nature biotechnology

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

The rapid expansion of sequenced bacterial and archaeal genomes in the past decade has enabled the construction of genome-based phylogenies[1–3] suitable for defining taxonomy. A robust taxonomy is needed to accurately describe microbial diversity, to interpret metagenomic data and to provide a common language for communicating scientific results[4]. Sequence-based phylogenetic trees provide a framework for the development of a taxonomy that takes into account both evolutionary relationships and differing rates of evolution. Current microbial taxonomies such as those provided by NCBI[5], SILVA[6], RDP[7], Greengenes[8] and EzTaxon[3] are often inconsistent with evolutionary relationships, because many taxa circumscribe polyphyletic groupings. This inconsistency is partly attributable to historical phenotype-based classification, as exemplified by the clostridia: microorganisms sharing morphological similarities have been erroneously classified in the genus *Clostridium*[9,10]. Modern microbial taxonomy is primarily guided by 16S rRNA relationships, and such discrepancies are observable in 16S rRNA gene trees[6,8], but most have not been corrected, owing to the scale of the task and the lengthy process of formally reclassifying microorganisms[11].

A second, less obvious, issue with existing sequence-based microbial taxonomies is the uneven application of taxonomic ranks across the tree. Regions that are the subject of intense study tend to be split into more taxa than other parts of the tree with equivalent phylogenetic depth; for example, the family Enterobacteriaceae (comprising dozens of genera) is equivalent to a single genus in other parts of the tree, such as *Bacillus*[12]. Conversely, understudied groups are often lumped together; for example, the phylum Synergistetes is currently represented by a single family[13] that would constitute multiple family-level groupings in more intensively studied parts of the tree. A proposal to standardize taxonomic ranks by using 16S rRNA sequence identity thresholds has identified a high degree of discordance between these thresholds and the SILVA taxonomy[11].

Current microbial taxonomies based on 16S rRNA gene relationships[3,6–8] have several limitations, including low phylogenetic resolution at the highest and lowest taxonomic ranks[14], missing diversity as a result of primer mismatches[15] and PCR-produced chimeric sequences that can corrupt tree topologies by drawing together disparate groups[16]. Trees inferred from the concatenation of single-copy vertically inherited proteins provide higher resolution than those obtained from a single phylogenetic-marker gene[17–19] and are increasingly representative of microbial diversity, as culture-independent techniques are now producing thousands of metagenome-assembled genomes (MAGs) from diverse microbial communities[20–22]. Despite some caveats of their own, including potential lateral gene transfer, differing rates of evolution, and recombination[19,23], concatenated protein trees have been extensively used in the literature[20,24,25] and have been proposed as the best basis for a reference bacterial phylogeny[26].

Here we present a phylogeny inferred from the concatenation of 120 ubiquitous single-copy proteins, and we used this phylogeny to propose a bacterial taxonomy that covers 94,759 bacterial genomes, including 13,636 (14.4%) from uncultured organisms (metagenome-assembled or single-cell genomes). Taxonomic groups in this classification describe monophyletic lineages of similar phylogenetic depth after normalization for lineage-specific rates of evolution. This taxonomy, which we have named the GTDB taxonomy, is publicly available at the Genome Taxonomy Database website (http://gtdb.ecogenomic.org/).

Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, University of Queensland, Queensland, Australia. Correspondence should be addressed to P.H. (p.hugenholtz@uq.edu.au).

Received 27 November 2017; accepted 27 July 2018; published online 27 August 2018; doi:10.1038/nbt.4229

RESULTS

Deriving the GTDB taxonomy

A data set comprising 87,106 bacterial genomes was obtained from RefSeq/GenBank release 80 and augmented with 11,603 MAGs recovered from Sequence Read Archive metagenomes according to the approach of Parks *et al.*[22]. After removal of 2,482 of these genomes on the basis of a completeness/contamination threshold and 1,468 genomes on the basis of a multiple sequence alignment (MSA) threshold, the resulting 94,759 genomes were dereplicated to remove highly similar genomes with high-quality reference material retained as representatives when possible (Online Methods). Nearly 40% (8,559) of the dereplicated data set of 21,943 genomes represents uncultured organisms reflecting the microbial diversity currently being revealed by culture-independent techniques[20–22]. A bacterial genome tree was inferred from the dereplicated data set by applying FastTree to a concatenated alignment of 120 ubiquitous single-copy proteins[22] (subsequently referred to as 'bac120') comprising a total of 34,744 columns after trimming of 1,021 columns represented in <50% of the genomes and 5,390 columns with an amino acid consensus <25% (Online Methods). The bac120 data set represents ~4% of an average bacterial genome and is comparable to other bacterial domain marker sets[27,28].

Having inferred the concatenated protein phylogeny, we annotated the tree with group names by using the NCBI taxonomy[5] standardized to seven ranks (Online Methods). Taxon names were overwhelmingly assigned to interior nodes with high bootstrap support (99.7% ± 2.9%) to ensure taxonomic stability. However, a few poorly supported nodes (<70%) in the bac120 tree were assigned names on the basis of independent analyses or to preserve widely used existing classifications (**Supplementary Table 1** and Firmicutes example below). Because more than one-third of the data set represents uncultured organisms, a substantial part of the tree was not effectively annotated with the NCBI genome taxonomy. Therefore, 16S rRNA gene sequences present in the MAGs were classified against the Greengenes[8] 2013 and SILVA[6] v123.1 taxonomies to provide additional taxonomic identifiers. Using a set of criteria to ensure accurate mapping between 16S rRNA and MAG sequences (Online Methods), we labeled 74 groups lacking cultured representatives with 16S rRNA-based names, including well-recognized clades such as SAR202 (ref. 29), WS6 (ref. 30) and ACK-M1 (ref. 31) (**Supplementary Table 2**). We term all such alphanumeric names nonstandard placeholders to be replaced with standard validated names in due course. Curation of the taxonomy then involved two main tasks: the removal of polyphyletic groups and the normalization of taxonomic ranks according to relative evolutionary divergence (RED).

