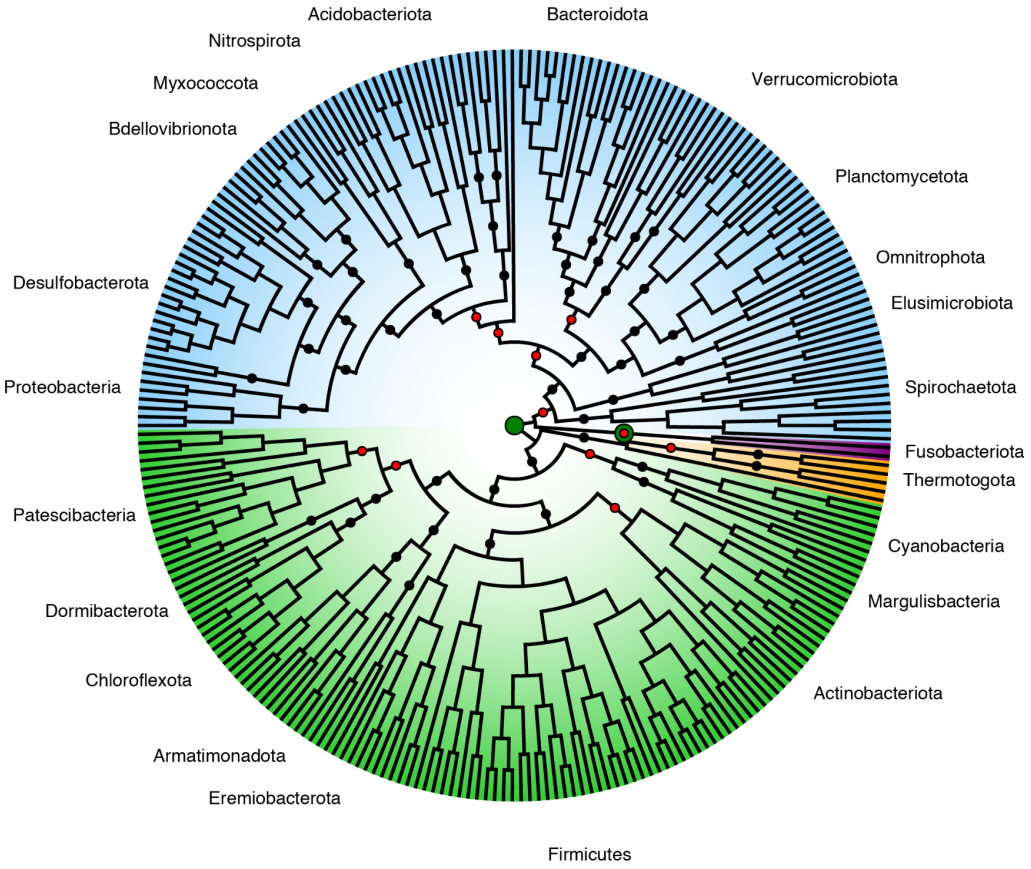
Parameter estimation and species tree rooting using ALE and GeneRax: Supplementary Material

Supplementary Figures



Supplementary Figure 1: Rooting analysis using different sets of parameters. We divided the gene families used in Coleman et al. (2021) into two groups: those that preferred a model with 3 parameters and those that preferred a model with 2 parameters in which the τ and δparameters are fixed to a constant ratio. The 5930 families that prefer the 3 parameters support a root region indicated by the large two green dots. The 5342 that preferred the 2 parameters model (represented by the red dots) are in general smaller families and the root region supported by them is more diffuse. The remaining tested roots that were rejected by both sets of families appear as black dots. In this analysis, we divided genes into the two sets using the two-step inference procedure of Bremer et al. (2022), but dividing families according to their preference after jointly optimising τ:δ (with 5908 families rejecting the simple model by AIC) produced the same result.

Supplementary Text

Genome evolutionary dynamics are best captured by branch-wise metrics

In the following, we provide a more detailed explanation of the distinction between δ, τ and λ parameters (estimated per-family in ALE/GeneRax); the number of inferred duplication, transfer and loss events per-family (which result from analysing the data using the model); and the number of inferred duplication, transfer and loss events per branch of the species tree, to supplement the discussion of Figure 3 in the main text.

