nature communications

Article

https://doi.org/10.1038/s41467-025-64881-2

# Augmenting microbial phylogenomic signal with tailored marker gene sets

Received: 17 March 2025

Accepted: 25 September 2025

Published online: 12 November 2025

Check for updates

Henry Secaira-Morocho[1,2], Xiaofang Jiang[2] & Qiyun Zhu[1]

Phylogenetic marker genes are traditionally selected from a fixed collection of whole genomes representing major microbial phyla, covering only a small fraction of gene families. However, most microbial diversity resides in metagenome-assembled genomes, which exhibit taxonomic imbalance and harbor gene families that do not fit the criteria for universal orthologs. To address these limitations, we introduce TMarSel, a software tool for automated, free-from-expert opinion, and tailored marker selection for deep microbial phylogenomics. TMarSel allows users to select a variable number of markers and copies based on KEGG and EggNOG gene family annotations, enabling a systematic evaluation of the phylogenetic signal from the entire gene family pool. We show that an expanded marker selection tailored to the input genomes improves the accuracy of phylogenetic trees across simulated and real-world datasets of whole genomes and metagenome-assembled genomes compared to previous markers, even when metagenome-assembled genomes lack a fraction of open reading frames. The selected markers have functional annotations related to metabolism, cellular processes, and environmental information processing, in addition to replication, translation, and transcription. TMarSel provides flexibility in the number of markers, copies, and annotation databases while remaining robust against taxonomic imbalance and incomplete genomic data.

Phylogenetic trees serve as the cornerstone for studies ranging from estimating the age of lineages[1,2] to comparative genomics[3] and microbial community ecology[4,5], as they recapitulate the evolutionary history of species[6,7]. Inference of phylogenetic trees relies on identifying phylogenetic markers from homologous sequences that descend vertically from a common ancestor (orthologs)[8–10]. In addition to orthologs, microbes harbor homologous genes that have been exchanged through horizontal gene transfer (HGT)[11,12]. The deep divergence times of microbes, estimated at around 4 billion years[1], have entangled homologs into complex relationships that obscure the precise identification of orthologs[8,10]. Nonetheless, new tree inference methods have been developed to bypass the identification of orthologs, thereby enabling the usage of all homologous sequences of a gene family as potential markers[13,14]. Because the downstream applications of phylogeny heavily depend on the tree quality, it is critical to select a combination of markers that yields the most accurate tree.

Metagenome-assembled genomes (MAGs) comprise the major genomic source of microbial diversity[15]. And yet, the gold standard 16S rRNA marker used for phylogenetic surveys of microbial diversity[16] is rarely recovered from shotgun metagenomic sequences[17,18]. Moreover, 16S rRNA-based trees reflect only the evolution of the gene rather than the set of species[19]. To improve the tree accuracy, modern phylogenetic surveys have adopted a larger number of markers involved in housekeeping functions, such as ribosomal proteins or aminoacyl-tRNA synthetases[20–24]. These markers have been selected from a fixed collection of whole genomes spanning the major microbial phyla, thereby biasing the representation of markers toward

[1]Center for Fundamental and Applied Microbiomics and School of Life Sciences, Arizona State University, Tempe, AZ, USA. [2]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. e-mail: qiyun.zhu@asu.edu

well-characterized taxa. In contrast, MAGs seldom contain the entire genomic repertoire of a population[25,26], and some even lack ribosomal proteins due to assembly errors[27], reducing the number of markers available for tree inference. Therefore, no one-size-fits-all set of markers exists. To account for the novel diversity and heterogeneous quality of MAGs, marker selection needs to be flexible and tailored to the input genome collection.

Although the inference of microbial trees has shifted from using a single marker[28,29] to multiple markers[20–24,30], their selection is restricted to universal orthologous genes, which are commonly defined as being present in 90% of genomes and existing as a single copy in at least 95% of them[31–33], severely limiting the number of markers considered. However, recent studies have shown that including gene families beyond standard universal orthologous genes leads to improvements in the accuracy of inferred trees[13,14]. While these studies highlighted the necessity for a comprehensive assessment of the phylogenetic signal from a broader set of genes, they did not attempt to select new markers. Here, we systematically select gene families to serve as markers for deep microbial phylogenomics tailored to the input genome collection. Leveraging recent advances in tree inference methods and genome annotation databases, we show that an expanded selection of markers yields species trees with higher accuracy than previous sets of markers in real-world datasets of whole genomes and MAGs derived from a wide range of environments. In addition to genes involved in replication, translation, and transcription, we found that markers have functional annotations related to metabolism, cellular processes, and environmental information processing, and all of them provide phylogenetic signals for tree inference. We also show that our automated, systematic, free-of-domain expertise, and tailored marker selection is robust against uneven taxon sampling and incomplete MAGs while remaining flexible in the number of markers to select and the choice of annotation database. Overall, we present a new method for Tailored Marker Selection (TMarSel), available as a software tool, that can be applied to modern genomic datasets, providing a foundation for more robust and accurate phylogenomic reconstruction.

## Results

### A vast yet unexplored gene family space for microbial phylogenomics

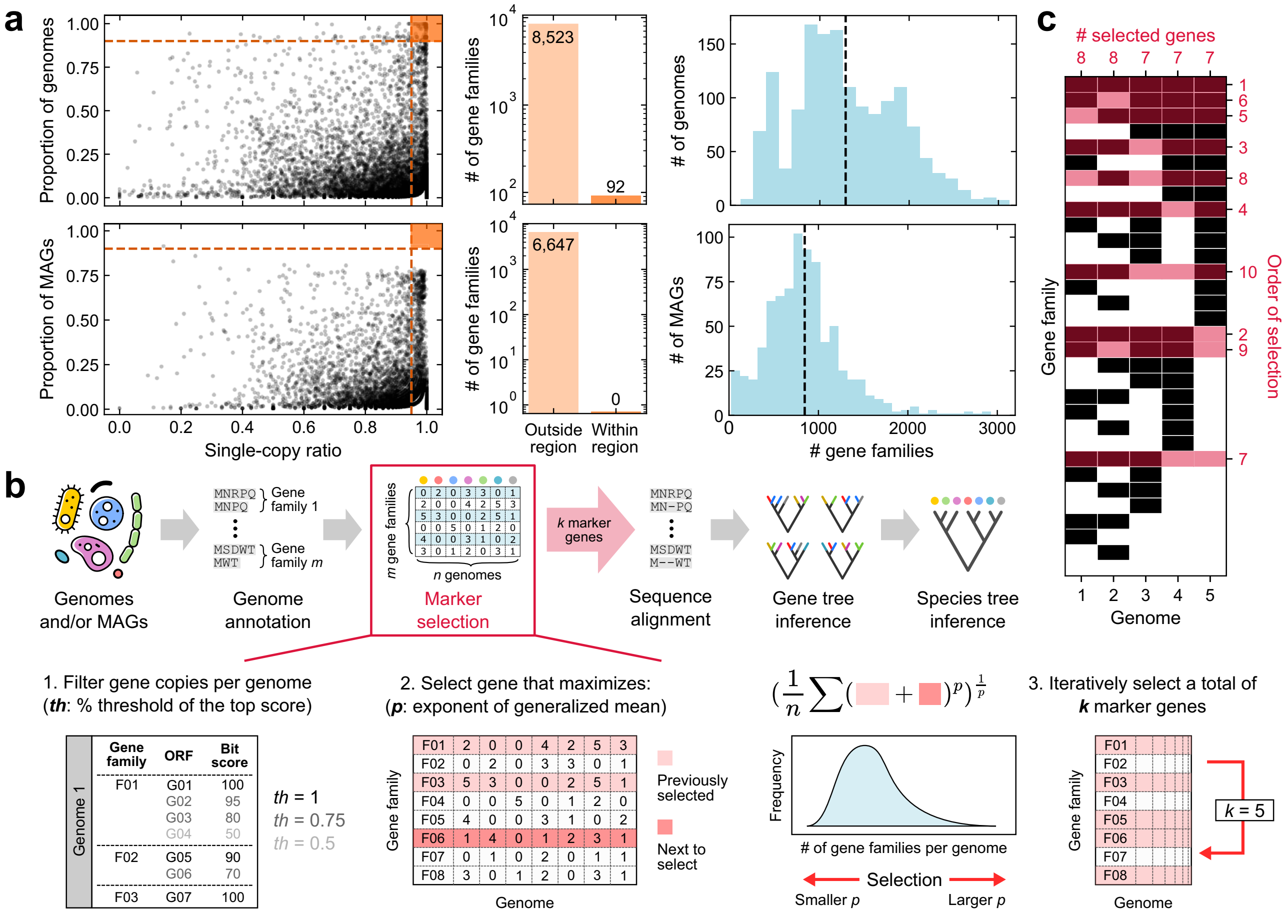
We surveyed a collection of 1510 whole reference genomes sampled across the microbial tree of life from the Web of Life 2 (WoL2) dataset to obtain an accurate representation of the gene family distribution in microbes. Specifically, we selected complete bacterial and archaeal genomes, ensuring that each taxonomic group, from phylum to family, had at least 10 taxa (Supplementary Fig. 1). We then annotated open reading frames (ORFs) of genomes into gene families using the KEGG and EggNOG databases, which resulted in 54% and 94% of ORFs annotated, respectively. KEGG gene families are scattered throughout genomes (Fig. 1a), ranging from universal to lineage-specific and from single-copy to multi-copy. The traditional criteria for marker selection are restricted to genes present in at least 90% of genomes and containing one copy in at least 95% of them[31–33]. We observe that only 1% of gene families annotated from the WoL2 genomes fall within the region defined by traditional criteria. This pattern highlights the limited number of gene families used for microbial phylogenomics. The limitation is further exacerbated in gene families annotated from 793 MAGs of the Earth Microbiome Project (EMP), despite a lower annotation rate, as only 47% and 87% of ORFs could be assigned to gene families with KEGG and EggNOG databases, respectively. This lower rate is mainly due to MAGs containing a large fraction of ORFs that cannot be assigned to gene families[34]. We observe that MAGs do not have gene families that conform to the traditional criteria. And yet, genomes and MAGs harbor, on average, 1289 and 846 gene families, respectively, that might add new phylogenetic signals to the tree inference process. EggNOG gene families share the same characteristics (Supplementary Fig. 2), suggesting that these trends are independent of the annotation database.