Removal of polyphyletic groups

Twenty phyla and 25 classes as defined by the NCBI taxonomy could not be reproducibly resolved as monophyletic in the bootstrapped bac120 tree (**Supplementary Table 3**). Most of these were the result of a small number of misclassified genomes; however, some taxa seemed to be truly polyphyletic, including well-known lineages such as the Firmicutes and Proteobacteria (**Supplementary Table 3**). The instability of the Firmicutes has previously been noted, primarily as a result of the Tenericutes and/or Fusobacteria moving into or out of the group[25,32]. In this prominent case, we chose to preserve the existing classification until more in-depth phylogenetic analyses are performed to resolve the issue (rationale described below). Other poorly supported lineages such as the Proteobacteria, which have been widely reported to be polyphyletic on the basis of the 16S rRNA gene[8,33] and protein markers[34,35], were conservatively divided into stable monophyletic groups. When possible, polyphyletic taxa containing the nomenclature type retained the name, and all other groups were renamed according to the International Code of Nomenclature of Prokaryotes (Online Methods). For lower-level ranks, notably genus, existing names were often retained with alphabetical suffixing to resolve polyphyly in the bac120 tree (for example, *Bacillus*_A, *Bacillus*_B and so forth). Only the group containing type material (if known) kept the original unsuffixed name to indicate the validity of the name assignment. This procedure serves two purposes: it preserves continuity in the literature, and it avoids the necessity to propose dozens of new names for highly polyphyletic groups, although we suggest that such renaming should ultimately be done. A total of 436 genera, 152 families and 67 orders were identified as polyphyletic in the tree, thus highlighting important deficiencies in the current taxonomy (**Supplementary Table 3**). The genus *Clostridium* was the most polyphyletic, representing 121 genera spanning 29 families, and was followed by *Bacillus* (81 genera across 25 families) and *Eubacterium* (30 genera across 8 families). However, these numbers were also influenced by rank normalization in some cases (described below).

Taxonomic-rank normalization

There is currently no accepted standardized approach for assigning species to higher taxonomic ranks (i.e., genus to phylum), although 16S rRNA sequence identity and amino acid identity (AAI) thresholds have been proposed[11,36,37]. The assignment of ranks within the NCBI taxonomy is highly variable under both these measures, because they have been proposed relatively recently and have not been widely adopted[2,11]. We normalized the assignment of higher taxonomic ranks by using RED calculated from the bac120 tree, an approach conceptually similar to that used by Wu *et al.*[38]. Our method provides an operational approximation of relative time with extant taxa existing in the present (RED = 1), the last common ancestor occurring at a fixed time in the past (RED = 0) and internal nodes being linearly interpolated between these values according to lineage-specific rates of evolution (**Fig. 1** and Online Methods). RED intervals for normalizing taxonomic ranks were defined as the median RED value for taxa at each rank ± 0.1 (**Fig. 1**). This procedure represents a compromise between strict normalization and the desire to preserve existing group names on well-supported interior nodes. Visualization of the NCBI taxonomy according to RED highlighted a substantial number of over- or underclassified taxa according to the proposed criteria (**Fig. 2a**). To correct these inconsistencies, we reassigned taxa falling outside of their RED intervals to either a new taxonomic rank (with appropriate nomenclatural changes) or a new node in the tree (**Fig. 2b**).

In contrast to 16S rRNA sequence identity or AAI thresholds, RED normalization accounts for the phylogenetic relationships between taxa and variable rates of evolution. For example, members of the rapidly evolving genus *Mycoplasma*[39] (**Fig. 1**) are sufficiently diverged to represent two phyla on the basis of a 16S rRNA gene sequence identity threshold of 75% (ref. 11). However, vertebrate-associated *Mycoplasma* and *Ureaplasma* diverged from their arthropod-associated sister families only 400 Ma (ref. 39), as is approximately consistent with the emergence of vertebrates[40]. This evolutionary event occurred much later than the primary diversification of bacterial phyla, which is estimated to have occurred between 2 and 3 Ga (ref. 41). The relatively recent emergence of *Mycoplasma* is more consistent with their RED-normalized ranking into a single order within the Firmicutes (**Fig. 2b**) than the two phyla that would be indicated by a 16S rRNA sequence identity of 75%.

Validation of the GTDB taxonomy

The robustness of the approach used to generate the GTDB taxonomy was evaluated with various tree-inference software, evolutionary

