuto, N. Ohkouchi, N. Fujita, Y. Takaki, H. Atomi, K. Takai, A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science* **359**, 559–563 (2018). doi:10.1126/science.aao3407 Medline

65. K. Schuchmann, V. Müller, Autotrophy at the thermodynamic limit of life: A model for energy conservation in acetogenic bacteria. *Nat. Rev. Microbiol.* **12**, 809–821 (2014). doi:10.1038/nrmicro3365 Medline

66. K. S. Makarova, D. H. Haft, R. Barrangou, S. J. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. M. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost, E. V. Koonin, Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011). doi:10.1038/nrmicro2577 Medline

67. E. V. Koonin, K. S. Makarova, Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. London Ser. B* **374**, 20180087 (2019). doi:10.1098/rstb.2018.0087 Medline

68. J. K. Nuñez, P. J. Kranzusch, J. Noeske, A. V. Wright, C. W. Davies, J. A. Doudna, Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014). doi:10.1038/nsmb.2820 Medline

69. K. S. Makarova, Y. I. Wolf, O. S. Alkhnbashi, F. Costa, S. A. Shah, S. J. Saunders, R. Barrangou, S. J. J. Brouns, E. Charpentier, D. H. Haft, P. Horvath, S. Moineau, F. J. M. Mojica, R. M. Terns, M. P. Terns, M. F. White, A. F. Yakunin, R. A. Garrett, J. van der Oost, R. Backofen, E. V. Koonin, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015). doi:10.1038/nrmicro3569 Medline

70. R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007). doi:10.1126/science.1138140 Medline

71. E. V. Koonin, The origins of cellular life. *Antonie van Leeuwenhoek* **106**, 27–41 (2014). doi:10.1007/s10482-014-0169-5 Medline

72. M. Krupovic, V. V. Dolja, E. V. Koonin, Origin of viruses: Primordial replicators recruiting capsids from hosts. *Nat. Rev. Microbiol.* **17**, 449–458 (2019). doi:10.1038/s41579-019-0205-6 Medline

73. A. C. J. Roth, G. H. Gonnet, C. Dessimoz, Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008). doi:10.1186/1471-2105-9-518 Medline

74. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:10.1093/molbev/mst010 Medline

75. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010). doi:10.1186/1471-2148-10-210 Medline

76. L. T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015). doi:10.1093/molbev/msu300 Medline

77. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015). doi:10.1038/nmeth.3176 Medline

78. A. A. Davín, T. Tricou, E. Tannier, D. M. de Vienne, G. J. Szöllősi, Zombi: A phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics* **36**, 1286–1288 (2020). doi:10.1093/bioinformatics/btz710 Medline

79. J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, P. Bork, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017). doi:10.1093/molbev/msx148 Medline

80. G. Coleman, A. Davín, T. Mahendrarajah, A. Spang, P. Hugenholtz, G. J. Szöllősi, T. Williams, Extended Data for A rooted phylogeny resolves early bacterial evolution, Version 9, Figshare (2020); https://doi.org/10.6084/m9.figshare.12651074.v9.

81. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of

gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020). doi:10.1038/s41467-020-17408-w Medline

82. M. Wilkinson, J. O. McInerney, R. P. Hirt, P. G. Foster, T. M. Embley, Of clades and clans: Terms for phylogenetic relationships in unrooted trees. *Trends Ecol. Evol.* **22**, 114–115 (2007). doi:10.1016/j.tree.2007.01.002 Medline

83. S. M. van Dongen, "Graph clustering by flow simulation," thesis, University of Utrecht, (2000).

84. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015). doi:10.1101/gr.186072.114 Medline

85. G. J. Szöllősi, E. Tannier, N. Lartillot, V. Daubin, Lateral gene transfer from the dead. *Syst. Biol.* **62**, 386–397 (2013). doi:10.1093/sysbio/syt003 Medline

86. H.-C. Wang, B. Q. Minh, E. Susko, A. J. Roger, Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018). doi:10.1093/sysbio/syx068 Medline

87. E. Susko, A. J. Roger, On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007). doi:10.1093/molbev/msm144 Medline

88. C. Zhang, E. Sayyari, S. Mirarab, "ASTRAL-III: Increased scalability and impacts of contracting low support branches" in *Comparative Genomics*, J. Meidanis, L. Nakhleh, Eds., vol. 10562 of *Lecture Notes in Computer Science* (Springer, Cham, 2017), pp. 53–75.

89. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011). doi:10.1371/journal.pcbi.1002195 Medline

90. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011). doi:10.1093/nar/gkr367 Medline

91. T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020). doi:10.1093/bioinformatics/btz859 Medline

92. P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014). doi:10.1093/bioinformatics/btu031 Medline

93. J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, P. Bork, eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019). doi:10.1093/nar/gky1085 Medline

94. J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, P. Bork, eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for

eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016). doi:10.1093/nar/gkv1248 Medline

95. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003). doi:10.1186/1471-2105-4-41 Medline