Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea

Qiyun Zhu1,19, Uyen Mai2,19, Wayne Pfeiffer3, Stefan Janssen1,4, Francesco Asnicar5, Jon G. Sanders1, Pedro Belda-Ferre1, Gabriel A. Al-Ghalith6, Evgenia Kopylova1, Daniel McDonald1, Tomasz Kosciolek1,7, John B. Yin8,9, Shi Huang1,10, Nimaichand Salam11, Jian-Yu Jiao11, Zijun Wu1,12, Zhenjiang Z. Xu1, Kalen Cantrell6, Yimeng Yang6, Erfan Sayyari8, Maryam Rabiee2, James T. Morton1,2, Sheila Podell13, Dan Knights6, Wen-Jun Li11, Curtis Huttenhower14,15, Nicola Segata5, Larry Smarr2,16,17, Siavash Mirarab7 & Rob Knight1,2,16,18*

Rapid growth of genome data provides opportunities for updating microbial evolutionary relationships, but this is challenged by the discordant evolution of individual genes. Here we build a reference phylogeny of 10,575 evenly-sampled bacterial and archaeal genomes, based on a comprehensive set of 381 markers, using multiple strategies. Our trees indicate remarkably closer evolutionary proximity between Archaea and Bacteria than previous estimates that were limited to fewer “core” genes, such as the ribosomal proteins. The robustness of the results was tested with respect to several variables, including taxon and site sampling, amino acid substitution heterogeneity and saturation, non-vertical evolution, and the impact of exclusion of candidate phyla radiation (CPR) taxa. Our results provide an updated view of domain-level relationships.

https://doi.org/10.1038/s41467-019-13443-4
The metaphor of a “tree of life” was used by Darwin in his *On the Origin of Species* in 1859. It came into its modern form when Carl Woese and co-workers used the new ability to genetically sequence the small subunit (SSU) ribosomal RNA gene from multiple different organisms to create a phylogenetic tree, thereby showing a scenario of three domains of life: Bacteria, Archaea, and Eukaryota. Recent years have seen discoveries of novel microbial groups enabled by culture-based and metagenomic methods, many of which represent previously unknown biodiversity, and keep updating our knowledge of the extent and relationships among domains as indicated by phylogenetics. Among these new discoveries is the candidate phyla radiation (CPR, also referred to as Patescibacteria), a highly diversified clade of mainly uncultivated microorganisms that may subdivide the domain of Bacteria, although this scenario remains controversial. Meanwhile, the discovery and analysis of multiple novel archaeal lineages have suggested an archaeal origin for eukaryotes, pointing to a two-domain scenario. The currently representative view of the tree of life, inferred based on the concatenation of ribosomal proteins, illustrates a bipartite pattern with distinct separation between Bacteria (including CPR) and Archaea (plus Eukaryota). More comprehensive work in both taxon and locus sampling exists, but the inter-domain relationships were not explored.

Reconstructing phylogenies typically relies on comparing homologous features. Although closely related organisms often share obvious genome-level homologies, building higher-level, especially cross-domain phylogenies has been challenging due to the rarity of clearly defined homologies. To date, many efforts rely on one, or a few, universal “core” genes that are usually involved in fundamental translation machinery. Examples include the SSU rRNA and several dozen ribosomal proteins. The choice of these “marker genes” is based on their universality, conservation, and the observation that they suffer from less frequent horizontal gene transfer (HGT). However, HGT is widespread across the domains, affecting even the most conserved “housekeeping” genes, and cannot be ruled out for even these markers. Furthermore, the reliance on a few marker genes limits the information (i.e., phylogenetic signal) available for resolving all relationships in the tree of life. Finally, it reduces applicability in metagenomics—increasingly the main source of novel genome data—where assembled genomes are frequently incomplete and error-prone. Maximizing the included number of loci, thus, is desirable. However, when dealing with many loci, reconciling discordant evolutionary histories among different parts of the genome can become challenging. Moreover, a practical dilemma is imposed by computational limitations: adding breadth across the phylogenetic space requires more computing effort, which leads to compromises with either the quantity of genes analyzed or the robustness of tree-building algorithms.

In this work, we build a reference phylogeny of 10,575 bacterial and archaeal genomes (Fig. 1). They are sampled from all 86,200 nonredundant genomes available from NCBI GenBank and RefSeq as of March 7, 2017 (Fig. 2), using a statistical approach that maximizes the covered biodiversity. Our phylogenetic reconstruction uses 381 marker genes, selected from whole genomes solely by sufficient sequence conservation to identify homology. The whole data set totals 1.16 trillion non-gap amino acids, making it among the largest single data sets upon which de novo phylogenetic trees have been built (Supplementary Table 1). To infer species trees, we use both a summary approach that accounts for discrepancy among the evolutionary histories of individual genes, and the conventional gene alignment concatenation approach. The resulting species trees provide high resolution of the basal relationships among microbial clades, which show that Bacteria and Archaea are in closer proximity compared with previous estimations (Fig. 1). The phylogeny also enable us to evaluate and revise previously established taxonomic hierarchies. We have made our data and protocols publicly available at https://biocore.github.io/wol/.

**Results**

**Comprehensive sampling of biodiversity and genes.** By using a purpose-built “prototype selection” algorithm to maximize evenness of genome sampling (Supplementary Fig. 1, detailed in Supplementary Note 1) and by incorporating multiple additional criteria, including marker gene presence, genome quality, and taxonomy, we selected 10,575 genomes, covering 146 of 153 phyla defined by NCBI, plus all 89 classes, 196 of 199 orders, 422 of 429 families, 2081 of 2117 genera, and 9105 of 20,779 species (Fig. 2a). A total of 2852 genomes (27.0%) are metagenome-assembled genomes (MAGs), while the remaining are from isolates and other sources (Fig. 2c). Meanwhile, 2267 genomes (21.4%) are complete genomes or chromosomes, while the remaining are scaffolds or contigs (Fig. 2d). Overall, the selected genomes are of high completeness and low contamination as evaluated based on known lineage-specific marker gene sets (Fig. 2b). By testing against the MAG quality standard established by Bowers et al., only 10.4% MAGs or 3.7% of all genomes fall within the low-quality draft category, while the remaining meet the criteria of either high- or medium-quality drafts (Fig. 2e). This balanced representation of known bacterial and archaeal diversity ensured the comprehensiveness and evenness of the resulting phylogeny.

Our phylogenomic analysis was based on the 400 marker genes originally proposed in PhyloPhlAn (Supplementary Fig. 2). The taxon sampling protocol ensured that all selected genomes contain at least 100 marker genes each. In the resulting data matrix, each marker gene is present in 7565 ± 1730 (mean and std. dev.) genomes (Supplementary Fig. 2a), while each genome contains 286.14 ± 80.23 (mean and std. dev.) marker genes (Supplementary Fig. 2b). These marker genes were further filtered down to 381, based on metrics of alignment quality (see the Methods section) across the sampled genomes (Supplementary Fig. 2d).

Assessing deep phylogeny using multiple strategies. We explored multiple tree inference methods (detailed in Supplementary Note 2, with selected ones compared in Fig. 3 and Supplementary Fig. 3), but will mostly focus on two strategies: CONCAT and ASTRAL. CONCAT concatenates gene alignments and infers a single tree using maximum likelihood (ML) performed using the robust implementation in RAxML. Computational limitations forced us (Supplementary Table 2) to use at most 100 sites per gene, selected either randomly (“concat.rand”) or based on maximum conservation (“concat.cons”). However, we also tested analyzing all sites, using the faster but less accurate ML program, FastTree (referred to as “fasttree”). In contrast, the ASTRAL tree (“astral”) is based on first inferring 381 gene trees and then summarizing them using the ASTRAL method. ASTRAL accounts for gene tree discordance due to divergent coalescent histories and has been shown in simulations to be more accurate than concatenation in the presence of highly frequent HGTs. Due to its inherent scalability, ASTRAL analyses were able to use all the data (i.e., all sites of every gene). For comparison with previous studies, a CONCat tree was also built using 30 ribosomal proteins (“concat.rpls”). We used ML to estimate branch lengths for the ASTRAL tree based on the same data used to infer the CONCat tree.