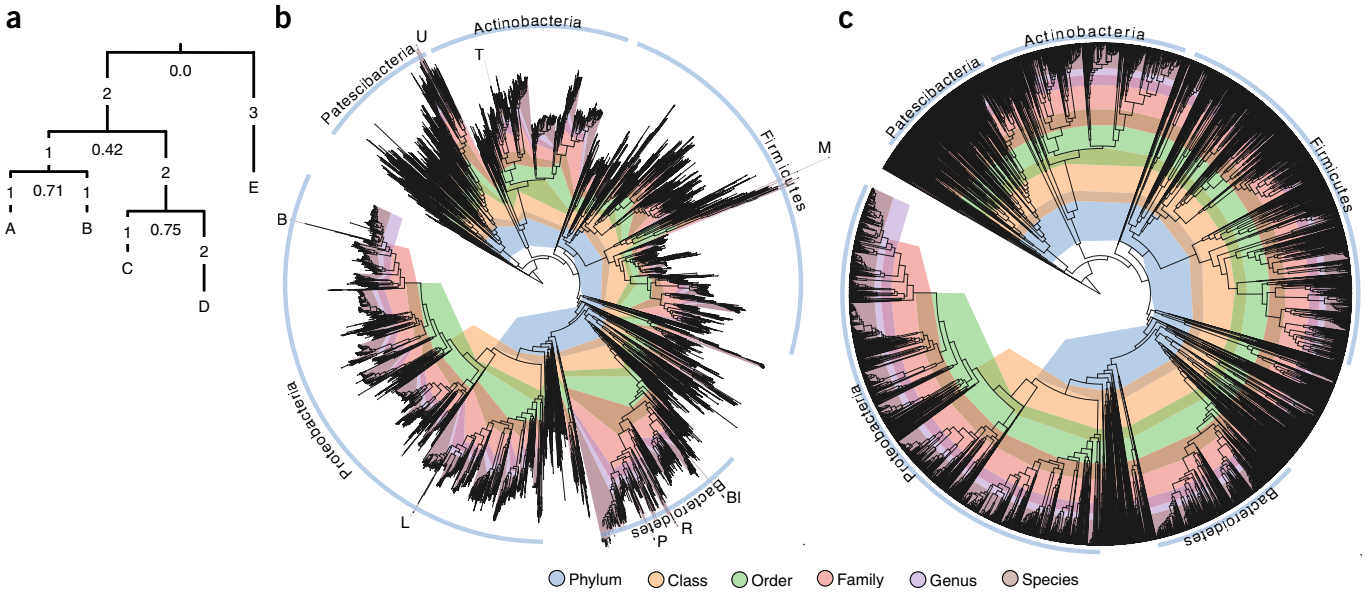

Figure 1 Rank normalization through RED. (**a**) Example illustrating the calculation of RED. Numbers on branches indicate their length, and numbers below each node indicate their RED. The root of the tree is defined to have a RED of zero, and leaf nodes have a RED of one. The RED of an internal node n is linearly interpolated from the branch lengths comprising its lineage, as defined by $p + (d/u) \times (1 - p)$, where p is the RED of its parent, d is the branch length to its parent, and u is the average branch length from the parent node to all extant taxa descendant from n. For example, the parent node of leaves C and D has a RED value of 0.75 ($0.42 + (2/3.5) \times (1 - 0.42)$), because its parent has a RED of $p = 0.42$, the branch length to the parent node is $d = 2$, and the average branch length from the parent node to C and D is $u = (3+4)/2 = 3.5$. (**b**) Bacterial genome tree inferred from 120 concatenated proteins (bac120) and contoured with the RED interval assigned to each taxonomic rank. Adjacent ranks overlap in some instances, because this permits existing group names to be placed on well-supported interior nodes. To accommodate visualizing the RED intervals, the initial tree inferred across 21,943 was pruned to 10,462 genomes by retaining one genome per species. The tree is rooted on the phylum Acetothermia for illustrative purposes. RED values used for rank normalization are averaged over multiple plausible rootings (Online Methods). Examples of taxa with high expected substitution rates are as follows: U, o__UBA9983; T, s__Tropheryma whipplei; M, o__Mycoplasmatales; Bl, f__Blattabacteriaceae; R, g__RC9; P, g__Porphyromonas; L, g__Liberibacter; and B, g__Buchnera. Prefixes indicate taxonomic ranks. (**c**) The bac120 tree, with branch lengths scaled by RED values, illustrating that rank normalization follows concentric rings that provide an operational approximation of the relative time of divergence.

models, marker sets and genome data sets. We first considered trees inferred with ExaML and IQ-TREE. Because these methods are computationally intensive, it was necessary to decrease the bac120 MSA from 34,744 to 5,038 columns by evenly sampling columns across each of the 120 proteins and to use subsampled sets of 4,985 or 10,462 genomes dereplicated to retain one genome per GTDB genus or species, respectively (Online Methods). We also inferred trees by using FastTree with the reduced MSA and subsampled genome sets to isolate the effect of inference software from data-set reduction. For each of these trees, we determined the optimal position of each GTDB taxon and classified a taxon as monophyletic, operationally monophyletic (defined as having an F measure ≥0.95) or polyphyletic (Online Methods). Most GTDB taxa above the rank of species and with two or more genomes were found to be monophyletic or operationally monophyletic, and only 79 of 2,586 (3.1%) taxa were polyphyletic in one or more of the species-dereplicated FastTree, IQ-TREE or ExaML trees (**Fig. 3a** and **Supplementary Table 4**). Notably, 44 of the 79 polyphyletic taxa were found to be polyphyletic in the species-dereplicated FastTree, suggesting that most of the identified incongruence with GTDB taxa was the result of using a subsampled MSA and a dereplicated set of genomes. On average, 95.2% (IQ-TREE), 96.5% (ExaML) and 96.9% (FastTree) of GTDB taxa at each taxonomic rank were classified as monophyletic or operationally monophyletic within the species-dereplicated trees (**Fig. 3a** and **Supplementary Fig. 1a**). Taxa that were not monophyletic within the species-dereplicated trees were most often a result of the incongruent placement of a small number of genomes, thus resulting in either direct conflict with the GTDB taxonomy or unresolved groups in the tree (Online Methods). Less than 0.1% of genomes had a conflicting taxonomic assignment at any rank in any of the three species-dereplicated trees, and <1.6% had an unresolved taxonomic assignment at any rank, with the exception of order-level assignments in the ExaML tree, for which 7.5% were unresolved (**Supplementary Fig. 1b** and **Supplementary Table 5**). This result was primarily due to fragmentation of the order Bacillales in the ExaML tree, which was one of the poorly supported nodes in the bac120 tree (**Supplementary Table 1**). Taxa at the same taxonomic rank were also observed to have similar RED values in all three species-dereplicated trees, thus indicating that rank normalization is robust to the maximum-likelihood method used, MSA subsampling and genome dereplication (**Fig. 3a**, **Supplementary Fig. 1c** and **Supplementary Table 1**). Similar results were observed for the genus-dereplicated trees and are summarized in **Supplementary Tables 1** and **4**. The GTDB taxonomy was also robust to model selection: only three taxa were polyphyletic in a tree inferred with FastTree under the LG protein-substitution model instead of the WAG model (**Supplementary Table 1**).

Having established that the GTDB taxonomy is robust across different maximum-likelihood-inference software, we next considered the effect of different marker sets. Applying FastTree to a concatenated alignment of 16 ribosomal proteins[20,25] (rp1) resulted in only 199 of the 4,501 (4.4%) GTDB taxa above the rank of species being classified as polyphyletic (**Fig. 3b** and **Supplementary Table 4**). On

average, 94.7% of GTDB taxa at each taxonomic rank were monophyletic or operationally monophyletic within the rp1 tree; the least was 92.7% at the class level, and the most was 96.5% at the order level (**Fig. 3b** and **Supplementary Fig. 2a**). Less than 0.5% of genomes had a conflicting taxonomic assignment at any rank, and <1.5% had an unresolved taxonomic assignment at any rank (**Supplementary Fig. 2** and **Supplementary Table 5**), with the exception of order-level assignments, which were unresolved for 4.0% of genomes. This result was largely due to an instability of the Enterobacterales probably caused by the inclusion of a highly reduced endosymbiont genome, '*Candidatus* Zinderia insecticola', in the rp1 tree. As with the inference-software comparisons, we observed that taxa at the same taxonomic rank had similar RED values, thus indicating that rank normalization was largely preserved in the rp1 tree (**Fig. 3b** and **Supplementary Fig. 2c**). Performing the same analysis on a 16S rRNA gene tree resulted in 387 of the 2,576 (15.0%) GTDB taxa above the rank of species, with two or more genomes being classified as polyphyletic; and 78.1% (species) to 90.8% (class) of GTDB taxa being recovered as monophyletic or operationally monophyletic (**Fig. 3b** and **Supplementary Fig. 3a**). Incongruent taxonomic assignments in the 16S rRNA tree were largely the result of unresolved taxa, and <1.1% of genomes had conflicting assignments at any taxonomic rank (**Supplementary Fig. 3b** and **Supplementary Table 5**). Taxa at the same rank had similar RED values in the 16S rRNA gene tree, though the spread of values was greater than observed on the bac120 or rp1 trees (**Fig. 3b** and **Supplementary Fig. 3c**).