To assess whether these overlooked gene families can contribute phylogenetic signals to the inference process, we developed a robust yet flexible method, TMarSel, to select gene families annotated with the KEGG and EggNOG databases as markers for deep microbial phylogenomics. TMarSel takes as input a file mapping ORFs to gene families. The user can then control the copy number per genome and gene family using a threshold ranging from 0 to 1, where 0 includes all ORFs and 1 includes only the ORFs with the highest bit scores obtained during genome annotation. Next, TMarSel builds a matrix containing the copy number of gene families across genomes upon which it employs an algorithm to iteratively select $k$ markers such that the generalized mean number of markers per genome is maximized (see Methods for details; Fig. 1b and 1c). TMarSel uses a single core and its runtime scales sublinearly with the number of markers, while the memory footprint remains constant (Supplementary Fig. 3). As a reference, selecting 1000 markers from WoL2 and EMP datasets takes 10 min and requires 10 GB of memory. The selected markers are fed into a standard pipeline to infer species trees (Fig. 1b). Briefly, for each marker, we generate a multiple sequence alignment, which is used to infer a gene tree. Because gene trees can include all the homologs of a gene family, we used the summary method ASTRAL-Pro2[35,36], which takes as input a set of gene trees and all their homologs, to infer a species tree. We then evaluate the quality of the inferred species trees as a direct measure of the goodness of the selected marker sets.

### A well-balanced marker selection produces highly accurate trees in simulated data

TMarSel performs an iterative selection of markers tailored to the input genome collection. Its behavior can be customized with two parameters: the total number ($k$) of markers to select and the exponent $p$ of the generalized mean. Practically, $p$ biases the selection of markers toward families present in genomes with fewer gene families (if $p < 0$) or toward families present in genomes with more gene families (if $p > 0$; see Methods). To assess the impact these parameters have on the inferred trees, we first simulated a toy dataset of 50 gene families from 10 genomes across 25 replicates. In each replicate, we built a matrix containing the copy number of gene families across genomes (Supplementary Fig. 4a). We derived a species tree from the matrix using neighbor-joining over the Jaccard distances between genomes, upon which gene trees were also derived (see Methods). We then performed a parameter sweep for $k$ and $p$ and varied the maximum number of copies for each gene family and the noise present in gene trees. Noise refers to the proportion of leaves that have been randomly shuffled. We gauged the error in inferred trees as the normalized Robinson-Foulds (nRF) distance[37] between inferred and real trees, where smaller distances indicate fewer errors and vice versa. Our simulations show that selecting a large number of markers reduces the error in the species trees, primarily when noisy gene trees are used for inference (Fig. 2a and standard deviations in Supplementary Fig. 5a). The parameter sweep shows that $p \leq 0$ yields the species trees with fewer errors, with $p = 0$ as an inflection point. Moreover, we observe that having multiple copies of the same marker does not improve the inference process. Instead, they negatively impact quality, as errors increase with the number of copies.

We further assessed the impact of parameters on tree quality using a gene family space that resembles real datasets (Supplementary Fig. 4b). We simulated prokaryotic gene families using realistic duplication, transfer, and loss values[38] across 25 replicates using the phylogenetic simulator Zombi[39]. Each replicate yielded a species tree, genomes for each taxon, and gene trees for each gene family. We then built a matrix containing the copy number of gene families across taxa, introduced noise to the gene trees, and performed a parameter sweep for $k$ and $p$ (see Methods). In addition to the nRF, we gauged the



| Gene family | ORF | Bit score |
|---|---|---|
| F01 | G01 | 100 |
|  | G02 | 95 |
|  | G03 | 80 |
|  | G04 | 50 |
| F02 | G05 | 90 |
|  | G06 | 70 |
| F03 | G07 | 100 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| F01 | 2 | 0 | 0 | 4 | 2 | 5 | 3 |
| F02 | 0 | 2 | 0 | 3 | 3 | 0 | 1 |
| F03 | 5 | 3 | 0 | 0 | 2 | 5 | 1 |
| F04 | 0 | 0 | 5 | 0 | 1 | 2 | 0 |
| F05 | 4 | 0 | 0 | 3 | 1 | 0 | 2 |
| F06 | 1 | 4 | 0 | 1 | 2 | 3 | 1 |
| F07 | 0 | 1 | 0 | 2 | 0 | 1 | 1 |
| F08 | 3 | 0 | 1 | 2 | 0 | 3 | 1 |

$$\left(\frac{1}{n}\sum\left(\square+\blacksquare\right)^{p}\right)^{\frac{1}{p}}$$

**Fig. 1 | Expanding the set of phylogenetic markers with TMarSel. a** Summary statistics of KEGG gene families annotated from 1510 WoL2 genomes (top) and 793 EMP MAGs (bottom). The left panels show the gene family space defined by the proportion of genomes in which each gene family is present and the proportion of genomes in which the gene family exists as a single copy. Dashed orange lines represent traditional marker selection criteria (i.e., present in 90% of genomes and containing a single copy in at least 95% of them), and the orange region highlights the area defined by these criteria. Each dot corresponds to a gene family. The middle panels show the number of gene families outside and within the region defined by traditional criteria. The left panels show the distribution of gene families across genomes and MAGs, with dashed vertical lines indicating the arithmetic mean. **b** Illustration of the pipeline used for species tree inference (top) and detailed overview of TMarSel's steps for marker selection (bottom). **c** A simple toy case with 5 genomes and all 32 combinations of gene presence (black squares) or absence (white squares). TMarSel iteratively selects $k = 10$ markers (crimson shade) that maximize the generalized mean of the number of selected genes per species, as denoted on top. Source data are provided as a Source Data file.

proportion of confident branches in the inferred trees as the number of branches with a Local Posterior Probability (LPP) score higher than 0.95 out of the total number of branches, as suggested in ref. 40. A higher proportion is an indicator of a more accurate tree topology. As in the previous simulations, our results indicate that a large number of markers reduces the error and reinforces $p = 0$ as an inflection point that achieves the lowest error in the inferred trees (Fig. 2b; standard deviations are provided in Supplementary Fig. 5b). The proportion of confident branches shows a similar pattern for $k$ and $p$ and is correlated with the error in the inferred trees. This indicates that trees with fewer errors have a higher proportion of confident branches and vice versa. The simulations also show that a saturation value occurs for $k$, at which the accuracy of inferred trees does not improve. In light of these results, we choose $p = 0$, which refers to maximizing the geometric mean number of markers per genome, for all subsequent analyses while varying $k$ until reaching a plateau in quality.

## Expanded marker selection improves the accuracy of the microbial tree of life

We next evaluated whether gene families annotated from the genomes of the WoL2 can serve as the foundational genetic elements for an accurate microbial tree of life. Because the simulation results suggested that multiple copies of the same gene family negatively impact the tree inference process (Fig. 2a), we first evaluated how many copies of each gene family should be included in the analyses. Using the bit score threshold assigned during genome annotation with KEGG and EggNOG databases, we controlled the number of copies of gene families (see Methods). Our results indicate that using a low number of copies yields the tree with the highest quality as measured by the proportion of confident branches (i.e., accuracy in topology) and nRF distance to the reference phylogeny (Supplementary Fig. 6). Consequently, we included only the ORFs with the highest bit score assigned to each gene family during genome annotation for marker selection.