Overall, ASTRAL (Fig. 1; Supplementary Fig. 4) and CONCat trees (Supplementary Figs. 5, 6) show congruence in topology
A new view of the bacterial and archaeal tree of life. The tree contains 10,575 evenly distributed bacterial and archaeal genomes, with topology reconstructed using ASTRAL based on individual trees of 381 globally sampled marker genes, and branch lengths estimated based on 100 most conserved sites per gene. Branches with effective number of genes (en) ≤ 5 and local posterior probability (lpp) ≤ 0.5 were collapsed into polytomies. Taxonomic labels at internal nodes and tips reflect the tax2tree curation result. Color codes were assigned to above-phylum groups and phyla with 100 or more descendants. For polyphyletic taxonomic groups, minor clades with <5% descendants of that of the most specious clade were omitted, while the remaining clades were appended a numeric suffix to reflect the tax2tree curation result. Color codes were assigned to above-phylum groups and phyla with 100 or more descendants.

To further evaluate the impact of taxon sampling, we tested a series of subsampled sets of taxa, selected so that they maximize the representation of large and deep-branching clades (Fig. 8). The ASTRAL tree, in particular, has high support among the early branching clades (Fig. 3a, b; Supplementary Note 2) when compared with trees derived from implicit (e.g., distance-based) analyses (Supplementary Fig. 7, Supplementary Note 2). The congruence is higher at the root (Figs. 1, 3, Supplementary Figs. 4–6) and the robustness of method (Supplementary Fig. 12, Supplementary Note 2).

To further evaluate the impact of taxon sampling, we tested a series of subsampled sets of taxa, selected so that they maximize the representation of large and deep-branching clades (Fig. 8). The ASTRAL tree, in particular, has high support among the early branching clades (Supplementary Figs. 3, 9, also see Supplementary Figs. 4–6). This high resolution is directly related to the large number of gene trees used in the inference, as using fewer loci notably decreased the branch support of the species tree (Supplementary Fig. 10, Supplementary Note 2). On the other hand, the evolutionary relationships recovered by CONCAT are impacted by the breadth of site sampling (Supplementary Fig. 11, Supplementary Note 2) and the robustness of method (Supplementary Fig. 12, Supplementary Note 2).

To further evaluate the impact of taxon sampling, we tested a series of subsampled sets of taxa, selected so that they maximize the representation of large and deep-branching clades (see the Methods section). Reducing taxon sampling changed the overall topology (Supplementary Fig. 13, Supplementary Note 2) and the inferred relationship between large groups (e.g., placement of Chloroflexi and Chlamydiae) (Supplementary Fig. 14), further highlighting the importance of our dense sampling of genomes.

Phylogenetic trees built by both strategies recapitulated clear separation between Archaea (669 taxa) and Bacteria (9906 taxa) at the root (Figs. 1, 3, Supplementary Figs. 4–6). Meanwhile, CPR...
(1454 taxa) forms a monophyletic group located at the base of the bacterial lineage in the ASTRAL tree and the CONCat trees that use global markers (Figs. 3c, 4a). Considering the potential impact of long-branch attraction, this placement will require further validation using more robust substitution models and controlled tests. The ASTRAL tree shows high consistency and congruence, though with a few exceptions at deep branches (Fig. 3a–d; Supplementary Figs. 15b, d, 16, elaborated in Supplementary Note 4). A detailed interpretation of our phylogeny in reference to taxonomy and multiple previous works is provided in Supplementary Note 5.

**Evolutionary proximity between Archaea and Bacteria.** ASTRAL and CONCat trees both reveal a relatively short branch connecting the most recent common ancestors of Archaea and Bacteria (Figs. 1, 4a, c; Supplementary Fig. 17). Its length is fractional comparing with the dimensions of both clades (appr. 0.13–0.14 by conserved sites, 0.09–0.11 by random sites) (Fig. 4c, e; Supplementary Table 4). This pattern is in contrast to previous trees built using fewer marker genes, all or most of which are ribosomal proteins formerly considered to be effective markers for assessing global microbial evolution (e.g.,13,19,30). To further test how the choice of marker genes affects the inter-domain distance, we estimated branch lengths of the ASTRAL tree using 30 ribosomal proteins extracted from curated NCBI taxonomy, showing frequent incongruences (Supplementary Fig. 15a, c, Supplementary Table 3), especially in metagenome-derived genomes (detailed in Supplementary Note 3). We further compared our trees with the recently developed GTDB taxonomy and trees,4 observed overall high congruence, though with a few exceptions at deep branches (Fig. 3a–d; Supplementary Figs. 15b, d, 16, elaborated in Supplementary Note 4). A detailed interpretation of our phylogeny in reference to taxonomy and multiple previous works is provided in Supplementary Note 5.
estimate using the 381 global marker genes (Fig. 4b, d, e; Supplementary Table 4). We also calculated the overall phylogenetic distance between taxa of the two domains, as relative to the intra-domain distances. This relative distance based on the ribosomal proteins (4.5–5.0) is around three times that of the distance by the global marker genes (1.5–1.6) (Fig. 4f; Supplementary Table 4).

Considering the special status of CPR, we performed an independent test with the 1454 CPR genomes removed from the data set prior to de novo phylogenetic inference, and we compared the results to the main results (Fig. 4c, f) with the CPR clade pruned from the tree. These trees continued to reveal the substantially shorter branch and tip-to-tip distances between the two domains as recovered by using the 381 global marker genes as compared with using the 30 ribosomal proteins (Supplementary Fig. 18, Supplementary Table 5).

We tested whether the potential saturation of amino acid substitution could cause an underestimation of the domain separation. The ratio between phylogenetic distance and sequence distance is similar between pairs of taxa selected both from Bacteria, both from Archaea, or one from each domain (Supplementary Fig. 19). This indicates that the relative length of the branch connecting the two domains compared with the intra-domain branches is not substantially impacted by saturation.

We further evaluated how individual gene trees impact the observed proximity between Bacteria and Archaea. Except for a few outliers, which include several “core” genes like rpoC (RNA polymerase subunit β1, 18.27), tsf (elongation factor Tu, 12.18), and fisA1 (elongation factor G, 9.54), most gene trees have relative Archaea–Bacteria distances between 1 and 3 (mean: 2.00) (Fig. 5a, b), which is consistent with that of the species tree
summarizing the global marker genes, and in contrast to that obtained using only the ribosomal proteins (Fig. 5a).

### Heterogeneity among individual genes’ evolutionary histories.

Because microbial genomes are highly dynamic and prone to HGTs, it is important to investigate the discrepancies among the evolutionary paths of individual gene families to better understand the evolution of genomes. To measure the topological concordance between two trees, we used the quartet score, which correlates well with the traditional Robinson–Foulds (RF) metric, resulting in a distribution of gene trees tightly centered around the species tree (Fig. 5d; Supplementary Figs. 20a, 21).

The discordance between the 381 single-gene trees and the species tree varied widely (Fig. 5c). The quartet scores (larger is more similar, with identical trees scoring 1.0) ranged from 0.372 (cmpD) to 0.943 (hslU), with the mean and standard deviation being 0.653 ± 0.136. Many of the individual trees with high similarity to the species tree belong to genes involved in the core machinery of genetic information processing, such as those encoding DNA/RNA polymerase subunits, ribosomal proteins, and elongation factors, while genes involved in peripheral functions such as membrane transport are frequently more distant from the species tree (Fig. 5c, d). This pattern is generally consistent with a previous study on a small taxon set. While determining the cause of discordance for individual genes is beyond the scope of this study, the pattern we observed is consistent with a reduced rate of HGT for fundamental genes compared with those with less conserved functional significance.