For comparison, we evaluated the congruence of the NCBI taxonomy with the trees inferred by using different inference software (species-dereplicated FastTree, IQ-TREE and ExaML) and marker sets (bac120, rp1 and 16S rRNA). In contrast to the GTDB taxonomy, all trees had numerous discrepancies with the NCBI taxonomy, in terms of both polyphyly and over- and underclassified taxa (**Figs. 2** and **3**). On average, 26.1% (rp1) to 28.0% (species-dereplicated FastTree) of NCBI taxa were classified as polyphyletic in these trees, and taxa at the same taxonomic rank had highly variable RED distributions (**Fig. 3** and **Supplementary Figs. 4**–**7**). Only 59.5% to 64.2% of genomes had NCBI taxonomy assignments congruent with the topology of these trees, whereas 76.1% to 96.8% had GTDB assignments in agreement with the tree topologies (**Table 1**).

Trees inferred from alternative-marker sets showed a higher degree of discordance with the GTDB taxonomy than those inferred by using alternative maximum-likelihood-inference software. To further explore the relationship between alternative-marker sets and inference methods (including neighbor joining), we calculated pairwise tree distances between all trees (**Fig. 3c**,**f** and **Supplementary Table 6**). These results showed that, in terms of both tree topology and supported splits, the maximum-likelihood-inference software used is less critical than the choice of marker set, and that genome dereplication and MSA subsampling also have a nontrivial effect on the inferred tree.

The stability of the GTDB taxonomy on trees inferred by using subsets of the bac120 marker set and under taxon subsampling was also evaluated in anticipation of decreasing computational burden as the database size increases. Subsampling of the 120 bacterial marker genes was performed 100 times with 60 of the markers randomly selected for each replicate. Notably, 96.7% of GTDB taxa were classified as monophyletic in ≥90% of the replicate trees, and only ten taxa (0.11%) were classified as polyphyletic in ≥50% of replicates (**Supplementary Table 1**). Given the lower phylogenetic resolution of individual genes[26,42], the results from individual gene trees were also highly robust: 86.1% of GTDB taxa were monophyletic in ≥50% of trees (**Supplementary Table 1**), and all gene trees recovered ≥51.6% of GTDB phyla and ≥82.0% of GTDB genera as monophyletic or operationally monophyletic (**Supplementary Table 7**). Taxon resampling with one genome per genus was performed 100 times, and representative genomes were randomly selected in each replicate. Across the 1,430 taxa with two or more genera, 97.5% were recovered as monophyletic in ≥90% of the taxon-resampled trees, and only four taxa were classified as polyphyletic in ≥50% of replicates (**Supplementary Table 1**).

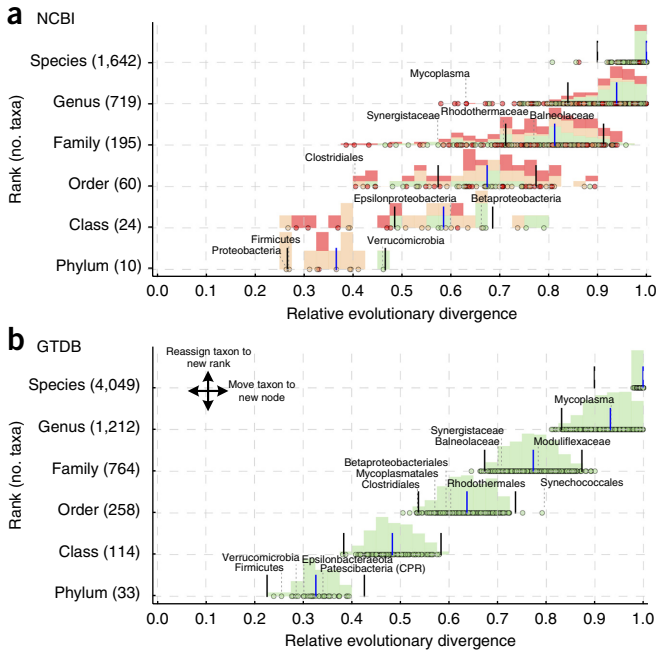

Figure 2 RED of NCBI and GTDB taxa in a genome tree inferred from 120 concatenated proteins. (**a**,**b**) RED of taxa defined by the NCBI (**a**) and GTDB (**b**) taxonomies. Each point represents a taxon distributed according to its rank (*y* axis) and is colored green, orange or red to indicate monophyletic, operationally monophyletic or polyphyletic in the genome tree, respectively. A histogram is overlaid on the points to show the relative density of monophyletic, operationally monophyletic and polyphyletic taxa. The median RED value of each rank is shown by a blue line, and the RED interval for each rank is shown by black lines. Only monophyletic or operationally monophyletic taxa were used to calculate the median RED values for each rank. The GTDB aims to resolve taxa that are over- or underclassified on the basis of their RED value by either reassigning them to a new rank (vertical shift in plot) or moving them to a new interior node (horizontal shift in plot). For example, the family Synergistaceae was normalized by reclassifying the family to encompass only the genera *Synergistes*, *Cloacibacillus*, *Thermanaerovibrio* and *Aminomonas*, rather than the 12 genera circumscribed by this family in the NCBI taxonomy. Only taxa with two or more subordinate taxa are plotted, because these taxa have positions in the tree indicative of their rank (for example, only 33 of the 99 phyla defined by the GTDB contain two or more classes, and a phylum with a single class consisting of multiple orders is expected to have a RED value commensurate with the rank of class). The number of taxa plotted at each rank is given in parentheses along the *y* axis.