We then benchmarked the performance of an expanded selection of markers for inferring the microbial tree of life. Fig. 3a, b show the quality of trees inferred using an increasing number of markers, ranging from $k = 10$ to 1000. For comparison, we also included four sets of previously proposed universal markers widely adopted for deep microbial phylogenomics[1,23,30,41]. Trees inferred with our markers exhibit the highest accuracy in topology and lower error to the WoL2 reference phylogeny compared to previous markers. However, the 400 universal markers from PhyloPhlAn3 achieve the lowest error, which is expected because the WoL2 phylogeny was reconstructed

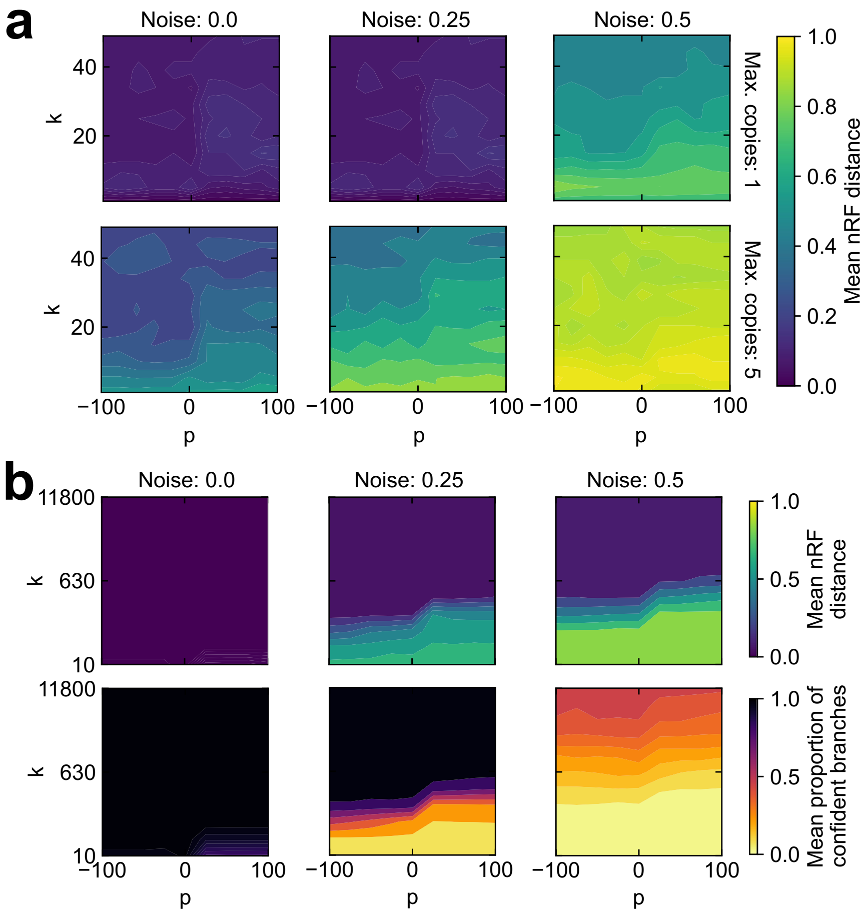


**Fig. 2 | Impact of marker selection parameters on trees inferred from simulated data. a** Contour plots of the mean normalized Robinson-Foulds (nRF) distance between inferred and real trees from simulated datasets of 50 gene families from 10 genomes across $n = 25$ replicates. **b** Contour plots of the mean nRF distance between inferred and real (top) and proportion of confident branches in inferred trees (bottom) from simulations of realistic prokaryotic gene families across $n = 25$ replicates. Marker selection was applied to each combination of the number of markers ($k$) and exponent ($p$). Each plot illustrates a scenario with different noise levels in gene trees (**a**, **b**) and maximum copies of gene families (only **a**). Source data are provided as a Source Data file.

using the PhyloplhAn3 markers. And yet, the high quality of trees inferred from either KEGG or EggNOG gene families speaks for the robustness of TMarSel for selecting markers. Moreover, the smooth saturation curve in tree quality suggests that for any given $k$, our method can select the best set of gene families for tree inference. Filtering genomes with less than 25% of markers leads to trees with slightly lower quality across marker sets. Nonetheless, more markers per genome result in better species placement within the tree (Supplementary Fig. 7). Additionally, trees inferred from all marker sets exhibit clades consistent with the GTDB taxonomy (Supplementary Fig. 8), according to taxonomic accuracy metrics that measure the consistency between taxonomy and phylogeny (see Methods). Altogether, these results suggest that all marker sets recover known relationships among clades, while markers derived from KEGG and EggNOG gene families yield more accurate trees compared to previous sets.

Because TMarSel iteratively selects more than twice the markers compared to the largest set available (i.e., 400 markers from PhyloPhlAn3), we investigated the functional classification of these new markers. Fig. 3b shows the abundance of higher functional categories of markers selected from KEGG gene families. 60%, 66%, and 53% of markers have functions related to genetic information processing when 10, 50, and 100 markers are selected, suggesting that genes involved in replication, transcription, and translation are preferred. Yet, as more markers are added, the selection shifts toward gene families annotated as carrier proteins, shape-determining proteins, and others that are not assigned to higher functional categories (see Supplementary Data 1 for a complete description). 38% of gene families are involved in metabolic functions, while 4% and 5% are involved in cellular processes and environmental information processing, respectively, when more than 600 markers are selected. We also observe an overlap in the functional descriptions of KEGG and EggNOG gene families (Supplementary Data 1), which speaks to the robustness of selection. These results highlight the diverse functional roles of these new markers that contribute to the phylogenetic signal.

The prevalence of mobile genetic elements (MGEs) in genomes from major microbial phyla[42], suggests that genome annotations are likely to yield multiple MGE gene families. Therefore, we assessed whether gene families annotated as MGEs were selected as markers by surveying their functional description terms (see Methods). We found that putative transposases, integrases, and competence proteins were selected. Nonetheless, they account for less than 2% of markers and represent less than 13% and 6% of the total number of gene families annotated with MGE descriptions in the KEGG and EggNOG databases, respectively (Supplementary Fig. 9). This showcases the robustness of TMarSel against the over-selection of MGEs.

We also assessed the impact of imbalanced taxonomic distributions on the tree topology. To achieve this, we generated skewed datasets with many poorly represented taxa and a few over-represented taxa at the phylum level, based on the WoL2 genomes,