To further test the impact of gene tree discordance on the species tree, we sequentially removed genes from the low end of...
the quartet score rank (Supplementary Note 2). ASTRAL produced stable topologies in this test (Supplementary Fig. 2A–C). We next tested the impact on phylogenetic distances. There was a weak positive correlation (Pearson correlation $R^2 = 0.157,$ $p = 1.88e-07$) between the quartet score and the relative Archaea–Bacteria distance (Fig. 5c). When the branch lengths of the species tree were estimated using genes with high quartet scores only, the distance moderately increased, yet remained far from the result by using the ribosomal proteins (Supplementary Table 6). This suggests that non-vertical transmission of genetic information has only a limited impact on our updated estimates of the inter-domain distance.

**Heterogeneity across sites.** Inferring phylogenetic trees at deep time scales, beyond the heterogeneity of gene histories, requires paying attention to the heterogeneity of substitution processes across the genome\(^{34,35}\). As recently as 2015, Gouy et al. declared the jury to still be out on the root of the tree of life\(^{36}\), partially due to difficulties in modeling heterogeneity of sequence evolution across sites. In particular, changes in amino acid frequency across sites of the same gene can exacerbate long-branch attraction\(^{37}\). To account for these difficulties, we tested whether our main conclusions stand if the data are analyzed with a recently developed model, PMSF, which considers heterogeneity in the amino acid substitution process\(^{38}\) (Fig. 3: “pmsf.cons” and “pmsf.rand”). Because of the computational complexity of this approach, we had to limit these analyses to 1000 taxa. At this sampling depth, we were also able to build a tree using all sites and the CONCAT method for comparison (“concat.allk”).

The topology of PMSF trees largely resembled the RAxML trees with the same taxon sampling (Fig. 3A, B). The impact of using the PMSF model instead of site homogeneous models on the topology and branch lengths was small compared with the impact of taxon, locus, and site sampling (Supplementary Fig. 23, Supplementary Note 2). The PMSF trees continued to support a large portion of relationships among deep branches recovered by the full-scale trees (Fig. 3C, D; Supplementary Fig. 3). The evolutionary proximity between Bacteria and Archaea continued to hold with the PMSF trees. Meanwhile, the PMSF tree based on ribosomal proteins (“pmsf.rpls”) also resembled the corresponding full-scale tree in
suggesting a long distance between Bacteria and Archaea (Fig. 4e, f; Supplementary Fig. 24, Supplementary Table 7). Taken together, this shows that our phylogenies and main conclusions are robust when considering site heterogeneity.

**Discussion**

The origin and evolution of life have been among the most intriguing scientific questions, with the current widely adopted notion being the three-domain system: Bacteria, Archaea, and Eukaryota\(^2\). Recent phylogenomics studies typically indicated a long distance between Bacteria and Archaea, with Eukaryota as an ingroup of the Archaea clade\(^10,11\). In this work, we built a reference phylogeny of over 10,000 bacterial and archaeal genomes, covering a significant proportion of the known biodiversity with available genome data. The result provides an updated view of microbial evolution, showing that Bacteria and Archaea, the two microbial domains conventionally but controversially grouped by the term “prokaryotes”\(^39\), are much closer in evolutionary proximity than estimates using a smaller number of “core” genes, such as the ribosomal proteins. This observation was further supported by extensive analyses using multiple tree-building methods, with consideration of taxon and site sampling, amino acid substitution heterogeneity and saturation, and non-vertical evolution, and was robust against the exclusion of CPR taxa. Interestingly, applying a simple universal molecular clock as well as relaxed clock rates to date, our trees resulted in divergence time estimates of major lineages that are compatible with geological timeline only when using the global markers, but not when trees are restricted to ribosomal proteins (Fig. 6; Supplementary Figs. 25, 26, Supplementary Tables 8-10, see full details in Supplementary Note 6). These comparisons suggest accelerated evolution in ribosomal proteins during the separation between Bacteria and Archaea. They show the limitation of using core genes alone to model the evolution of the entire genome, and highlight the value in using a more diverse marker gene set.

Our work highlights the value of even taxon sampling, a global marker gene set representing the larger average of genome content, and comparative phylogenetic analyses. These procedures largely reduced the bias of gene choice in exploring genome evolution, and allowed us to characterize the evolutionary discrepancies of individual gene families. Despite these efforts, some lineages are still underrepresented in our sampling, such as DPANN\(^4\), which has genomes that are often missing many of the 381 marker genes (detailed in Supplementary Note 5). Moreover, the rapid growth of genomic data has led to the absence of some newly discovered groups from our tree. While it is impractical to repeat all of our analyses to include all new genomes, it is of interest to assess whether the newly discovered microbial diversity may impact our results. Prior to submission of this article, we updated the genome collection from NCBI on May 23, 2019, and selected 187 new genomes representing phyla as defined by the newest NCBI and GTDB taxonomies that are absent or underrepresented in the current set of 10,575 genomes (see the Methods section). Phylogenetic trees built using the extended genome set continued to support the domain-level relationships in both topology and evolutionary distance as recovered by the main analysis (Supplementary Fig. 27, Supplementary Table 11, see Supplementary Note 7 for full details). Finally, we note that the inclusion of eukaryotes is challenging with the current marker gene set due to the overall sparsity of detectable homology. Further improvements in methodology are important in order to deliver a robust phylogeny that encompasses all forms of life.

**Methods**

**High-performance computing.** Analyses of the genome data sets in this study were computationally intensive. Heavy computations used the “Comet” supercomputer located at the San Diego Supercomputer Center (SDSC). Each standard
node is equipped with 24 Intel Haswell CPU cores and 128 GB of DDR3 memory, while each GPU node is equipped with four NVIDIA P100 GPUs, plus 28 Broadwell CPU cores at 128 GB memory. A small proportion of the computations used the “Barnacle” computer system operated by the Knight Lab, each node of which has 32 Haswell CPU cores and 256 GB DDR3 memory. Whenever possible, all CPU cores were used in a typical multi-processing task to minimize run time. Tasks that required >128 or 256 GB memory used the large-memory nodes of Comet. All used 64 CPU cores and 1.5 TB memory per node. Benchmarks of the prototype selection algorithms and some local developments were performed on the “WarpsDrw” workstation, equipped with 32 Intel Sandy Bridge CPU cores and 256 GB of DDR3 memory.

Retrieval of genome data and metadata

Microbial genomes were downloaded from the NCBI genome database (GenBank and RefSeq) as of March 7, 2017. We used and provided updates related to this work to the automated workflow RepoPhlAn (https://bitbucket.org/nosegata/reophlan; commit 03614ac) to download genome from that GenBank and RefSeq database. Each small proportion of the computations used the NCBI accession of the corresponding assembly, but without version number. For example, a genome with assembly accession “GCF_000125456.1” was identified as “G000125456.1” in this study. In cases when the same genome was present in both GenBank (accession starting with “GCA_”) and RefSeq (accession starting with “GCF.”), only the RefSeq version was kept.

Annotation and classification of marker genes

The functional annotation of the 400 PhyloPhlAn marker genes33 was performed by aligning the protein sequences of the 400 marker genes (inferred from 2887 bacterial and archaeal genomes as described in Segata et al.25) against the UniRef50 database (March 2018 release) using BLASTp. The best hit for each gene was taken and queried against the UniProt database for gene and protein names. To categorize genes by function, the UniProt entries were translated into Gene Ontology (GO) terms40 with the “subset_prokaryote” tag (March 2018 release). Because not all UniProt entries have corresponding GO terms, manual curation was involved to pick the most appropriate GO terms for those cases by examining the BLAST hit table. GO terms were further translated into GO slim terms to obtain higher-level functional categories. Note that this analysis is independent from the phylogenetic analysis of the current genome data set, and the result can be used as a reference for PhyloPhlAn users.