Comparison of GTDB with other classifications

Overall, 58% of the 84,634 genomes with an NCBI taxonomy had one or more changes to their classification above the rank of species (**Fig. 4a**). These changes included both reclassification of taxa and filling in

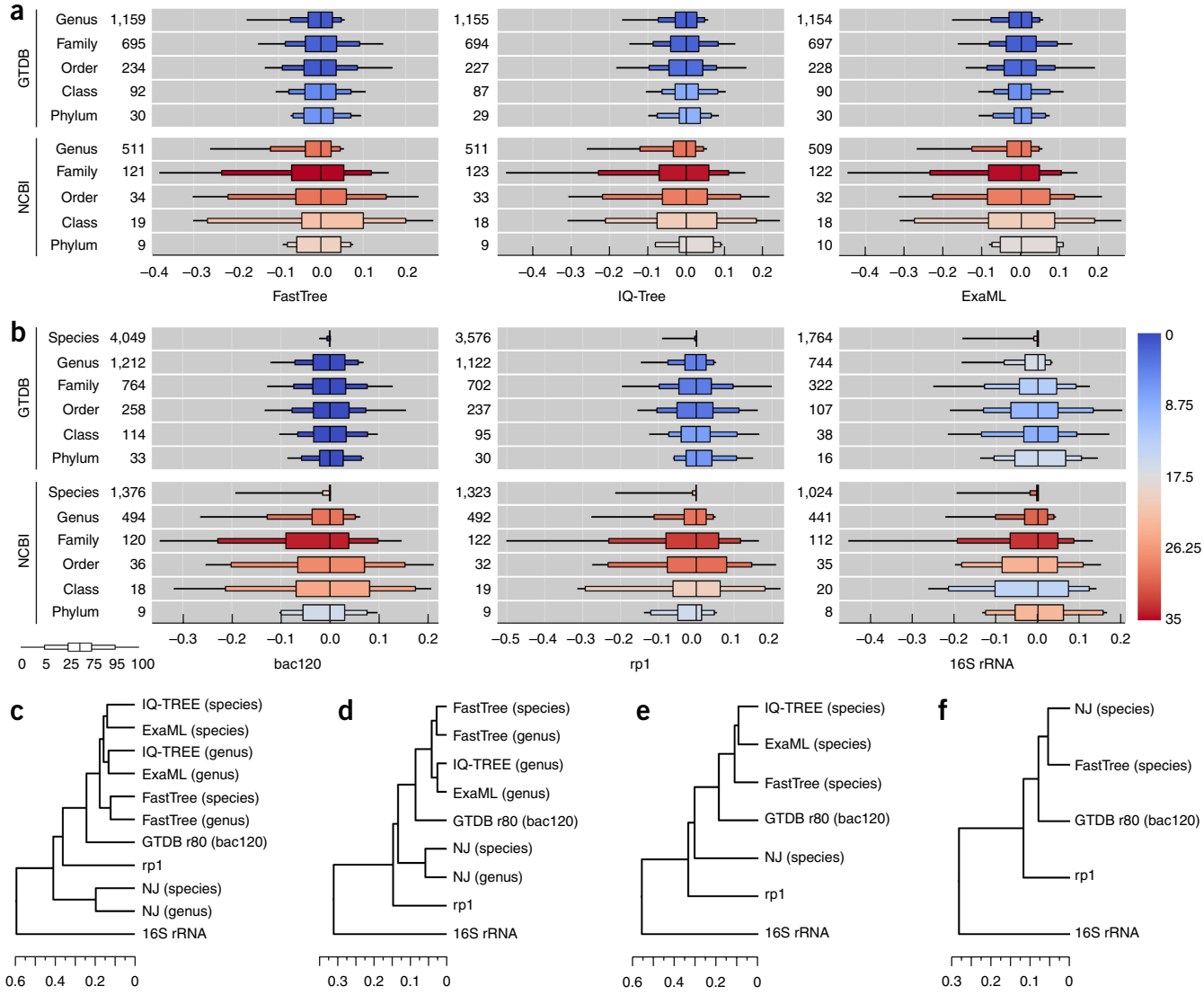

Figure 3 RED and polyphyly of GTDB and NCBI taxa on trees inferred by using varying inference methods and marker sets. (**a**) Trees inferred with FastTree, IQ-TREE and ExaML from the concatenated alignment of 120 bacterial proteins and spanning 10,462 genomes dereplicated to one genome per species. RED distributions for taxa at each rank are shown relative to the median RED value of the rank. Results are summarized in box-and-whisker plots indicating percentiles 0/100, 5/95, 25/75 and 50. Distributions were calculated over monophyletic and operationally monophyletic taxa with two or more subordinate taxa, because these taxa have positions in the tree indicative of their rank. The number of taxa comprising each distribution is shown next to each box-and-whisker plot. The percentage of taxa classified as polyphyletic in each tree at each rank is indicated by a color gradient from blue to red. (**b**) Analogous results for trees inferred with FastTree by using 120 bacterial proteins (bac120), 16 ribosomal proteins (rp1) or the 16S rRNA gene and spanning the dereplicated set of 21,943 genomes used to define the GTDB. Plots showing the RED values of individual GTDB and NCBI taxa are shown in **Figure 2** and **Supplementary Figures 1–7**. (**c**) Hierarchical-cluster tree illustrating the Robinson–Foulds distance between trees inferred with different maximum-likelihood methods, neighbor joining (NJ) and alternative-marker sets (rp1 and 16S rRNA) over a common set of 4,985 genomes constructed by sampling one genome per GTDB genus. The alternative inference methods were also applied to trees originally dereplicated to one genome per species, which were subsequently pruned to the common set of 4,985 genomes. The bac120 tree was used to define the GTDB r80 taxonomy. (**d**) Hierarchical-cluster tree illustrating the proportion of supported splits in common among trees over the common set of 4,985 genomes. (**e**,**f**) Analogous plots to **c** (**e**) and **d** (**f**), except that pairwise distances were calculated over trees defined on a common set of 10,462 genomes constructed by sampling one genome per GTDB species. Because nonparametric bootstraps could not be determined for IQ-TREE and ExaML when dereplicated at the species level, these trees do not appear in **f**.