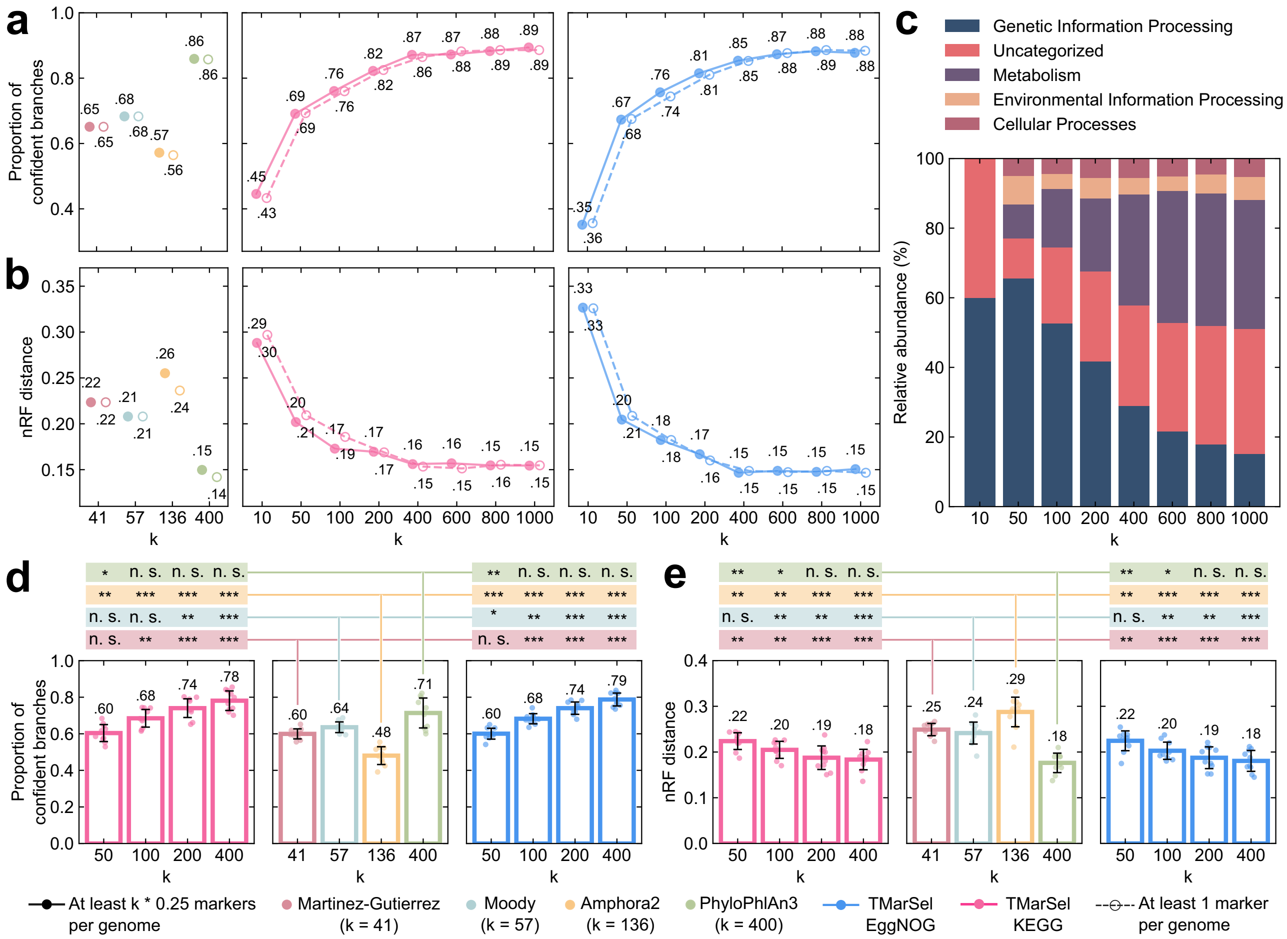


**Fig. 3 | Evaluating the performance and functional makeup of TMarSel-selected markers on genomes from the WoL2 dataset. a** Topological accuracy, measured as the proportion of confident branches, of inferred trees from 1510 WoL2 genomes. **b** nRF distance between inferred trees and WoL2 reference phylogeny. Left panels indicate the tree quality of previous marker sets, whereas middle and right panels show the tree quality of KEGG and EggNOG markers, respectively. Each dot corresponds to a tree inferred from different sets of markers (color-coded). Solid lines and filled symbols indicate trees derived from genomes containing at least 25% of markers, while dashed lines and unfilled symbols indicate trees from genomes with at least one marker. **c** Relative abundance of KEGG higher functional categories (color-coded) across markers selected from KEGG gene families. **d** Topological accuracy of trees inferred from taxonomically imbalanced datasets generated from the WoL2 genomes across $n = 10$ replicates and **e** nRF distance between inferred trees and the WoL2 reference phylogeny. Unfilled bars indicate that trees were inferred from genomes with at least one marker, and their color displays the marker set (color-coded). Data are presented as the mean ± standard deviations. Asterisks show the significance level of two-sided two-sample Mann–Whitney U tests after Benjamini–Hochberg correction between trees inferred from KEGG (left) and EggNOG (right) markers versus previous markers. Significance levels: n. s.: $p > 0.05$; *: $p \le 0.05$; **: $p \le 0.01$; ***: $p \le 0.001$. Source data and test statistics are provided as a Source Data file.

across 10 replicates (see Methods). We gauged the taxonomic imbalance of a dataset using the Gini index, as suggested in ref. 23. Index values range from 0 to 1, where 1 indicates a highly imbalanced dataset and vice versa. The WoL2 dataset has a Gini index of 0.59, while the generated datasets show an average Gini index of 0.66 ± 0.04, indicating a greater imbalance. For each dataset, we then selected an increasing number of markers from KEGG and EggNOG gene families and inferred trees from genomes containing at least one marker. We also included universal marker sets for comparison. As previously suggested, taxonomic imbalance decreases the quality of tree topology[43]. However, trees inferred from markers selected with our method show a higher quality than trees from previous sets. Specifically, as the number of KEGG markers increases, the average proportion of confident branches increases from 0.60 ± 0.05 to 0.78 ± 0.05 compared to 0.71 ± 0.08 from the best-performing marker set (two-sided two-sample Mann–Whitney U test, $p = 0.121$; Fig. 3d). The error relative to the WoL2 reference phylogeny decreases from 0.22 ± 0.02 to 0.18 ± 0.02, matching the error of PhyloPhlAn3 markers at 0.18 ± 0.02 (two-sided two-sample Mann–Whitney U test, $p = 0.650$; Fig. 3e), from which the WoL2 phylogeny was reconstructed. Trees inferred from EggNOG markers show the same trends. Overall, these results suggest that TMarSel markers mitigate the impact of taxonomic imbalance on tree quality, making it particularly suitable for metagenomic datasets where taxonomic imbalance is prevalent. For instance, the EMP dataset has a Gini index of 0.73, indicating a greater taxonomic imbalance than the WoL2 dataset.

## Robust marker selection yields accurate trees for MAGs despite incomplete genomic data

Most of the microbial diversity comes from MAGs derived from environmental samples[15]. Microbial communities from environments as diverse as seawater, soil, sediment, or animal gut sequenced in the EMP dataset have contributed to the genomic corpus of microbial diversity[44]. To test whether marker selection with TMarSel yields accurate phylogenies for MAGs, we used 793 high-quality MAGs from the EMP dataset annotated with KEGG and EggNOG databases. Because no reference phylogeny exists for the EMP dataset, we evaluated the quality of the inferred trees using only the proportion of

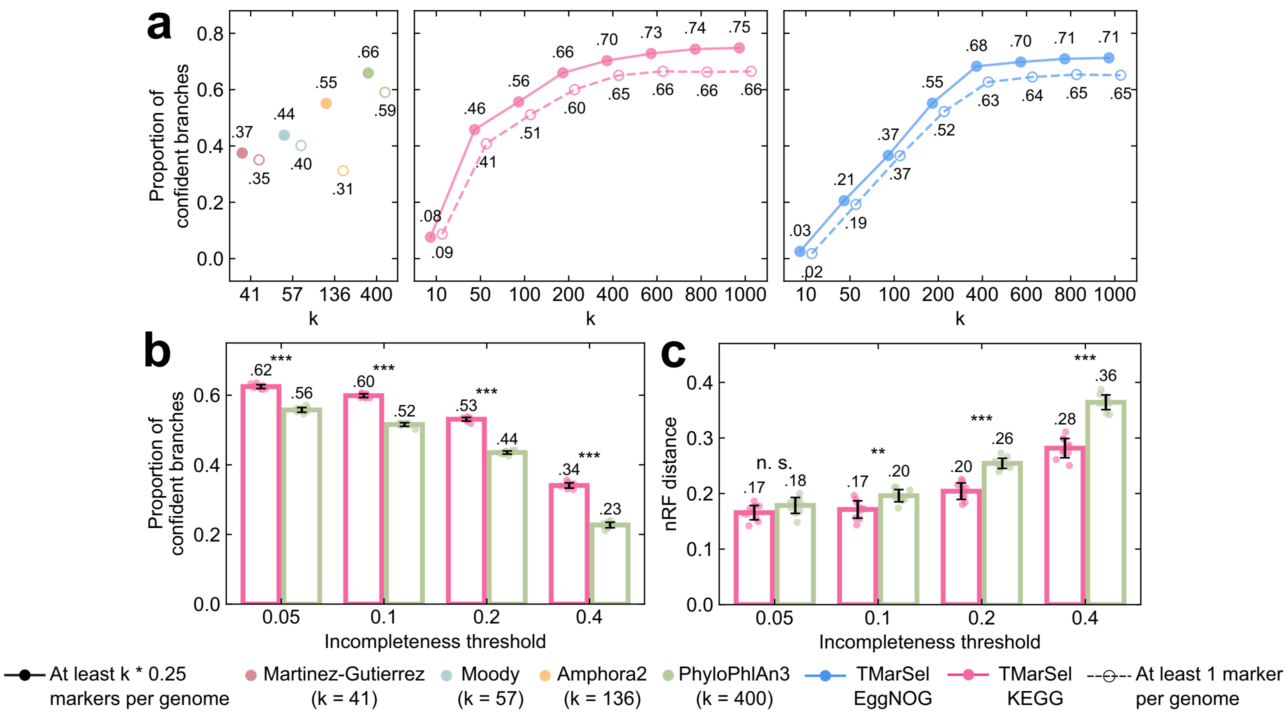


**Fig. 4 | Performance of TMarSel-selected markers on MAGs from the EMP dataset. a** Topological accuracy, measured as the proportion of confident branches, of trees inferred from 793 complete MAGs from the EMP dataset. Left panels indicate the tree quality of previous marker sets, whereas middle and right panels show the tree quality of KEGG and EggNOG markers, respectively. Symbols indicate trees inferred from different marker sets (color-coded), while line style and filled or unfilled symbols distinguish trees inferred from MAGs containing at least 25% of markers versus those containing only one marker. **b** Mean topological accuracy of trees inferred from simulations of increasingly 793 incomplete MAGs across $n = 10$ replicates. **c** nRF distance between trees inferred from incomplete MAGs to complete MAGs. The incompleteness threshold refers to the proportion of ORFs removed from each MAG. Unfilled bars indicate that MAGs contained at least one marker. Data are presented as the mean ± standard deviations. Asterisks show the significance level of two-sided two-sample Mann-Whitney U tests after Benjamini–Hochberg correction between the quality of trees inferred from KEGG versus PhyloPhlAn3 markers. Significance levels: n. s.: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$. Source data and test statistics are provided as a Source Data file.

confident branches, which measures the accuracy of topology. Similar to the WoL2 results, we found that using the ORFs with the highest bit score for each marker results in more accurate topologies (Supplementary Fig. 10). Despite the overall quality of trees being lower than those from the WoL dataset, the TMarSel markers selected from KEGG and EggNOG gene families produce trees with the highest accuracy compared to previous sets of universal genes (Fig. 4a). The increase in accuracy is more pronounced when more markers are used in the inference process ($k \geq 400$), though a plateau is reached at $k = 800$. Moreover, filtering MAGs with less than 25% of marker yields trees with ~5% more confident branches. Although the filtering step decreases the number of MAGs across all marker sets, trees inferred with TMarSel retain more MAGs than the next-best set of universal genes from PhyloPhlAn3 (Supplementary Fig. 11). The functional annotation and proportion of MGEs in TMarSel markers are congruent with those selected from the WoL2 dataset (Supplementary Fig. 12a, b).