Analyses of genome sequences and identification of marker genes

The DNA sequences of the 86,200 bacterial and archaeal genomes were subjected to the following analyses:

1. The quality scores for DNA, protein, tRNA, and rRNA were calculated following Land et al.31.
2. A MinHash sketch was built for each genome using Mash 1.1.42, with default settings (sketch size = 1000, k-mer size = 21), based on which a pairwise distance matrix was built for the entire genome pool. In brief, MinHash is a k-mer hashing technique that enables the quantification of genome-to-genome distance. It is efficient for very large genome sets, and it correlates well with the conventionally used average nucleotide identity (ANI)42.
3. Although NCBI provides genome annotations, we chose to re-annotate the genomes using a uniform protocol to ensure consistency. Specifically, open-reading frames (ORFs) were predicted using Prodigal 2.6.355 in the single-gene mode, and aligning ORFs to putative ORF regions identified using modules. It finally uses transitivity to perform the subsequent merging.
4. Based on the predicted ORFs, the 400 marker genes were inferred and extracted using the phylogenomics pipeline PhyloPhlAn (commit 2c06e1a3) in which the 400 marker genes were originally introduced. In brief, we used USEARCH v9.1.13 to align ORFs against the reference marker gene sequences (see above) at an E-value threshold of 1e-40. It then selected the highest bit score hit of each ORF. Should more than one hit per marker genome be observed, the highest bit score hit was selected as the representative of that marker gene.
5. The completeness, contamination, and strain heterogeneity scores were computed using CheckM 1.0.724 with the default protocol (“lineage_wf”).

Prototype selection and genome sampling

Protein taxonomic is a key prerequisite to inferring an unbiased view of organism evolution46,47. Beyond computational challenges in robust tree-building, the highly uneven distribution of known biodiversity (e.g., 40.0% of all genomes (34,507) belong to the nine most-sequenced species) requires deliberate subsampling to reduce the bias from the resulting phylogeny in representing a global view of evolution. We therefore applied the data-reduction strategy of “prototype selection”47, which subsamples genomes from the pool such that they represent the largest possible biodiversity—in terms of maximized sum of pairwise distances as defined by k-mer signatures (Supplementary Fig. 1a). We developed a heuristic (detailed in Supplementary Note 1), capable of handling the size of the current genome pool, with results comparable with or better than published alternatives (Supplementary Fig. 1b–c).

Using this algorithm and by applying multiple criteria, we downsampled the 86,200 bacterial and archaeal genomes to 11,079. The procedures are detailed below.

1. Excluded genomes with marker gene count < 100 or contamination > 10%. The marker gene count threshold 100 was chosen because it is sufficiently large to yield high resolution of the tree using ASTRAL (Supplementary Fig. 10a, c). The contamination threshold 10% is inline with the medium- and low-quality draft genome standards proposed by Bowers et al.5. Nevertheless, we did not adopt the completeness and tRNA/rRNA coverage thresholds5, because the 400 protein-coding marker genes are more relevant for phylogenetic reconstruction.
2. Included the NCBI-defined reference and representative genomes (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/).
3. Included genomes that are the only representative of each taxonomic group from phylum to genus.
4. Included genomes that are the only representative of each species without defined lineage (no classification other than species).
5. Executed the prototype selection algorithm developed in this work: “distractive_maxdist” (see Supplementary Note 1) based on the distance matrix defined by MinHash signatures, with the already included genomes as seeds, to obtain a total of 11,000 genomes.
6. For each phylum to genus, and species without classification from phylum to genus, selected one with the highest marker gene count. This added 79 genomes to the selection.

These 11,079 genomes were subjected to our phylogenetics protocol, during which further filtering was performed based on sequence alignment quality (see below). Eventually, 10,575 genomes were retained.

Impact of alternative genetic codes

We chose to uniformly apply the standard archaean and bacterial genetic code table 11 to all genomes in order to minimize bias. Reports have shown that several lineages, such as Mycoplasma/Spiroplasma48,49, Helobdella50, and Absconditabacteria51, use alternative genetic code tables 4, 25 and others, in most of which a stop codon is repurposed to encode for an amino acid, resulting in ORF elongation. We did not incorporate alternative genetic codes, however, because there is no accurate way to associate each of the 86,200 genomes with its true genetic code. Incorrect truncation of ORFs may unnecessarily exclude genes and taxa, whereas incorrect elongation of ORFs could result in artificially long branches, because the amino acid sequence after a true stop codon is likely relaxed from selective pressure. Considering our goal of inferring phylogenetic topoloy and distances, we decided to only use the standard genetic code.

However, we did test the impact of using alternative genetic code on the gene and taxon sampling. We ran Prodigal 3.0.0-re1, which automatically switches from genetic code 11 to 4 if the average ORF length is too short. This resulted in altered gene calling results in 453 out of the 86,200 genomes, of which 63 had overly short ORF lengths even when using genetic code 4. PhyloPhlAn marker gene discovery on the other 390 genomes with genetic code 4 suggested marginal increase in the extracted number of the 400 marker genes per genome (1.23 ± 5.28, mean and std. dev.). Only seven additional genomes which had <100 marker genes managed to pass this threshold (see above) after switching to genetic code 4. Therefore, omitting alternative genetic code has little impact on the inclusion of genomes.

Metric multidimensional scaling (mMDS) of genome distances

The effect of prototype selection was visualized using the mMDS technique, which renders a low-dimensional plot that minimizes the loss of information when transforming from the high-dimensional data. We performed mMDS using the “nmds” function implemented in scikit-learn 0.19.231 on the genome distance matrix, using the default setting, to compute the coordinates at the top five axes. The resulting coordinates were visualized with the interactive tool Emperor52 as bundled in QIIME 2 release 2017.1272.

Protein sequence alignment and filtering

Protein sequences of each of the 400 marker gene families were aligned using UPP v2.054, a phylogeny-based and fragmentary-aware alignment tool. UPP consists of several sequentially connected modules. It first identifies suspected fragmentary sequences, then calls FASTA v1.8.65 to align the remaining sequences and build a phylogeny (backbone tree) based on them. Then it builds an ensemble of HMMs using HMMER56 based on the phylogeny. Finally, it aligns the fragmentary sequences to the HMMs and selects the one with the best match. Sequences that are 25% longer or shorter than the median sequences were considered as fragments and excluded from the backbone. More specifically, if we first builds a starting tree, performs a tree-based clustering of the sequences, and builds a spanning tree from these clusters. Then it calls MAFFT v7.149b57 to align the sequences in each cluster, and calls OPAL58 to merge the alignments of adjacent clusters according to the spanning tree, and finally uses transitivity to perform the subsequent merging.

In the dataset, we filtered out extremely gappy sites and sequences: sites with >90% gaps were deleted from the alignments, followed by the dropping of sequences with >66% gaps.
Filtering of marker genes. To ensure the quality of the species tree built upon these marker genes25, we filtered out the genes that were not aligned reliably by UPP. As such, the marker genes with ≥75% gaps in the aforementioned alignments were excluded from the pool, leaving 381 marker genes in total. The threshold 75% was chosen based on the distribution pattern of per-gene alignment quality (Supplementary Fig. 2d).

Filtering of outlier taxa from gene trees. We removed suspected outliers by detecting the taxa on disproportional long branches and filtering them out from the phylogeny inferred by FastTree27. To do this, we applied TreeShrink60 v1.1.0, a method that simultaneously detects long branches on a set of gene trees by identifying a set of taxa that could be removed from each gene so that the gene trees are maximally reduced in diameter. We used FastTree 2.1.9 to infer preliminary gene trees of the 381 selected genes, then ran TreeShrink to detect outlier long branches in these trees, with the per-taxa test with α = 0.05 (5% false-positive tolerance). Finally, we dropped genomes that contained <100 marker genes post gene tree filtering.