Frequencies of duplication, transfer and loss vary across gene families (e.g., ribosomal proteins are transferred less frequently than metabolic genes (Jain, Rivera, and Lake 1999)) and across lineages (gene transfers appear to be more frequent in Bacteria than in eukaryotes). The δ, τ and λ parameters in ALE are estimated for each family, but do not vary over the species tree. Starting from gene trees, doing inference with this model gives rise to inferred duplication, transfer and loss events that map to branches of the species tree. These lineage-specific events can then be summarised to give branch-wise counts of inferred duplication, transfer and loss events that can be used to compare genome dynamics between lineages. The δ, τ and λ parameters, however, are not representative of lineage-level dynamics (or across-lineage variation) because they are estimated family-wise (Figure 3). Each family-wise ALE output file lists inferred events for each branch of the species tree, and branch-wise counts of inferred duplications, transfers and losses are a natural way to summarise these inferences; these various parameters and summary statistics are plotted in Figure 2; a Python script to perform these calculations is provided at https://github.com/AADavin/ALEtutorial. In their critique, Bremer et al. (2022) compared family-wise δ and τ parameters to per-genome counts of events from previous analyses.

A second potential source of error relates to using overall averages to summarise gene family dynamics, when in fact gene family size and δ, τ and λ parameter values vary greatly across families. For example, the global means of τ and δ across all families in the bacterial dataset are 0.42 and 0.19, respectively, and this might indeed suggest that frequencies of transfer and duplication are roughly 2:1 in this dataset. However, the global means (or medians) of these parameters do not capture the substantial variation in frequencies of duplication, transfer and loss across families: note that the means of the τ/δ ratios for Bacteria and opisthokonts are both very far from the ratio of the means (Figure 3), and it is these per-family dynamics that underlie inferred events. We also note that Bremer et al. seem to have calculated metrics based on model parameters, not on model inferences — that is, inferred events. We would recommend metrics based on inferred events, rather than model parameters. The fundamental reason is that inferences have a clear biological interpretation, whereas model parameters are of secondary interest and, in a data-homogeneous model, do not reflect biologically interesting variation in the process we are attempting to model. Consider an analogy to studying substitution histories in standard phylogenetics: we might wish to know which substitutions occurred where (on which branch of the tree), or how many substitutions occurred per year or along a branch of interest. In a

homogeneous substitution model, the equilibrium frequencies of the amino acids and the instantaneous rates of change among them are fixed. Despite this, inference using the model reveals variation in branch lengths, rates and patterns of substitution across the tree. These patterns reflect information in the data that is captured by the model. In the same way, the most relevant output from fitting a reconciliation model is the evolutionary scenario and the associated inferred duplication, transfer and loss events which, like substitutions in the standard phylogenetic analysis, occur along branches of the species tree.

Beyond the parameter/inference distinction, there is an additional drawback to metrics based on per-family ratios: many families experience 0 duplications (e.g., 5,314/11,272 families in the bacterial dataset), so their T/D (expressed as inferred events) is infinite, and τ/δ tends to infinity (δ ~ 0 when zero duplications are inferred); similarly, families with 0 transfers (of which there are 9673 in the opisthokont dataset) will have τ/δ ~ 0. Thus, family-wise T/D and τ/δ range from 0 to infinity. This does not present a problem for inference using ALE, because the per-family τ and δ parameters are estimated separately; however, it makes the per-family parameter ratios a noisy, unstable and potentially misleading metric. For example, the median per-family ratio of τ to δ is 0 for the opisthokont dataset and 7 for the Bacteria dataset, while the means are 7×10^7 for opisthokonts and 1×10^9 for Bacteria; the enormous difference between means and medians demonstrates that these summaries inadequately describe the evolutionary dynamics of the underlying gene families. Bremer et al.'s remark (Bremer et al. 2022) that the range of per-family T/D rate ratios inferred by ALE is unreasonably wide is most likely based upon conflating per-family τ/δ ratios (which they calculated for the ALE analyses) with per-*genome* counts of inferred *events* — transfers and duplications — to which they compared ALE's per-family values. Perhaps the clearest example of why expressing evolutionary dynamics as a τ/δ ratio can be unhelpful comes from the analysis of the 117 single-copy marker genes of the opisthokont dataset. Bremer et al. (2022) reported that the mean τ/δ for these genes was 26968:1, which they regarded as abnormally high. However, the high ratio simply results from the fact that ALE infers 0 duplications for all 117 families; ALE estimates verticality at 0.99 for these genes, which seems reasonable for a set of generally vertically-evolving marker genes (Figure 3).