MAGs represent draft microbial genomes, but most MAGs do not capture all the genomic content of a microbe[25]. To assess whether accurate trees can still be inferred, we simulated increasingly incomplete MAGs across 10 replicates. The incompleteness threshold is defined as the proportion of ORFs removed from a MAG (see Methods). We then selected markers ($k = 400$) from KEGG-derived gene families, as they produced the most accurate trees. To obtain a comprehensive picture of the impact of incompleteness, MAGs with at least one marker were included. We then measured the accuracy in the topology of trees inferred from incomplete MAGs and the mean nRF distance (error) between trees inferred from complete MAGs and incomplete MAGs. We also included trees inferred with the 400 universal genes from PhyloPhlAn3 for comparison and performed Mann-Whitney U tests with multiple testing corrections to assess whether differences in tree quality were significant. As expected, the accuracy of the trees decreases with the incompleteness threshold because fewer gene families are available for selection. Nonetheless, on average, trees inferred from TMarSel markers have 8% more confident branches than trees inferred from PhyloPhlAn3 universal genes across incompleteness thresholds (Fig. 4b). In addition, the robustness of our markers is evident in the low errors of inferred trees, which increase slower at higher thresholds compared to the trees inferred from PhyloPhlAn3 markers (Fig. 4c). Overall, these results suggest that a tailored marker selection from MAGs yields more accurate trees compared to previous sets of markers, even when MAGs lack a fraction of ORFs.

## Discussion

In this study, we assessed the phylogenetic signal of gene families for microbial phylogenomics using TMarSel, a robust yet flexible method for marker selection. We show that an expanded and tailored selection of markers can improve the accuracy of phylogenetic trees across simulated and real-world datasets of whole genomes as well as incomplete MAGs.

TMarSel provides a systematic exploration of the gene family space because it expands the source of markers to gene families instead of only orthologs. Among the currently available methods that model the evolutionary histories of gene families for tree inference, only ASTRAL-pro2 scales efficiently with a large number of genomes and genes[14], allowing us to assess the impact of different parameter combinations on the quality of the species trees. Although ASTRAL-pro2 only accounts for gene duplication and loss, its quartet-based approach is robust against HGT[45]. Moreover, the taxonomic consistency of clades, from the phylum to the genus levels, in trees inferred from TMarSel markers speaks for the robustness of our results despite the pervasiveness of HGT among microbes. Because we select gene families as markers, TMarSel is also compatible with other tree inference methods that actively account for gene duplication, loss, and transfer[46–48].

While previous marker sets were selected from a pre-defined collection of sequences and had a fixed number of markers, we offer more flexibility in both aspects. First, we rely on functional databases to obtain gene families from the input genomes, upon which TMarSel is applied. Our results show that markers selected from either KEGG or EggNOG yield trees with similar accuracy. This suggests a potential generalization to alternative databases, such as PFAM, UniRef, and MetaCyc. The choice of the database may depend on the focus of the study and downstream applications. For instance, KEGG links gene families to metabolic pathways, chemical reactions, enzymes, and other high-level functions[49–51]. Whereas EggNOG, in addition to functional annotations, provides evolutionary details of gene families[52,53]. Because TMarSel relies on functional annotations, ORFs not assigned to gene families, the so-called microbial dark matter[34], are not considered during the selection of markers. Therefore, novel potential marker genes are omitted. The prevalence of the microbial dark matter is more pronounced on MAGs, as only 49% and 87% of ORFs have been annotated with KEGG and EggNOG databases, respectively. However, an increasing number of databases are being developed to identify a greater fraction of the microbial dark matter[54–56], which TMarSel can leverage to provide a state-of-the-art set of markers. Moreover, since TMarSel requires a file mapping ORFs to gene families, users may also provide gene families inferred de novo. It is worth noting, however, that the quality of families can greatly impact the tree topology. Therefore, we suggest prioritizing the use of sound gene family inference methods.

Regarding the second aspect, rather than setting a fixed number of markers, TMarSel can select the best $k$ markers despite taxonomic imbalance (Fig. 3d, e). We show that increasing the number of markers yields more accurate trees. The increase in accuracy follows a saturation curve that suggests an optimal range for $k$, which can be identified by selecting an increasing number of markers until reaching a plateau. Since our selection method follows an iterative approach, smaller sets are distilled versions of larger marker sets. The iterative nature of TMarSel also suggests that the first selected markers contribute more phylogenetic signal, as shown in the saturation curves of tree quality (Figs. 3a, b, and 4a). Therefore, if computational resources are limited, especially for tree inference, a small set of markers still yields accurate trees. It has been suggested that at least 30 markers should be used for microbial phylogenomics[57]. However, the minimum number depends on the input data, as 50 markers yield trees with different proportions of confident branches when genomes or MAGs are used. As a result, the minimum number of markers increases when inferring phylogenies from MAGs.

MAGs comprise a large portion of the prokaryotic diversity, and their representation in genomic databases will only grow as more environmental samples are sequenced[15,58]. However, the integration of MAGs into phylogenomic pipelines comes with challenges. First, as we have shown, MAGs do not have gene families that conform to the traditional criteria for selection. Second, even high-quality MAGs, gauged as complete by CheckM, do not contain the entire genomic diversity of a microbial population[25,26], and some MAGs even lack ribosomal RNA and ribosomal protein genes due to assembly errors[27]. TMarSel effectively addresses the first challenge, as it can systematically explore the gene family space and select the best $k$ markers tailored to the input set of MAGs. The second challenge is more complex because incomplete MAGs inherently lack genomic data. And yet, we show that it is still possible to select markers with strong phylogenetic signals, yielding trees with high accuracy while allowing for partial recovery of the tree topology.

TMarSel effectively selects the optimal combination of markers from gene families that serve as the foundational genetic elements for inferring accurate phylogenetic trees from a tailored input of whole genomes and MAGs, despite taxonomic imbalance and incomplete genomic data. Since TMarSel primarily relies on the presence-absence patterns of gene families, it remains agnostic to the taxonomy of the input genomes or MAGs, as well as functional constraints, allowing for the selection of markers with diverse functional roles. We also show that marker selection can be agnostic to evolutionary rate, alignment quality, and non-vertical evolution. However, further evaluations of these properties are desirable to identify rogue markers that are detrimental to the inference process. In addition, we have only assessed the performance of our method for taxa spanning across the microbial tree of life. And yet, the robustness of our results indicates a potential generalization to fine-grained taxonomic groups, as well as other annotation databases and tree inference methods.

# Methods

## Marker gene selection

We represented gene families $\mathbf{U} = \{u_1, \ldots, u_m\}$ across genomes $\mathbf{V} = \{v_1, \ldots, v_n\}$ as a 2D matrix $\mathbf{A}_{m\times n}$ where entries are positive if gene family $u_i$ exists in genome $v_j$ and its values correspond to the number of times the gene is identified in the genome (i.e., copy number). Gene families not existing in a given genome were represented with a zero in $A$. To select a set of $k$ marker genes $\mathbf{G} = \{g_1, \ldots, g_k\}$, $k < m$, we devised an algorithm that, in each $k$ iteration, selects the gene $g$ that maximizes the objective function: $\arg\max\left(\left(\frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{c}_{k-1} + \mathbf{A}_{i,j}\right)^{p}\right)^{\frac{1}{p}}\right)$, where $\mathbf{c}_{1\times n}$ is a cumulative vector containing the copy number of genes already selected that were removed from $\mathbf{A}$ (see Appendix 1 in Supplementary Information). Note that our cost function is analogous to the generalized mean with exponent $p$. Small values of $p$ shift the cost function toward small values, selecting genes present in genomes with fewer gene families. In contrast, large values of $p$ will select genes from genomes with many genes. Because $\mathbf{A}$ contains zeroes, our cost function will return zero for $p \leq 0$ (Supplementary Fig. 4A). To avoid this issue, we added a pseudocount of 0.1 to $\mathbf{A}$ when selecting marker genes for all values of $p$.