Gene tree reconstruction. Gene tree topologies were reconstructed using the maximum likelihood (ML) method as implemented in the state-of-the-art phylogenetic inference program RAxML 8.2.1026. The best amino acid substitution model for each of the 381 universal marker genes was inferred using RAxML’s built-in script ProteinModelSelection.pl. Three phylogeny trees were reconstructed for each gene family: one using a starting tree computed by the fast ML approach implemented in FastTree and the other two using parsimony trees built with random seeds 12345 and 23456. RAxML was executed with the ML search convergence criterion (-D) and the CAT rate heterogeneity model without final optimization (-F) to reduce the execution time.

For each of the 1143 topologies (3 × 381), another RAxML run was executed to optimize branch lengths and to compute likelihood scores under the robust, but expensive Gamma rate heterogeneity model. Because of numerical instability, at least one of the RAxML runs failed for 39 of the 381 gene families. For those cases, IQ-TREE 1.6.191, an alternative and faster maximum likelihood program, was used instead to optimize branch lengths using the same model (G4). The tree with the highest likelihood score among the three runs was retained for downstream applications. In 161 gene families, this tree was from the run with the FastTree starting tree, while in the remaining gene families the best tree was from either one of the random seeds.

Species tree reconstruction by summarizing gene trees (ASTRAL). A species tree was reconstructed by summarizing the 381 gene trees, using ASTRAL-MA52 (implementing ASTRAL-III algorithm)25 5.12.6a. This analysis was run on the Corcent supercomputing cluster using 24 cores and 4 GPU acceleration. In the resulting tree, the branch lengths represent the units of coalescence. Each branch has three support values: (1) effective number of genes (EN): the number of gene trees that contain some quartets around that branch; (2) quartet score (QT): proportion of the quartets in the gene trees that support this branch; (3) local posterior probability: the probability that the true tree has given, at the set of gene trees (computed based on the quartet score and using incomplete lineage sorting (ILS))31.

Branch length estimation for the ASTRAL tree. The branch lengths of a summary tree generated by ASTRAL are in coalescent units and only for internal branches. In order to obtain conventional branch lengths, i.e., the expected number of amino acid substitutions per site, we ran IQ-TREE using the concatenated alignment (most conserved or randomly selected sites as described below) as input, the ASTRAL tree as the topological constraint, and the LG + Gamma model. Branch lengths obtained using both site categories were highly correlated (Supplementary Note 2).

Species tree reconstruction based on the concatenated alignment (CONCAT). The alignments of the 381 marker genes were concatenated into a supermatrix. Due to the computational challenge in running classical maximum likelihood tree reconstruction on the full-scale data set, we had to downsample it to around 38 k amino acid sites. In order to explore the impact of site sampling on tree topology and branch lengths, we separately adopted two strategies for site sampling: (1) selected up to 100 most conserved sites per gene, the degree of conservation was estimated using the “trident” metric63, which is a weighted composition of three functions: symbol diversity, stereochemical diversity, and gap distribution. The PFASUM60 substitution matrix was used for computing the stereochemical diversity64. (2) randomly selected 100 sites per gene from sites with <50% gaps. For the downsampled supermatrix, a maximum likelihood tree was first built using FastTree, with LG model for amino acid substitution and Gamma model for rate heterogeneity. Using this FastTree tree as the starting tree, plus two maximum parsimony trees generated from random seeds (12345 and 23456), we performed three independent runs using RAxML, with the LG + CAT models (PROTGAMMLG + CAT), with rapid hillclimbing (-f D) and without final Gamma optimization (-F). With the resulting topologies, we performed branch length optimization and likelihood score calculation using IQ-TREE, with the LG + Gamma models (LG + G4). We further performed 100 rapid bootstraps using RAxML to provide branch support values.

Species tree reconstruction based on ribosomal proteins. To test the impact of choice of marker gene set on the topology and relative distances among major taxonomic groups, we conducted a separate analysis in which the species tree was built using ribosomal protein sequences. We identified and extracted 30 ribosomal protein families using the program PhyloSift 1.0159 with its marker database released on August 8, 2017. If more than one copy of a marker protein was detected in a genome, all copies were discarded. After this filtering, genomes with fewer than 25 marker proteins were dropped from the data set, resulting in a total of 9814 genomes out of the original 10,575. Sequences of each ribosomal protein family were aligned using UPP as described above. The resulting alignments were concatenated and subjected to RAxML tree reconstruction using the LG model for amino acid substitution66 (which is the best model for 304 out of the 381 genes based on RAxML’s model selection) and the CAT model for rate heterogeneity (PROTCATLG). The resulting tree was then fed into IQ-TREE for branch length optimization, with the Gamma model for rate heterogeneity. One hundred rapid bootstraps were executed in RAxML to provide branch support.

The same concatenated alignment was also used to estimate the branch lengths for the ASTRAL tree based on the 381 marker genes. Because the quality of an ASTRAL tree improves as the number of gene trees increases (Supplementary Fig. 10a, c), running ASTRAL on only 30 trees of structurally and functionally highly related genes is of limited validity. Thus we did duplicate the ASTRAL de novo, but only to assess the impact of ribosomal proteins on the branch lengths of the existing ASTRAL tree.

Species tree reconstruction and branch length estimation with CPR taxa excluded. We followed the same protocol as stated above to reconstructed species trees and estimate branch lengths based on the protein sequence alignments with the ribosomal proteins excluded. The aim was to determine how much the inclusion of the CPR taxa removed, leaving 9121 taxa. Only one model was used for each tree in the branch length estimation routine in order to reduce the computational expense for reconstructing the 381 gene trees: Instead of running RAxML three times per gene and selecting one tree with the highest Gamma likelihood, we ran RAxML once per gene using the random seed 12345. The two alternative site sampling schemes: most conserved (“cons”) and randomly selected (“rand”) as demonstrated in the main result were both tested, using the same amino acid sites as in the main protocol in each scheme.

Species tree reconstruction using site heterogeneous models (PMSF). We built alternative CONCAT trees using the posterior mean site frequency (PMSF) method68 implemented in IQ-TREE, which considers mixture classes of rates and substitution models (here the LG model) across sites. Because this method is computationally expensive, we downsampled the 10,575 taxa to 1000 (see below for the taxon downsampling strategy). ModelFinder (which is part of IQ-TREE)67 was used to select an optimal model among the empirical profile mixture models C10 to C6069, plus the site homogenous model (with Gamma rate across sites) as a control. In this analysis, the tree was considered as C60 as the optimal model for all tests. Therefore, we used the LG + C60 model for PMSF phylogenetic reconstruction. PMSF requires a guide tree, which we obtained from ModelFinder results. Computational challenge limited this analysis to at most 1000 taxa (which consumed 1.43 TB memory, close to the 1.5 TB physical memory equipped in our high-memory configurations). Branch support values were computed using UFBoot (UFBoot)69 method implemented in IQ-TREE. In parallel to this analysis, we performed phylogenetic inference using the Gamma model (+ G) or the Free-Rate model (+ R) on the same 1000-taxa input data for comparison purpose.

Species tree reconstruction using implicit methods. We applied two implicit strategies for inferring the evolutionary relationships among the sampled genomes. They are not based on the alignment of homologous features across multiple genomes, but instead are based on the predefined distances among genomes. Specifically, they are the Jaccard distances defined by the MinHash signature (see above), and by the presence/absence of the 400 marker genes (see above). The conventional neighbor joining (NJ) method as implemented in ClearCut 1.0.971 was used to reconstruct phylogenetic trees from the two distance matrices, respectively.