missing rank name information (~3% of genus to phylum names are currently undefined across the 84,634 genomes with an NCBI taxonomy). On average, 19% of names were changed per rank, the least being 7% at the phylum level and the most being 31% at the order level (**Fig. 4a**). A total of 199 NCBI names above the rank of species were 'retired' from the GTDB taxonomy mostly as a result of RED normalization (**Supplementary Table 8**). An analogous comparison to the SILVA taxonomy also showed substantial differences across all taxonomic ranks: 66% of genomes had one or more changes above the rank of species (**Supplementary Table 9** and **Supplementary Fig. 8**). Many of these differences are in common with the NCBI taxonomy, owing to the GTDB rank normalization process; however,

Table 1 Congruency of GTDB and NCBI taxonomic classifications with tree topology

Tree	No. NCBI genomes[a]	GTDB (%)	NCBI (%)
bac120	10,411	100	64.1
FastTree (species dereplicated)	8,905	96.0	61.1
IQ-TREE (species dereplicated)	8,905	96.8	64.2
ExaML (species dereplicated)	8,905	90.3	61.0
rp1	9,815	89.9	60.2
16S rRNA	7,243	76.1	59.5

[a]Number of genomes with an NCBI classification. These genomes were used for comparing the congruencies of the taxonomies with tree topology.

there are also many documented differences between NCBI and SILVA[43].

Only 18% of taxon names in the GTDB taxonomy above the rank of species have been validly published; a further 19% have been proposed but not validated; and the remaining 63% are currently nonstandard placeholder names (**Fig. 4b**), thus indicating the scope of the task remaining to produce a fully standardized taxonomy consisting of validated names. This task will be greatly facilitated by recent proposals to use genome sequences as type material for as-yet-uncultured lineages, which in principle would allow for validation of names[44,45].

Genus- and species-level classifications

Genera and species comprise 84% of the 16,924 defined taxon names in the bac120 tree. Misclassified species in the public repositories are an area of particular concern to researchers, because they can introduce noise into a variety of analyses, including strain typing[46], biogeographic distributions of species[47] and pangenome analyses[48]. Moreover, classification errors can propagate over time as incorrectly labeled genomes are used as reference material to identify novel sequences. A small number of microbial genera have been rigorously examined for this problem, and taxonomic corrections have been proposed, including *Aeromonas*[49] and *Fusobacterium*[50]. We compared the results of these analyses to the GTDB taxonomy as a means of providing an independent verification of our results. On the basis of multilocus sequence analysis and average nucleotide identity (ANI) comparisons, Beaz-Hidalgo *et al.*[49] have proposed that nine *Aeromonas dhakensis* genomes are incorrectly classified as *Aeromonas hydrophila*. All nine of these genomes were reclassified as *A. dhakensis* in the bac120 tree, and an additional four genomes not included in the Beaz-Hidalgo study were also reclassified as *A. dhakensis* (**Supplementary Table 10**). Kook *et al.*[50] have recently recommended the reclassification of *Fusobacterium nucleatum* subspecies *animalis*, *nucleatum*, *polymorphum* and *vincentii* as separate species, on the basis of ANI and genome distance metrics. Rank normalization of the GTDB taxonomy by using RED values largely reproduced this finding without prior knowledge of the authors' work (**Supplementary Table 10**). Reclassification of species according to the bac120 tree is also consistent with recent efforts to objectively define bacterial species according to barriers to homologous recombination estimated against the core genome of each species[51]. In that study, 23 of 91 bacterial species have been proposed to contain one or more members not belonging to their respective species ('excluded taxa'). We found that almost all comparable instances of excluded taxa were due to misclassification in the NCBI taxonomy (**Supplementary Table 10**). These results suggest that the bac120 tree topology and RED estimates of species-level groups based on ~4% of the genome (120 conserved markers) are consistent with alternative analytical approaches using larger fractions of the genome.

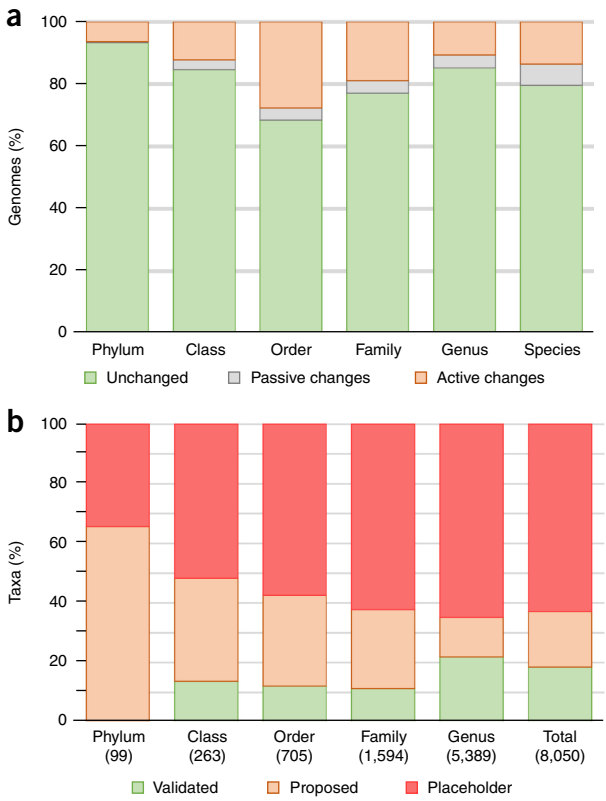

Figure 4 Comparison of GTDB and NCBI taxonomies and naming status of GTDB taxa. (**a**) Comparison of GTDB and NCBI taxonomic assignments across 84,634 bacterial genomes from RefSeq/GenBank release 80. For each rank, a taxon was classified as being unchanged if its name was identical in both taxonomies; passively changed if the GTDB taxonomy provided name information absent in the NCBI taxonomy; or actively changed if the name was different between the two taxonomies. Changes between the GTDB and NCBI taxonomies are fully listed in **Supplementary Table 3**. (**b**) Percentage of GTDB taxa at each rank that are validly published and approved; proposed but not validated; or nonstandard placeholder names. The number of taxa at each rank is shown in parentheses.