Reconciliation-based estimates of genome size evolution

In a further critique of reconciliation-based analyses, Bremer et al. investigated the per-family ratio of the λ to gene gain rate (which they defined as $\tau+\delta$) in the 11,272 families of the bacterial dataset. A “loss/gain ratio” > 1 denotes a gene family with a general tendency to contract over time, while a ratio <1 indicates a family that tends to expand over time (that is, where family expansion by gene duplication, or acquisition by transfer, outweighs losses). They reported a mean loss/gain ratio of 0.76, which they argued was unrealistic given that bacterial genes are lost more often than they are acquired (Bremer et al. 2022). This calculation and inference appear to be incorrect, as we briefly explain below.

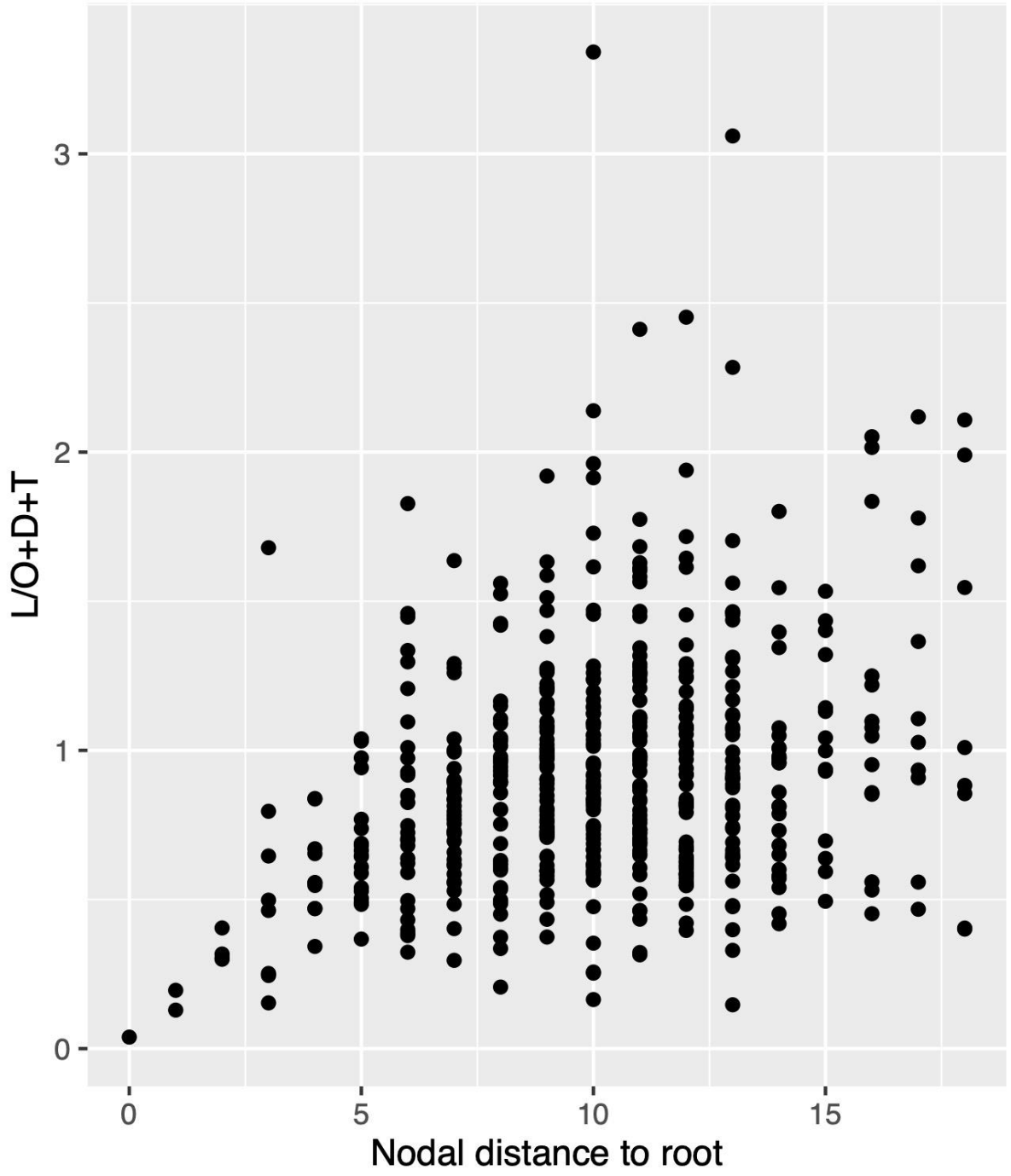
Bremer et al. (2022) calculated the loss/gain ratio by dividing the mean λ by the sum of the mean of δ and the mean of τ, to obtain an estimate of 0.76. However, the ratio of two means is not the mean of their ratios, and in fact the actual mean $\lambda/\delta+\tau$ for the bacterial dataset ranges from 1.23×10^7 - 1.37×10^7 across the three branches of the root region; the median is 1 for all three root branches. The enormous difference between the means and

medians arises because the data are not normally distributed: as discussed above, many families experience 0 Ds, Ts or Ls, and so family-wise ratios of parameters or events are highly unstable. Thus, judged by this $\lambda/\delta+\tau$ metric, losses generally equal or exceed gains among the gene families of the bacterial dataset.