Rooting and post-manipulation of species trees. We rooted the species tree at the branch connecting the Archaea clade and Bacteria clade, according to the widely adopted hypothesis of life evolution72–75. This hypothesis is not without controversy75,76. The discovery and study of CPR life forms raised the possibility of an additional root (i.e., root → R) on the same 1000-taxon input data for comparison purpose.

Species tree reconstruction using implicit methods. We applied two implicit strategies for inferring the evolutionary relationships among the sampled genomes. They are not based on the alignment of homologous features across multiple genomes, but instead are based on the predefined distances among genomes. Specifically, they are the Jaccard distances defined by the MinHash signature (see above), and by the presence/absence of the 400 marker genes (see above). The conventional neighbor joining (NJ) method as implemented in ClearCut 1.0.971 was used to reconstruct phylogenetic trees from the two distance matrices, respectively.
node IDs can be used as unique identifiers in downstream analyses and applications.

Phylogeny-based downsampling of taxa. We designed a protocol to downsamples taxa from the 10,575 genomes for further phylogenetic analyses. We adopted the relative evolutionary divergence (RED) metric, as it is the core of our subsampling strategy. This metric allowed us to select the leaf set of the tree that best reflects the deep phylogeny. Specifically, we calculated RED for all nodes (terminal and internal) of the ASTRAL tree (i.e., the tree shown in Fig. 1) using TreeNode functions implemented in scikit-bio 0.5.2.27. Nodes were selected iteratively from the low end of the RED list, with ancestral nodes (if any) of the current node dropped from the selection at each iteration, until the desired number of clades n was achieved.

Within each selected clade, four criteria were sequentially applied to the descendants until one taxon was selected: (1) contains the most marker genes; (2) contamination level is the lowest; (3) DNA quality score is the highest; (4) random selection (if there were still more than one taxa after applying the other three criteria). This enabled the subsequent computation of the three metrics.

Comparison of multiple trees. We used both the classical Robinson–Foulds (RF) metric (calculated using scikit-bio’s “compare_rfd” function) and the quartet score (calculated using ASTRAL) to quantify the topological concordance between a pair of trees. A further challenge, we used the “tip distance” (TT), calculated using scikit-bio’s “compare_tip_distances” function, to measure the correlation of the phylogenetic distances among taxa in a pair of trees. It equals \(1 - \frac{r}{2}\), where r is the Pearson correlation coefficient between the tip-to-tip distance (i.e., total length of branches connecting two tips) matrices of the two trees. Because the two trees might have different sets of taxa, we first truncated them using the “shear” function implemented in scikit-bio so that they both only contained the shared taxa. This enabled the subsequent computation of the three metrics.

For a set of multiple trees (species trees or gene trees), a matrix of the pairwise RF distance, quartet distance (1–quartet score) or tip distance was constructed, based on which subsequent statistical analyses were performed to assess the clustering pattern of trees, as stated below.

Clustering analysis of multiple trees. We used several statistical approaches to assess the clustering pattern of multiple trees based on the RF, tip or quartet distance matrices built as stated above:

1. Hierarchical clustering, using the “linkage” function implemented in SciPy 1.1.0.
2. mMDS, as detailed above.
3. Principal coordinate analysis (PCOA), performed using QIIME 2’s “pcoa” command, and visualized using Emperor. This method aims to visualize the biggest variance in a few dimensions, as compared with mMDS as described above.
4. Permutational multivariate analysis of variance (PERMANOVA)82, performed using QIIME 2’s “beta-group-significance” command, with 999 permutations (the default setting). This method evaluates the statistical significance of grouping of trees by a certain variable such as method, site sampling and taxon sampling.

Cross-comparison of the ASTRAL and CONCAT trees. The first challenge for this comparison was that the branch support values were estimated using completely different methods (local posterior probability vs. rapid bootstrap) and so are not directly comparable. We manipulated the trees so that they have the same overall resolution: First, we collapsed the low-supported branches in the CONCAT tree (by conserved sites), using the commonly accepted bootstrap threshold: 50.

This left 9595 internal nodes. Then we performed branch collapsing to the ASTRAL tree, from the low end of the range of local posterior probability (lpp), until it reached 0.68057, also leaving 9595 internal nodes.

The second challenge was that large-scale trees are difficult to align and to display. We collapsed the two trees so that they have 50 pairs clades with at least 50 descendants each. For each pair of clades, the descendants are identical. The remaining tips were pruned. This operation left 7764 taxa in each tree. The sizes of the 50 chosen clades are 155.3 ± 106.9 (mean and standard deviation).

A tanglegram of the resulting collapsed trees was reconstructed using Dendroscope 3.5.98. In our case, the clades were fully aligned. The tanglegram was then rendered back-to-back without the need for displaying the connector lines.

Calculation of the relative Archaea-Bacteria distance. We calculated the phylogenetic distance (sum of branch lengths) between every pair of taxa in a tree using scikit-bio’s “tip distances” function. The pairs were divided into three groups: A–A, A–B, and B–B (A and B are abbreviations for Archaea and Bacteria). Within each group, the mean distance was calculated. Then the overall relative A–B distance was calculated as: mean (A–B)/(\(\text{mean (A–A)} \times \text{mean (B–B)}\)). Note that due to HGT and other reasons, archaeal and bacterial taxa are rarely perfectly separated in individual gene trees. Therefore, the calculated distance should be interpreted as the average evolutionary distance between archaeal and bacterial genomes, instead of the distance between the two clades.

Test for amino acid substitution saturation. We followed the principle introduced by Jeffroy et al.84 to test for the saturation. Specifically, we wanted to test whether the degree of saturation on inter-domain taxa pairs (Bacteria vs. Archaea) is larger than that on intra-domain pairs. For each domain, 100 taxa were randomly sampled for this analysis. We plotted the phylogenetic distance, i.e., the sum of branch lengths between two tips, as the x-axis, versus the Hamming distance of gap-free sites per each alignment between a pair of sequences, as the y-axis (Supplementary Fig. 19a–d). Because the three categories of taxon pairs have different distribution on the x-axis, we further binned on the x-axis and performed comparison within each bin (Supplementary Fig. 19c, f).

Phylogenetic analysis with latest genome availability. We made several modifications to the main protocol to reduce the computational burden of the rapid test of the extended set of 10,762 (10,575 + 187) genomes: UPP was called in “insertion” mode to update the existing amino acid sequence alignments. In-house scripts were used to locate the same set of sites instead of performing de novo site sampling. Both ASTRAL and CONCAT methods were used to build species trees. For ASTRAL, we used IQ-TREE with “fast” mode to build maximum likelihood trees from concatenated alignments without using a predefined starting tree. For ASTRAL, we kept the same analysis parameters to build a species tree from the 381 gene trees, whereas the gene trees were built as follows to save computation while maintaining high quality:

1. We used the gene trees as topological constraints (“g”) to incorporate the new taxa using RAXML. Then we used those trees as starting trees (“t”) to perform de novo ML searches using RAXML. This way, we only did de novo ML search once instead of three as previously, but we argued that the generated gene trees would have comparable ML score as in the previous procedure. To test this hypothesis, we randomly selected ten genes to generate four trees: (1) RAXML with FastTree as starting tree; (2) & (3) RAXML with random starting trees with two different random seeds; (4) RAXML tree generated using the described procedure. Note that the tree having highest likelihood score among (1), (2), (3) and (4) defines the ML tree in the previous procedure. Our results showed that the gene trees generated by (4) have higher likelihood scores than the best of (1), (2), and (3) in six of ten of the tested genes. Besides, we use a \(\chi^2\) test to show that the trees (4) have higher chance to be the best tree than (1), (2), and (3). In this test, the null hypothesis \(H_0\) is that (4) has the same chance to be the best tree among the four trees. Applying the test on the ten selected genes, we rejected \(H_0\) with \(p\)-value = 0.011.