The genus *Clostridium* is widely acknowledged to be polyphyletic, and efforts have been made to rectify this problem, including a global attempt to reclassify the genus by using a combination of phylogenetic markers[9]. The authors of that study have proposed the reclassification of 78 *Clostridium* species, and nine other species, into six novel genera[9,52]. Of these, we confirmed that *Erysipelatoclostridium* (with the exception of *Clostridium innocuum* str. 2959), *Gottschalkia* and *Tyzzerella* (excepting *Clostridium nexile* CAG:348) represent monophyletic genus-level groups. The remaining three genera proposed by Yutin and Galperin[7] represent multiple genera in the GTDB taxonomy, including genera with validly published names (**Supplementary Table 11**). This result is consistent with recent analyses of individual taxa in these groups[53,54]. The GTDB taxonomy is also largely in agreement at the genus level with a recent global genome-based classification of the Bacteroidetes[55]. Of the 122 genera addressed in that study, six were found to be in need of reclassification; *Chryseobacterium*, *Epilithonimonas*, *Aequorivita, Vitellibacter*, *Flexibacter* and *Pedobacter*. All six were similarly identified as polyphyletic in the GTDB taxonomy and reclassified accordingly. These findings demonstrate that our

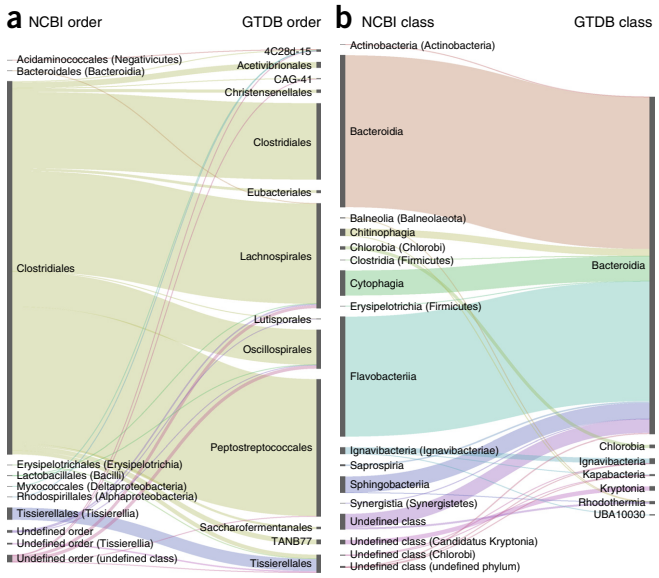

Figure 5 Comparisons of NCBI and GTDB classifications of genomes designated as Clostridia or Bacteroidetes in the GTDB taxonomy. (**a**) Comparison of NCBI (left) and GTDB (right) order-level classifications of the 2,368 bacterial genomes assigned to the class Clostridia in the GTDB taxonomy. Genomes classified in a class other than Clostridia by NCBI are indicated in parentheses. (**b**) Comparison of NCBI and GTDB class-level classifications of the 2,058 bacterial genomes assigned to the phylum Bacteroidetes in the GTDB taxonomy. Genomes classified in a phylum other than the Bacteroidetes by NCBI are indicated in parentheses.

methods are broadly consistent with rigorous independent analyses of problematic genera and species.

Taxonomic changes at higher ranks

A number of notable taxonomic changes at higher ranks are proposed for well-studied groups. For example, the class Betaproteobacteria was reclassified as an order within the class Gammaproteobacteria because it is entirely circumscribed within the latter group and is closer to the median RED value for an order than a class (**Fig. 2a**). This change is consistent with the original 16S rRNA gene topology of the Proteobacteria and subsequent trees[6,8,56], although such a rank change has not been proposed in these studies. The Deltaproteobacteria and Epsilonproteobacteria were removed entirely from the Proteobacteria, because this phylum is not consistently recovered as a monophyletic unit, as found in many previous 16S rRNA and other marker gene analyses[11,57,58]. In the case of the Epsilonproteobacteria, this class was combined with the order Desulfurellales (Deltaproteobacteria) to form a new phylum[58].

The Firmicutes also underwent extensive internal reclassification. As a clade, this phylum is typically monophyletic but poorly supported in most trees (**Supplementary Table 1**), and it has a RED in the phylum range, albeit to the left of the median for this taxonomic rank (**Fig. 2b**). The Firmicutes were therefore retained as a phylum-level lineage, although future revision of this status may be warranted. This phylum was divided into 34 classes including the mycoplasmas, which are currently classified as a separate phylum, the Tenericutes[59] and 14 classes exclusively comprising MAGs. Incorporation of the Tenericutes within the Firmicutes is consistent with single-gene phylogenies[6,8,32,53] and is further supported by recent evidence based on multiple molecular markers[25,26,60]. Similarly to its type genus, the order Clostridiales was extensively subdivided (**Fig. 5a**), largely as a consequence of an anomalous RED for this rank (**Fig. 2a**).

On the basis of robust monophyly, taxonomic rank normalization and naming priority in the literature, the phylum Bacteroidetes is proposed to encompass the Chlorobi and Ignavibacteriae as class-level lineages. Concomitantly, several former classes of Bacteroidetes were amalgamated into the class Bacteroidia as order-level lineages, including the Chitinophagales, Cytophagales, Flavobacteriales and Sphingobacteriales (**Fig. 5b**). These proposed changes are in contrast to recent reclassifications, in which Bacteroidetes is divided into three major lineages by promoting the families Rhodothermaceae and Balneolaceae to phyla[55,61] (**Fig. 2a**). In the GTDB taxonomy, these were retained as families within their own orders in the class Rhodothermia, according to their RED values (**Fig. 2b**). The higher-level taxonomy of the phylum Actinobacteria was largely unchanged. The five classes Actinobacteria, Acidimicrobiia, Coriobacteriia, Thermoleophilia and Rubrobacteria were retained, and the sole change at the class level was the downgrading of the Nitriliruptoria to an order within the class Actinobacteria according to rank normalization. Changes to other major lineages are summarized in **Supplementary Table 3**.