To assess the wall clock time and memory usage of our selection method, we simulated multiple copy number matrices with a varying number of gene families ($m = 50$ to $100,000$) and genomes ($n = 707$ to $50,118$) and a maximum copy number of five following the protocol described in the next section. For each combination of $m$ and $n$, we selected an increasing number of markers from $k = 10$ to $m - 1$ if $n < 10{,}000$ or up to $k = 10,000$ otherwise. All steps were executed on a computer with an Xeon E5-2697 v2 processor, using the time command to record both statistics. The scaling exponent of each line depicting the computational requirements as a function of the number of markers was calculated using the curve fit function implemented in SciPy v1.11.3[59]. For this, we fitted a power law of the form $k^b$ to the data and took the average exponent across $m$, where $k$ is the number of markers and $b$ is the scaling exponent (Supplementary Fig. 3).

## Toy simulations

To assess the impact of parameters $p$ and $k$, we simulated multiple copy number matrices as follows: (i) Fix the number of gene families ($m = 50$) and genomes ($n = 10$). (ii) For each genome, randomly draw a number between zero and one from a uniform distribution to represent the fraction of gene families in the genome. For instance, a value close to zero corresponds to a genome with few gene families and vice versa. (iii) For each genome, generate a presence-absence vector of gene families, where a value of one (i.e., presence) is sampled with a probability equal to the fraction of gene families from step ii. In contrast, a value of zero (i.e., absence) is sampled with a probability of 1 - fraction of gene families. Thus, step iii produces a presence-absence matrix of gene families across genomes. (iv) To add copy numbers over the presence-absence matrix from step iii, we sampled a number between one and a maximum number of copies, set beforehand, for each gene present within a gene family. All steps were implemented using NumPy v1.26.0[60] while fixing a seed for reproducibility. This procedure was repeated for 25 replicates, yielding 25 different copy number matrices later used to select marker genes. For each replicate, we then performed a parameter sweep for 11 values of $k$ and $p$ ranging from $k=1$ to $k=m-1$ and from $p=-100$ to $p=100$ (Supplementary Fig. 4a).

To generate a species tree from the copy number matrix, we first calculated the Jaccard distance for each pair of genomes to obtain a distance matrix. We then applied neighbor-joining implemented in Scikit-bio v0.6.2 over the distance matrix to generate a species tree. To obtain gene trees, we took the species tree as a template, and for each gene family, we removed the genomes that did not contain the gene family. Hence, at this stage, gene trees contain only a single copy of each genome. We then added gene copies as sister branches of a given genome. Thus, for example, if a gene family is present in four out of ten genomes with copy numbers ranging from one to five (e.g., 1, 3, 5, 2), the resulting gene tree will have 11 leaves in total from four genomes, representing all the copies that stem from duplication or transfer events. Because we do not infer gene trees from sequences, we introduced noise into gene trees to simulate uncertainties that may arise during multiple sequence alignment and impact the gene tree inference process. Specifically, noise here refers to the proportion of branches that are randomly shuffled. Scikit-bio v0.6.2 was used to manipulate trees.

### Realistic simulations

Because the toy simulations are unrealistic, we simulated realistic prokaryotic gene families using Zombi[39]. We first generated a species tree under a birth-death model for 500 units of time with speciation and extinction rates per unit of time set to 0.04 and 0.03 per unit of time, as suggested by Louca et al.[61]. We then simulated genomes along the branches of the species tree using 10,755 and 5577 bacterial and archaeal gene family-wise values of duplication (D), transfer (T), and loss (L) benchmarked in ref. 38 while leaving the rest of the parameters as default. At the end of the simulation, Zombi outputs the real species tree and the genomes of each taxon along with real gene trees of each gene family. From the genomes, we built the copy number matrix that was later used to select marker genes. We repeated the simulations for 25 replicates while fixing a seed for reproducibility. For each replicate, we then performed a parameter sweep for 11 values of $k$ and $p$ ranging from $k=1$ to $k=m-1$ and from $p=-100$ to $p=100$. We also introduced noise to the gene trees to simulate uncertainties that may arise during multiple sequence alignment.

### Web of Life 2 and earth microbiome project datasets

The Web of Life (WoL2) contains 15,953 prokaryotic species with a single representative high-quality genome and taxonomic annotations from the Genome Taxonomy Database (GTDB) R207, as well as a reference phylogeny[62]. The WoL2 dataset is publicly available at https://ftp.microbio.me/pub/wol2/. Due to the burden of computational resources, we selected complete reference genomes from the WoL2, ensuring that each taxonomic group, from phylum to family, had at least 10 taxa. This resulted in 1510 genomes spanning the entire microbial diversity that will be used for genome annotation and marker selection.

The Earth Microbiome Project (EMP)[44] contains 811 high-quality Metagenome Assembled Genomes (MAGs) from 32 environments across the globe and are publicly available at https://www.globus.org/, endpoint emp_500_public. After genome annotation (see below), we inspected whether the number of Open Reading Frames (ORFs) matched the number of gene families. We discarded 18 poorly annotated MAGs with a low number of gene families compared to the number of ORFs. This resulted in 793 MAGs that will be used for all subsequent analyses.

### Genome annotation

To generate the copy number matrix for marker gene selection, we need to annotate the genomes and MAGs. First, we used Prodigal v2.6.3[63], in single-genome mode with the genetic code table specified according to taxonomy, to predict ORFs for genomes of the WoL2. ORFs from MAGs of the EMP dataset were predicted with Prokka v1.14.6[64]. We then used the KEGG Orthology release 102.0+[51] and evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG) v5.0[52] databases to annotate the ORFs into gene families with KOfam-Scan and EggNOG-mapper v2[65], respectively. In the KEGG annotation, we selected only prokaryotic KEGG Orthologs (KOs) with an e-value lower than the threshold defined in the database[66], thus minimizing false positive assignments. In the EggNOG annotation, we only selected the assignments at the broadest taxonomic level since we are interested in inferring a tree for diverse microbial species. The annotation of ORFs into gene families was then used to generate a copy number matrix $\mathbf{A}_{m\times n}$ with $m$ gene families and $n$ genomes.

We used the hierarchical classification from KEGG to map KOs to higher functional categories. For this, we counted the number of times a given gene is mapped onto a functional category. The counts in each category were then normalized by the total number of counts and multiplied by 100 to obtain a relative abundance. EggNOG only provides a functional description for each gene family. Therefore, we could not map its gene families to higher functional categories.

In addition, we identified Mobile Genetic Elements from KEGG and EggNOG gene families using the following description terms: baseplate, capsid, excisionase, DUF4102, pf00665, KilA-N, ORF11CD3, phage, portal, tail, terminase, tape, T5orf172, viral, virion, conjugal, conjugation, conjugative, DotD, IV secretory, IV secretion, MobA, mobilization, mobilization, MobL, DUF955, plasmid, relaxase, TcpE, TraG, TraL, TraM, DDE, pf01609, IS66, IstB, transposase, transposon, transposition, anti-restriction, antirestriction, integrase, integration, K02238, K02242, K02243, K02244, K02245, K02246, K12296, K04096, K06198, K07343, as suggested in ref. 67.

### Controlling copy numbers

Since ORFs mapped to gene families by KOfamScan or EggNOG contain summary statistics, we focused on the bit score value to control the number of copies of a gene family present within a genome. We used the bit score rather than the e-value, as the former is independent of the database size. For every genome, we identified the maximum bit score assigned to each gene family. We then excluded ORFs that had a bit score below a certain threshold that represents proximity to the maximum bit score. For example, a threshold of one will only include the best hits of each gene family per genome. In contrast, a zero threshold will include all the ORFs. Thus, the threshold represents the number of gene copies of each gene family per genome to include for marker selection.

### Impact of taxonomic imbalance on tree topology

We first calculated the Gini index to measure the taxonomic imbalance, as suggested in ref. 23, at the phylum level. Index values range from 0 to 1, where 0 indicates a perfectly balanced set and vice versa. We observed that the 1510 genomes from the WoL2 belong to 17 bacterial and archaeal phyla, and have a Gini index of 0.59.

To generate datasets with a Gini index greater than the WoL2 tested in the manuscript, we focused on the entire WoL2 dataset. We set the maximum number of genomes to 500 and the number of phyla to 17 for each dataset. We then sampled the genome abundance of each phylum from a $\beta$-distribution with parameters $a = 0.1$ and $b = 1$, ensuring highly right-skewed distributions where a few phyla contain the majority of genomes. The minimum genome abundance was set to four because the ASTRAL-Pro2 inference method is based on quartets. For each abundance, we then randomly sampled a phylum and checked whether the phylum had enough genomes to sample. If yes, then we randomly sampled genomes according to the abundance. Otherwise, we sampled another phylum until it had enough genomes. This procedure was repeated for ten replicates while fixing a seed for reproducibility. The generated datasets have an average Gini index of 0.66 ± 0.04, indicating a greater imbalance than the WoL2 dataset, but lower than the EMP dataset, which has a Gini index of 0.73.