Divergence time estimation using maximum likelihood. We used the maximum likelihood method (rs 1.8180) to estimate the divergence times based on the species trees. Specifically, we used the Langley–Fitch (LF) method86, which assumes a universal molecular clock (substitution rate) for the entire tree, with the truncated-Newton (TN) method for optimizing the likelihood of branch lengths87. A recent study showed that this method has comparable estimation accuracy when benchmarked against the more sophisticated Bayesian framework, but its computation is significantly faster88, thus suitable for the size of our data set. Near-zero branches were collapsed to avoid numerical errors. Ten replicates with random initial conditions were performed for each setting. In each replicate, three restarts were executed after the initial optimization with a random perturbation factor of 5%. Replicates that failed to pass the gradient check were discarded. The divergence times estimated by the run with the highest likelihood score, and the mean and standard deviation of those by all successful runs were reported.

Divergence time estimation using Bayesian inference. We used the Bayesian tool BEAST 1.10.489 to estimate divergence times. Considering the computational expense, we randomly selected 5000 amino acid sites from the full-length alignment, and downsamplled the original 10,575 taxa to 100. Taxon sampling was performed using the same RED-guided protocol (see above), but was manually modified afterwards to ensure sufficient sampling according to the calibration point. Two alternative molecular clock models were used: the strict clock model, or the uncorrelated relaxed clock model with a lognormal distribution (UCLD)90. The species tree was modeled using a Yule process82, with topology fixed to the ASTRAL tree. Logs of MCMC runs were examined using Tracer 1.7.180. Burn-in was randomly sampled for this analysis. We plotted the phylogenetic distance, i.e., the sum of branch lengths between two tips, as the x-axis, versus the Hamming distance of gap-free sites per each alignment between a pair of sequences, as the y-axis (Supplementary Fig. 19a–d). Because the three categories of taxon pairs have differential distribution on the x-axis, we further binned on the x-axis and performed comparison within each bin (Supplementary Fig. 19c, f).

Testing for amino acid substitution saturation. We followed the principle introduced by Jeffroy et al.84 to test for the saturation. Specifically, we wanted to test whether the degree of saturation on inter-domain taxa pairs (Bacteria vs. Archaea) is larger than that on intra-domain pairs. For each domain, 100 taxa were randomly sampled for this analysis. We plotted the phylogenetic distance, i.e., the sum of branch lengths between two tips, as the x-axis, versus the Hamming distance of gap-free sites per each alignment between a pair of sequences, as the y-axis (Supplementary Fig. 19a–d). Because the three categories of taxon pairs have differential distribution on the x-axis, we further binned on the x-axis and performed comparison within each bin (Supplementary Fig. 19c, f).
Tree-based taxonomic curation and annotation. We used the program tax2tree (commit 99f19be) to curate the original NCBI taxonomy assignment of genomes based on the phylogenetic trees and to annotate the internal nodes of the tree using most appropriate taxonomic labels. The same program was used to curate multiple databases, such as the classical Greengenes and the recent GTDB. The program took as input the species tree and the original NCBI taxonomy and inferred the most plausible taxonomic annotation at every node of the tree, as determined using an F-measure scoring system across candidate taxonomic terms. In scenarios where one term was estimated to be the best candidate for multiple, independent clades (i.e., paralogy/polyphyly), a numeric suffix was appended to the term to indicate the grouping and order (from more descendants to less) of those clades. For example, Firmicutes_3 is the largest clade assigned to the parapylogenetic phylum Firmicutes, followed by Firmicutes_2, Firmicutes_3, etc. Based on the decorated tree, correct taxonomic names were re-generated for unclassified and mis-annotated genomes. Taxonomic groups represented by only one genome in this work were back-filled post tax2tree annotation.

Assessment of cladistic properties of taxon sets. The cladistic property of a taxonomic group (or an arbitrarily defined taxon set) with reference to a species tree was evaluated using three methods:

1. The strict definition of “monophyley”: when a clade contains all genomes assigned to a single taxonomic group and no other genomes, this taxonomic group is considered monophyletic. Further, we identified “relaxed” monophyletic groups compared to the aforementioned “strict” scenario. In the “relaxed” scenario, if a clade consists of genomes assigned to a taxonomic group, and genomes without assignments at the same taxonomic rank (i.e., unclassified), this taxonomic group is still considered monophyletic.
2. tax2tree’s classification consistency score, representing the fraction of tips within that clade relative to the total number of tips in the tree which are of the same taxon. Consistency = 1 is equivalent to strict monophyletic.
3. The ASTRAL-computed quartet score of this taxonomic group, i.e., the fraction of quartets in the tree that supports this taxonomic group as monophyletic, i.e., separates this taxonomic group from the others.

An approach introduced in DiscoVista which evaluates and visualizes the compatibility between a given taxon set and a tree with branch support values. It computes a “support” or “rejection” degree as follows: If the taxon set constitutes a monophyletic clade in the tree, it is supported; and the support degree (green) is the support value of the branch connecting the lowest common ancestor of the clade to its parent. On the other hand, if it is not a monophyletic group in the original tree, but after contracting branches with support values below a threshold, the monophyly can no longer be rejected due to polytomy, the lowest threshold is considered the rejection degree (with a negative sign) (maagenta).

Evaluation of GTDB taxonomic groups. We downloaded GTDB release: 86.1 from https://gtdb.ecogenomic.org/. The format of genome identifier in GTDB was matched by that of our database (e.g., GR_GCA_000123456.1 was translated in the NCBI taxonomy G000123456). Following the protocols described above, we evaluated the GTDB phylogeny and taxonomic units, and annotated our species trees using the GTDB taxonomy.

Statistics. Statistical analyses and plotting were performed using Python 3.6 and QIIME 2 release 2017.12. Specifically, PERMANOVA test was performed using QIIME 2’s “beta-group-significance” command. Independent or paired two-sample t test was performed using scpy 1.0.0’s “ttest_ind” and “ttest_rel” commands, respectively. Fisher’s exact test was performed using scpy’s “fisher_exact” function. Linear regressions were performed using scpy’s “linregress” function. The p-value was computed using a two-sided Wald test, in which the null hypothesis was slope = 0. Gaussian kernel density estimations were performed using seaborn 0.9.0’s “distplot” function. Hierarchical clustering was performed using scpy’s “linkage” function. Quantile-quantile (Q–Q) plot was computed using scpy’s “probplot” command. Redundancy analysis (RDA) was performed using vegan 2.4.4’s “rda” and “ordiR2step” commands. Dimension reductions were performed using mDS implemented in scikit-learn 0.19.2, or PCoA implemented in QIIME 2 (both detailed above). Pairwise distances based on k-mer signatures and on marker gene presence/absence were computed using the Jaccard index (see above). Branch supports in the phylogenetic trees were computed using rapid bootstrap implemented in RAxML 8.2.10, and ultrafast bootstrap implemented in IQ-TREE 1.6.1, and local posterior probability implemented in ASTRAL 5.12.6a (detailed above). Robinson–Foulds (RF) distance and “tip distance” were calculated using scikit-bio 0.5.2. Quartet scores were calculated using ASTRAL.

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

Data availability
The data sets generated and analyzed during the current study are publicly available at GitHub (https://github.com/biocore/wol) and Zenodo (https://doi.org/10.5281/zenodo.3524546), under the BSD 3-Clause license. The source data underlying Figs. 1–6 and Supplementary Figs. 1–27 are provided as a Source Data file. All relevant data are available from the corresponding author.

Code availability
The Python implementations of the prototype selection algorithms for genome subsampling are publicly available from GitHub (https://github.com/biocore/wol) and Zenodo (https://doi.org/10.5281/zenodo.3524546), under the BSD 3-Clause license. A copy of the code is provided in Supplementary Software.