However, it is important to note that this per-family $\lambda/\delta+\tau$ metric does not provide insight into changes in gene content over time, because it refers to per-family (not per-genome) δ, τ and λ parameters (not events). In order to track changes in gene content, we need to map gain and loss events onto the species tree. We also need to consider gene originations, an important source of new genes that is not considered in the $\lambda/\delta+\tau$ metric.

To summarise ALE-based estimates of gene content change through time, we calculated the branch-wise number of gene losses, divided by the sum of gene originations, duplications and transfers (L/O+D+T). The mean and median branch-wise values are 0.9213 and 0.8777, respectively, suggesting that gains slightly outweigh losses overall in the bacterial dataset (that is, on average ~0.92 losses occur for every gene gain resulting from an origination, acquisition by transfer, or gene duplication). Does this result imply that bacterial gene repertoires have increased continuously through time? No, because these values have been inferred from gene families that have survived to be sampled in the present day. Note that, for a gene family to go extinct, within-family losses must equal or exceed gains. Gene families with L > O+D+T are the only ones that go extinct, and so our present-day sample of families is biased towards families with lower L/O+D+T values; in other words, we would expect these values, given that we base our inferences on extant gene families. This effect can be readily seen in the bacterial dataset by plotting L/O+D+T as a function of their distance from the root of the tree (Figure S2): the ratio is lowest near the root, and rises above 1 on recent branches. This is because the families which inform L/O+D+T on a branch must have survived from that branch to the present day; the ratio therefore appears lower on more ancient branches.

This is not to say that current reconciliation models fully capture changes in genome dynamics. For example, in the ALE/GeneRax model, δ, τ and λ parameters are homogeneous over the tree as well as independent of gene family copy number. In reality, not only do rates of duplication, transfer and loss vary across branches, but there is strong evidence that they also depend on the number of homologous gene copies in a particular genome (Huynen and van Nimwegen 1998). Nonetheless, published ALE results seem plausible, identifying the large loss of gene content in the common ancestor of the CPR clade ((Brown et al. 2015; Méheust et al. 2019) and major gene losses near the tips of the tree leading to parasitic and endosymbiotic lineages, in agreement with alternative (non-reconciliation) approaches (McCutcheon and Moran 2011).



Supplementary Figure 2: Survivorship bias in estimates of gene loss and gain in the bacterial dataset. When calculated from gene families that have survived to the present day, the ratio of gene losses to gains (originations, duplications, and acquisition by transfer) along the species tree shows a clear survivorship bias: branches closest to the root experience more gains than losses, while the balance shifts on recent branches. This is because gene families present in the past that survived to the present day tend to experience more gains than losses; gene families in which losses exceed gains go extinct and are not sampled by present-day phylogeneticists.

Supplementary References

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, and Jillian F. Banfield. 2015. “Unusual Biology across a Group Comprising More than 15% of Domain Bacteria.” *Nature* 523 (7559): 208–11.

Huynen, M. A., and E. van Nimwegen. 1998. “The Frequency Distribution of Gene Family Sizes in Complete Genomes.” *Molecular Biology and Evolution* 15 (5): 583–89.

Jain, R., M. C. Rivera, and J. a. Lake. 1999. “Horizontal Gene Transfer among Genomes: The Complexity Hypothesis.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (7): 3801–6.

McCutcheon, John P., and Nancy A. Moran. 2011. “Extreme Genome Reduction in Symbiotic Bacteria.” *Nature Reviews. Microbiology* 10 (1): 13–26.

Méheust, Raphaël, David Burstein, Cindy J. Castelle, and Jillian F. Banfield. 2019. “The Distinction of CPR Bacteria from Other Bacteria Based on Protein Family Content.” *Nature Communications* 10 (1): 4173.

Szöllősi, Gergely J., Adrián Arellano Davín, Eric Tannier, Vincent Daubin, and Bastien Boussau. 2015. “Genome-Scale Phylogenetic Analysis Finds Extensive Gene Transfer among Fungi.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1678): 20140335.