For each dataset, we selected an increasing number of markers ($k = 50$, 100, 200, and 400) with exponent $p = 0$ from KEGG and EggNOG gene families with TMarSel. We also extracted previously proposed markers for comparison. Species trees were inferred with the pipeline described below. The quality of trees was then assessed as described below. We also performed two-sided two-sample Mann–Whitney U tests with multiple testing correction using the Benjamini–Hochberg method to test for significant differences in quality. Scikit-bio v0.6.2 was used to calculate the Gini index, while Scipy v1.11.3 was used for statistical analyses.

## Pipeline for species tree inference

Once markers have been selected, we retrieved all the ORF sequences associated with them. Depending on the experimental condition, we excluded those genomes with less than $k \times 0.25$ marker genes. Protein sequences of each marker gene were then aligned using UPP2[68] with default parameters. UPP2 is a multiple-sequence aligner designed for large datasets containing sequence length heterogeneity that might arise under large insertion or deletion events or due to incomplete assembly[69,70]. In the first stage, UPP2 selects a set of full-length sequences and computes a backbone alignment and an unrooted tree. A collection of profile Hidden Markov Models (pHMM) is then built for each subset of sequences in the backbone tree, and the remaining sequences are inserted into the backbone alignment. We chose the optimal number of backbone sequences using the backbone query split algorithm from UPP2. Data pipeline errors, such as sequencing, assembly, genome annotation, or alignment errors, substantially impact phylogenetic reconstruction as they increase the noise in the dataset. We used TAPER[71] with default parameters to remove such errors within alignments. TAPER is an outlier section algorithm that removes amino acids based on a divergence score computed along genomic positions and species[71].

We used the clean alignments for Maximum Likelihood tree reconstruction with FastTree v2.1[72] using the Lee-Gascuel (LG) model of amino acid substitution, as a previous study showed that the LG model best explains substitutions in a majority of prokaryotic marker genes[62]. This step produced a set of $k$ gene trees. We used TreeShrink[73] to remove pipeline errors that have escaped detection at the sequence level and resurfaced as suspiciously long branches in the phylogeny. TreeShrink removes leaves that inflate the tree diameter, defined as the maximum distance between any two leaves in the tree[73]. In our case, a leaf corresponds to a protein sequence of a gene within a genome.

We then used ASTRAL-Pro2[36] to estimate a species tree from the set of gene trees. ASTRAL-Pro2 combines information from gene trees to maximize a measure of quartet similarity between them and the species tree[35]. The quartet similarity measure is defined as the fraction of partitions between all combinations of four species, with a common ancestor originating from a speciation event that shares the same branching structure (i.e., topology) as the real species tree under gene duplication and loss scenarios[35]. Because only speciation events count toward the measure, ASTRAL-Pro2 identifies orthologs from gene trees containing all the homologs of a gene family[35,36]. This flexibility is ideal because the genome annotation step produces gene families that contain orthologs, paralogs, and xenologs.

## Sets of previous markers

To benchmark the usefulness of our marker genes for inferring phylogenetic trees, we compiled four datasets of marker genes previously proposed and used for deep microbial phylogenomics. (i) 41 single-copy universal marker genes tested for their phylogenetic signal in multidomain phylogenetic reconstruction[23]. (ii) 57 single-copy universal markers, which were used to estimate the age of the Last Universal Common Ancestor[1]. In the case of the 41 universal markers from Martinez-Gutierrez, the pHMMs of each gene were already provided, whereas, for the 57 universal markers from Moody, we built the pHMMs from the multiple sequence alignments with hmmbuild from HMMER v3.4[74]. We then used hmmsearch to search for homologs of the markers in the genomes of the WoL2 and MAGs from the EMP dataset. For each gene, only the ORF with the top bit score was extracted and used in the pipeline for species tree inference. Other marker sets we benchmarked are: (iii) 136 universal markers from AMPHORA2[30] available in PhyloPhlAn3[41], and (iv) 400 single-copy universal markers first proposed in refs. 24,75 and part of PhyloPhlAn3. To identify homologs of these two last sets of universal genes on the WoL2 and EMP datasets, we ran PhyloPhlAn3 v3.1.68 with high diversity and fast parameters. We then took the identified ORFs and used them in the pipeline for species tree inference. This approach ensures that species trees from all marker genes were inferred using the same pipeline, thus guaranteeing a fair comparison.

## Quality of species trees

To evaluate the quality of species trees inferred with our marker genes, (i) we calculated the Robinson-Foulds distance between the inferred tree and the WoL2 reference phylogenetic. The RF distance counts the different number of ways to divide a set of taxa by removing a branch[37]. We further normalized the RF distance by the sum of internal branches between the two trees to obtain a number between zero and one, where zero indicates two identical trees and vice versa. DendroPy v4.6.1[76] was used for all tree distance calculations. (ii) We also measured the quality of the inferred trees using the Local Posterior Probability (LPP) scores from ASTRAL-Pro2, which is a measure of confidence of each branch based on gene tree quartet frequencies. An LPP higher than 0.95 was used to classify a branch as highly confident, as suggested in ref. 40.

(iii) We gauged the taxonomic consistency of clades in a phylogenetic tree using the taxonomic accuracy metrics from ref. 75 and the standard microbial GTDB taxonomy R207. Taxonomic precision captures the notion that phylogenetically closer species must share a common taxonomic label. The precision of a clade is calculated as $\frac{L_C}{L_{LCA}}$, where $L_C$ is the sum of branch lengths of the members of the clade ($C$) and $L_{\mathrm{LCA}}$ is the sum of branch lengths of all the descendants of the Lowest Common Ancestor (LCA) of the clade. Thus, if all the members of the clade form a monophyletic subtree, the precision is one (highest possible). If any member of the clade is scattered, the denominator grows faster than the numerator, thereby reducing the precision. Taxonomic recall, on the other hand, captures whether taxonomically similar taxa are grouped close in the phylogeny. The recall of a clade is calculated as $\frac{|\{c \in C : d(c, \mathrm{LCTS}) \leq D_{\mathrm{LCTS}}\}|}{|C|}$, where $d$ is the distance between a member of the clade $c$ and the Longest Consistent Taxonomic Subtree (LCTS) and $D$ is the diameter (i.e., the branch length separating the most distant leaves) of the LCTS. The recall, therefore, calculates the number of taxa outside the LCTS. The LCTS is defined as the internal node with the largest number of children that are part of the clade and are monophyletic themselves.

(iv) In addition, we assessed the placement of each taxon from an inferred tree relative to the reference tree. We first calculated a distance matrix from the number of edges (i.e., internal branches) connecting every pair of taxa in the inferred and reference trees. Since each row in the distance matrix represents a vector containing the distance of a taxon to the other taxa in the tree, we calculated the cosine distance between the vectors of the same taxa in inferred and reference trees. Thus, if a taxon has the same placement in the inferred tree as in the reference tree, the distance vectors have the same direction, resulting in a cosine distance of zero. We expect well-placed taxa to have lower cosine distances to the reference tree. To assess the overall trend of placement, we applied a least squares regression implemented in SciPy v1.11.3.

### Simulations of incomplete MAGs

To simulate genome incompleteness, we randomly sampled contigs and sequentially selected ORFs in each contig until satisfying an incompleteness threshold, defined as the proportion of ORFs to be removed. The threshold is specified as a proportion of the total number of ORFs in the MAG. For instance, a threshold of 0.1 in a MAG containing 1000 ORFs will result in 100 ORFs removed. Once all the ORFs to be removed had been selected, we removed them from the genome annotation files that relate ORFs to gene families produced by KOfamScan and EggNOG mapper. We then used the remaining gene families to build the copy number matrix for marker gene selection. For each incompleteness threshold, we repeated the simulation for ten replicates while fixing a seed for reproducibility. In each replicate, a different contig and, subsequently, ORFs were sampled. Thus, our simulations yield MAGs with different genomic compositions.

For each replicate, we selected markers ($k = 400$) from the incomplete set of MAGs and inferred a species tree. We also inferred trees using the 400 universal markers from PhyloPhlAn3 for comparison. The trees were then evaluated for the proportion of confident branches. In addition, we calculated the nRF between trees inferred from incomplete MAGs and trees inferred from complete MAGs. To assess whether the differences in quality between our markers and PhyloPhlAn3 markers were statistically significant, we performed Mann-Whitney tests and corrected the $p$-values with the Benjamini–Hochberg method implemented in SciPy v1.11.3.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Source data are provided as a Source Data file. The datasets generated and analyzed, as well as the code used to produce the results and all the figures presented in this manuscript, are publicly available on Zenodo[78], under the BSD 3-Clause license. Exact $p$-values of statistical tests are provided in the Source Data file. Source data are provided with this paper.

## Code availability

The code used to produce the results and all the figures presented in this manuscript are publicly available on GitHub (https://github.com/HSecaira/AugmentingPhyloSignalMicrobes[78]), under the BSD 3-Clause license. The source code of TMarSel is hosted on GitHub (https://github.com/HSecaira/TMarSel/tree/main), together with documentation and test data sets.