Received: 8 June 2019; Accepted: 6 November 2019; Published online: 02 December 2019

References
1. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl Acad. Sci. USA 74, 5888–5890 (1977).
2. 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).
3. Pace, N. R. A molecular view of microbial diversity and the biosphere. Science 276, 734–740 (1997).
4. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431 (2013).
5. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32, 822–828 (2014).
6. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35, 725–731 (2017).
7. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. 37, 179–185 (2019).
8. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208 (2015).
9. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353 (2017).
10. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Martin Embley, T. The archaeabacterial origin of eukaryotes. Proc. Natl Acad. Sci. USA 105, 20356–20361 (2008).
11. Hug, L. A. et al. A new view of the tree of life. Nat. Microbiol. 1, 16048 (2016).
12. Guy, L. & Ettema, T. J. G. The archaeal ‘TACK’ superphylum and the origin of eukaryotes. Trends Microbiol. 19, 580–587 (2011).
13. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell 172, 1181–1197 (2018).
14. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004 (2018).
15. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504, 231–236 (2013).
16. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. Nat. Rev. Microbiol. 15, 711 (2017).
17. O’Malley, M. A. & Koonin, E. V. How stands the tree of life a century and a half after the origin? Biol. Direct 6, 32 (2011).
18. Harris, J. K., Kelley, S. T., Spiegelman, C. B. & Pace, N. R. The genetic core of the universal ancestor. Genome Res. 13, 407–412 (2003).
19. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287 (2006).
20. Puigbò, P., Wolf, Y. I. & Koonin, E. V. Search for a ‘Tree of Life’ in the thicket of the phylogenetic forest. J. Biol. 8, 59 (2009).
21. Gogarten, J. P., Peter Gogarten, J., Ford Doolittle, W. & Lawrence, J. G. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226–2238 (2002).
22. Ramulu, H. G. et al. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? Mol. Phylegenet. Evol. 75, 103–117 (2014).
23. Creevey, C. J., Doerks, T., Fitzpatrick, D. A., Raes, J. & Bork, P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. PLoS ONE 6, e22099 (2011).
24. Haft, D. H. et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 46, D851–D860 (2018).
25. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of genomes. Nat. Commun. 4, 2934 (2013).

26. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).

27. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490 (2010).

28. Zhang, C., Babiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinforma. 19, 153 (2018).

29. Davidson, R., Vachaspati, P., Mirarab, S. & Warnow, T. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. BMC Genomics 16(Suppl 10), S1 (2015).

30. Yutin, N., Puigbò, P., Kooin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE 6, e9792 (2011).

31. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668 (2016).

32. Roch, S. & Snir, S. Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. J. Comput. Biol. 20, 93–112 (2013).

33. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl Acad. Sci. USA 96, 3801–3806 (1999).

34. Philippe, H. et al. Resolving phylogenetic heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109 (2004).

35. Gouy, R., Baurain, D. & Philippe, H. Rooting the tree of life: the phylogenetic jury is still out. Philos. Trans. R. Soc. Lond. B Biol. Sci. 376, 20140329 (2015).

36. Lartillot, N., Bruneel, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(Suppl 1), S4 (2007).

37. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling species tree heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenetic estimation. Syst. Biol. 67, 216–235 (2018).

38. Whitman, W. B. The modern concept of the prokaryote. J. Bacteriol. 191, 2000–2005 (2009).

39. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000).

40. Land, M. L. et al. Quality scores for 32,000 genomes. Stand. Genom. Sci. 9, 20 (2014).

41. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132 (2016).

42. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinforma. 11, 119 (2010).

43. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. Bellerophon: an improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320 (2008).

44. Valdar, W. S. J. Scoring residue conservation. Proteins: Struct., Funct., Genet. 76, 227–241 (2002).

45. Keul, F., Hess, M., Goesel, M. & Hamacher, K. PFASUM: a substitution matrix from Pfam structural alignments. BMC Bioinforma. 18, 293 (2017).

46. Lartillot, N., Brinkmann, H. & Philippe, H. Empirical protein substitution models for phylogenetic reconstruction. Bioinformatics 24, 2177–2183 (2008).

47. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFB002: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522 (2018).

48. Caviedes-Vidal, T., Yang, W. & Shi, X. Rooting the tree of life by transition analyses. Biol. Direct 1, 3 (2006).

49. Virtanen, P. et al. SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python. arXiv:1907.10121 (2019).

50. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32–46 (2001).

51. Jeffroy, O., Brinkmann, H., Deluc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231 (2006).

52. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19, 301–302 (2003).

53. Langley, C. H. & Fitch, W. M. An examination of the constancy of the rate of nucleotide substitution. Evolution 39, 130–148 (1985).

54. Lartillot, N., Brinkmann, H. & Philippe, H. Rooting the ribosomal tree of life. Proc. Natl Acad. Sci. USA 106, 7192–7198 (2009).

55. Nash, S. G. A survey of truncated-Newton methods. J. Comput. Appl. Math. 124, 45–59 (2000).

56. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data analysis using Bayesian inference with BEAST 1.7.1. PLoS Comput. Biol. 4, e1000181 (2008).

57. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88 (2006).
91. Gerhard, T. The conditioned reconstructed process. J. Theor. Biol. 253, 769–778 (2008).
92. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in bayesian phylogenetics using tracer 1.7. Syst. Biol. 67, 901–904 (2018).
93. McDonald, D. et al. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610–618 (2012).
94. Federhen, S. The NCBI taxonomy database. Nucleic Acids Res. 40, D136–D143 (2012).
95. Sayyari, E., Whitfield, J. B. & Mirarab, S. DiscoVista: interpretable visualizations of gene tree discordance. Mol. Phylogenet. Evol. 122, 110–115 (2018).

Acknowledgements
We are grateful to Victor Nizet, Eric Allen, Milton Saier, Jürgen Schulze, Toni Gabaldón Estevan, Benjamin Hillmann, Arturo Medrano-Soto, Madhusudan Gujral, Monika Kumaraswamy, Jeff Dereus, Anupriya Tripathi, Amnon Amir, Serene Jiang, Yoshiki Vásquez-Baera, Thant H. Zaw for insightful discussions on this study and additional assistance. This study was supported in part by National Science Foundation grant 1565057 and the Alfred P. Sloan Foundation under grant number G-2017-9838 awarded to R.K. This work used the Comet supercomputer at the San Diego Supercomputer Center through allocation BIO150043, awarded to R.K., L.S. and S.M. through the Extreme Science and Engineering Discovery Environment (XSEDE). Both Comet and XSEDE are funded by the National Science Foundation. S.M., U.M., and M.R. were supported by the National Science Foundation grant III-1845967 to S.M. W.L. was supported by National Natural Science Foundation of China (91951205).

Author contributions
Q.Z., R.K. and S.M. conceived the study; Q.Z. and U.M. led data analysis; U.M. and S.J. led algorithm development and implementation; W.P., U.M. and F.A. led high-performance computing; Q.Z. led the results interpretation and paper writing; Q.Z., G.A.A., J.B., Z.W., E.S., M.R., K.C., Y.Y. and S.M. contributed to algorithm development and implementation; U.M., S.J., J.G.S., P.B., D.M., S.H., N.S., J.J., S.P., C.H., S.M. and R.K. contributed to the result interpretation; F.A., T.K., I.T.M. and S.M. contributed to data analysis; Q.Z., F.A., E.K. and Z.Z.X. contributed to data collection; J.G.S., D.M., D.K., W.L., N.S., L.S., S.M. and R.K. contributed to study design; all coauthors contributed to the writing and discussion of the paper.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13443-4.

Correspondence
Correspondence and requests for materials should be addressed to R.K.

Peer review information
Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission 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.

Open Access
This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019