## References

1. Moody, E. R. et al. The nature of the last universal common ancestor and its impact on the early earth system. *Nat. Ecol. Evol,* **8**, 1654–1666 (2024).
2. Martinez-Gutierrez, C. A., Uyeda, J. C. & Aylward, F. O. A timeline of bacterial and archaeal diversification in the ocean. *eLife* **12**, RP88268 (2023).
3. Dewar, A. E., Belcher, L. J. & West, S. A. A phylogenetic approach to comparative genomics. *Nat. Rev. Genet.* **26**, 395–405 (2025).
4. Lozupone, C. & Knight, R. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
5. Ladau, J. & Eloe-Fadrosh, E. A. Spatial, temporal, and phylogenetic scales of microbial ecology. *Trends Microbiol.* **27**, 662–669 (2019).
6. Baldauf, S. L. Phylogeny for the faint of heart: a tutorial. *Trends Genet.* **19**, 345–351 (2003).
7. Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128 (1999).
8. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).
9. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013).
10. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).
11. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
12. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).
13. Smith, M. L., Vanderpool, D. & Hahn, M. W. Using all gene families vastly expands data available for phylogenomic inference. *Mol. Biol. Evol.* **39**, msac112 (2022).
14. Weiner, S., Feng, Y., Gogarten, J. P. & Bansal, M. S. A systematic assessment of phylogenomic approaches for microbial species tree reconstruction. *Peer Community J.* **5**, e72 (2025).
15. Wu, D., Seshadri, R., Kyrpides, N. C. & Ivanova, N. N. A metagenomic perspective on the microbial prokaryotic genome census. *Sci. Adv.* **11**, eadq2166 (2025).
16. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
17. Yuan, C., Lei, J., Cole, J. & Sun, Y. Reconstructing 16s rNA genes in metagenomic data. *Bioinformatics* **31**, i35–i43 (2015).
18. Yang, C. et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
19. Ludwig, W. & Klenk, H.-P. Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. *Bergey's Manual® of Systematic Bacteriology* 2nd edn, Vol. 2, 49–66 (Springer-Verlag, Berlin, 2005).
20. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
21. Ramulu, H. G. et al. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet. Evol.* **75**, 103–117 (2014).
22. Patwardhan, A., Ray, S. & Roy, A. Molecular markers in phylogenetic studies-a review. *J. Phylogenet. Evolut. Biol.* **2**, 131 (2014).
23. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).
24. Segata, N. & Huttenhower, C. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLOS ONE* **6**, e24704 (2011).
25. Meziti, A. et al. The reliability of metagenome-assembled genomes (mags) in representing natural populations: insights from comparing mags against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* **87**, e02593–20 (2021).
26. Eisenhofer, R., Odriozola, I. & Alberdi, A. Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun.* **3**, 12 (2023).
27. Mise, K. & Iwasaki, W. Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. *ISME Commun.* **2**, 118 (2022).

28. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**, 5088–5090 (1977).
29. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579 (1990).
30. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with amphora2. *Bioinformatics* **28**, 1033–1034 (2012).
31. Tian, R. & Imanian, B. Vbcg: 20 validated bacterial core genes for phylogenomic analysis with high fidelity and resolution. *Microbiome* **11**, 247 (2023).
32. Na, S.-I. et al. Ubcg: up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* **56**, 280–285 (2018).
33. Wang, S., Ventolero, M., Hu, H. & Li, X. A revisit to universal single-copy genes in bacterial genomes. *Sci. Rep.* **12**, 14550 (2022).
34. Escudeiro, P., Henry, C. S. & Dias, R. P. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Curr. Res. Microb. Sci.* **3**, 100159 (2022).
35. Zhang, C., Scornavacca, C., Molloy, E. K. & Mirarab, S. Astral-pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* **37**, 3292–3307 (2020).
36. Zhang, C. & Mirarab, S. Astral-pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**, 4949–4950 (2022).
37. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
38. Coleman, G. A. et al. A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).
39. Davín, A. A., Tricou, T., Tannier, E., de Vienne, D. M. & Szöllősi, G. J. Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics* **36**, 1286–1288 (2020).
40. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
41. Asnicar, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using phylophlan 3.0. *Nat. Commun.* **11**, 2500 (2020).
42. Khedkar, S. et al. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.* **50**, 3155–3168 (2022).
43. Rosenberg, M. S. & Kumar, S. Taxon sampling, bioinformatics, and phylogenomics. *Syst. Biol.* **52**, 119 (2003).
44. Shaffer, J. P. et al. Standardized multi-omics of earth's microbiomes reveals microbial and metabolite diversity. *Nat. Microbiol.* **7**, 2128–2150 (2022).
45. Davidson, R., Vachaspati, P., Mirarab, S. & Warnow, T. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genom.* **16**, 1–12 (2015).
46. Morel, B. et al. Speciesrax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *Mol. Biol. Evol.* **39**, msab365 (2022).
47. Morel, B., Williams, T. A., Stamatakis, A. & Szöllősi, G. J. Alerax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics* **40**, btae162 (2024).
48. Weiner, S., Feng, Y., Gogarten, J. P. & Bansal, M. S. Assessing the potential of gene tree parsimony for microbial phylogenomics. In *RECOMB International Workshop on Comparative Genomics*, 129–149 (Springer, 2024).
49. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
50. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. Kegg: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
51. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
52. Huerta-Cepas, J. et al. 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).
53. Hernández-Plaza, A. et al. eggnog 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).
54. Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
55. Baltoumas, F. A. et al. Nmpfamsdb: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res.* **52**, D502–D512 (2024).
56. Rodríguez del Río, Á. et al. Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* **626**, 377–384 (2024).
57. Riesco, R. & Trujillo, M. E. Update on the proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evolut. Microbiol.* **74**, 006300 (2024).
58. Nayfach, S. et al. A genomic catalog of earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
59. Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272 (2020).
60. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
61. Louca, S. et al. Bacterial diversification through geological time. *Nat. Ecol. Evol.* **2**, 1458–1467 (2018).
62. Zhu, Q. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nat. Commun.* **10**, 5477 (2019).
63. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 1–11 (2010).
64. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
65. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggnog-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
66. Aramaki, T. et al. Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
67. Vatanen, T. et al. Mobile genetic elements from the maternal microbiome shape infant gut microbial assembly and metabolism. *Cell* **185**, 4921–4936 (2022).
68. Park, M., Ivanovic, S., Chu, G., Shen, C. & Warnow, T. Upp2: fast and accurate alignment of datasets with fragmentary sequences. *Bioinformatics* **39**, btad007 (2023).
69. Shen, C., Zaharias, P. & Warnow, T. Magus+ ehmms: improved multiple sequence alignment accuracy for fragmentary sequences. *Bioinformatics* **38**, 918–924 (2022).
70. Nguyen, N.-pD., Mirarab, S., Kumar, K. & Warnow, T. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.* **16**, 1–15 (2015).
71. Zhang, C., Zhao, Y., Braun, E. L. & Mirarab, S. Taper: pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods Ecol. Evol.* **12**, 2145–2158 (2021).
72. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010).

73. Mai, U. & Mirarab, S. Treeshrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genom.* **19**, 23–40 (2018).
74. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
75. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
76. Sukumaran, J. & Holder, M. T. Dendropy: a python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
77. Jennewein, D. M. et al. The Sol Supercomputer at Arizona State University. In *Practice and Experience in Advanced Research Computing*, PEARC '23, 296–301 (Association for Computing Machinery, 2023).
78. Secaira-Morocho, H., Jiang, X. & Zhu, Q. Augmenting microbial phylogenomic signal with tailored marker gene sets. https://doi.org/10.5281/zenodo.17069928 (2025).

## Acknowledgements

We thank Dr. Siavash Mirarab for valuable suggestions on the selection method and tree inference pipeline. We thank Dr. Justin Shaffer for organizing and sharing the EMP dataset. We are grateful to Matthew Aton for sharing computational resources to run analyses and to members of the Jiang lab for their insightful feedback on the manuscript. This work utilized the computational resources of the Sol Supercomputer at Arizona State University[77]. This research was partly supported by the Arizona Biomedical Research Centre (ABRC) under award no. RFGA2023-008-15 to Q.Z. Q.Z. was partly supported by the Division of Biological Infrastructure, National Science Foundation, under award no. 2412446. H.S.-M. and X.J. were supported by the Division of Intramural Research of the National Institutes of Health, National Library of Medicine.

## Author contributions

Conceptualization, H.S.-M. and Q.Z.; investigation, H.S.-M. and Q.Z.; formal analyses, H.S.-M. and Q.Z.; methodology, H.S.-M. and Q.Z.; software, H.S.-M. and Q.Z.; data curation, H.S.-M. and Q.Z.; writing - original draft, H.S.-M.; writing - review & editing, H.S.-M., Q.Z. and X.J.; visualization, H.S.-M., Q.Z. and X.J.; supervision, Q.Z. and X.J.; funding acquisition, Q.Z. and X.J.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64881-2.

**Correspondence** and requests for materials should be addressed to Qiyun Zhu.

**Peer review information** *Nature Communications* thanks Ana Pérez Cobas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.