Rank normalization of uncultured microbial diversity

Having normalized the taxonomy on existing isolate-based classifications, we were able to calibrate the taxonomic ranks of uncultured lineages. Candidate phylum KSB3 was initially proposed on the basis of comparative analysis of environmental 16S rRNA gene sequences[62,63], and more recently two near-complete MAGs belonging to this phylum have been reconstructed from a bulking sludge metagenome, for which the names '*Candidatus* Moduliflexus flocculans' and '*Candidatus* Vecturathrix granuli' have been proposed[64]. These genomes were further classified into separate families, orders and classes within the phylum; however, by rank normalization, they represent separate genera belonging to a single family. The group still retains a phylum-level status, because it is not reproducibly affiliated with other bacterial lineages[36]; however, we propose that the phylum (Modulibacteria) is currently genomically represented by a single class (Moduliflexia), single order (Moduliflexales) and single family (Moduliflexaceae; **Fig. 2b**).

As part of a single-cell-genomics study, the superphylum Patescibacteria has been proposed to encompass the candidate phyla Parcubacteria (OD1), Microgenomates (OP11) and Gracilibacteria (GN02)[57]. These candidate phyla have been further subsumed within the Candidate Phyla Radiation (CPR) on the basis of the addition of 797 MAGs[20]. Currently, there are at least 65 candidate phyla proposed to belong to the CPR[20,21], and the justification of individual phyla has been based primarily on a 16S rRNA sequence-identity threshold of 75% (ref. 11). The CPR has been consistently recovered as a monophyletic group by using concatenated protein markers in this and previous studies[20,22,25]. However, rank normalization suggests that the CPR should be reclassified as a single phylum, for which we suggest reimplementing the name Patescibacteria (**Fig. 2b**), although ultimately the group should be named according to the nomenclature type material[65].

DISCUSSION

We present the GTDB taxonomy, which aims to provide an objective, phylogenetically consistent classification of bacterial species. We show that this taxonomy is largely congruent with the topology and substitution rates of phylogenies inferred by using different marker sets and maximum-likelihood-inference methods. Although we preserved

existing taxonomic classifications when possible, a substantial number of modifications were required to resolve polyphyletic groups and to normalize taxa at each taxonomic rank on the basis of our operational approximation of relative time of divergence.

The GTDB taxonomy covers 94,759 bacterial genomes, but we expect the number of available reference genomes to expand rapidly and to encompass new lineages[21,22]. In anticipation of this expansion, we will curate the taxonomy biannually to incorporate new genomes and proposed taxonomic groups, while retaining a phylogenetically consistent classification. Subsampling of the bac120 data set suggests that subsets of these marker genes could be used in the future to produce reliable phylogenies that better scale with the projected increase in the reference-genome database[2]. Some incongruencies between genome trees inferred for each biannual update are expected to affect the GTDB taxonomy, as has already been observed for well-established groups such as the Firmicutes, which may require reclassification in subsequent iterations. A small number of GTDB taxa were also not recovered as monophyletic groups under trees inferred with different inference methods or marker sets. Such regions of instability should be addressed individually with more in-depth analyses to establish the most suitable classification, as for example, has been done recently with the class Epsilonproteobacteria[58].

The GTDB taxonomy is available through the Genome Taxonomy Database website (http://gtdb.ecogenomic.org/), and we are facilitating its incorporation into other public bioinformatic resources. We are also developing a standalone tool, GTDB-Tk (https://github.com/Ecogenomics/GtdbTk/), to enable researchers to classify their own genomes according to the GTDB taxonomy and its classification criteria. The methods reported here are applicable to any taxonomically annotated phylogenetic tree, and we are in the process of expanding the GTDB to include Archaea and double-stranded DNA viruses. We anticipate that the availability of an up-to-date normalized genome-based classification should greatly facilitate the analysis of microbial genome data and communication of scientific results.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Yilmaz for helpful discussions on the proposed genome-based taxonomy; QFAB Bioinformatics for providing computational resources; and members of ACE for beta-testing GTDB. The project was primarily supported by an Australian Research Council Laureate Fellowship (FL150100038) awarded to P.H.

AUTHOR CONTRIBUTIONS

D.H.P., D.W.W. and P.H. wrote the paper, and all other authors provided constructive suggestions. D.H.P. and P.H. designed the study. M.C. and P.H. performed the taxonomic curation. D.H.P., D.W.W., C.R., A.S., and P.-A.C. performed the bioinformatic analyses. P.-A.C. designed the website.

COMPETING INTERESTS

The authors declare no competing interests.

1. Garrity, G.M. A new genomics-driven taxonomy of Bacteria and Archaea: are we there yet? *J. Clin. Microbiol.* **54**, 1956–1963 (2016).
2. Hugenholtz, P., Sharshewski, A. & Parks, D.H. Genome-based microbial taxonomy coming of age. in *Microbial Evolution* (ed. Ochman, H.) 55–65 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2016).
3. Yoon, S.H. *et al.* Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617 (2017).
4. Godfray, H.C.J. Challenges for taxonomy. *Nature* **417**, 17–19 (2002).
5. Federhen, S. The NCBI taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).
6. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
7. Cole, J.R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2014).
8. 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).
9. Yutin, N. & Galperin, M.Y. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ. Microbiol.* **15**, 2631–2641 (2013).
10. Beiko, R.G. Microbial malaise: how can we classify the microbiome? *Trends Microbiol.* **23**, 671–679 (2015).
11. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).
12. Abbott, S.L. & Janda, J.M. in *The Prokaryotes* 3rd edn. (eds. Dworkin, M. *et al.*) 72–89 (Springer, New York, 2006).
13. Jumas-Bilak, E., Roudière, L. & Marchandin, H. Description of *'Synergistetes'* phyl. nov. and emended description of the phylum *'Deferribacteres'* and of the family *Syntrophomonadaceae*, phylum *'Firmicutes'*. *Int. J. Syst. Evol. Microbiol.* **59**, 1028–1035 (2009).
14. Janda, J.M. & Abbott, S.L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**, 2761–2764 (2007).
15. Schulz, F. *et al.* Towards a balanced view of the bacterial tree of life. *Microbiome* **5**, 140 (2017).
16. DeSantis, T.Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
17. Brochier, C., Forterre, P. & Gribaldo, S. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol. Biol.* **5**, 36 (2005).
18. Ciccarelli, F.D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
19. Thiergart, T., Landan, G. & Martin, W.F. Concatenated alignments and the case of the disappearing tree. *BMC Evol. Biol.* **14**, 266 (2014).
20. Brown, C.T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
21. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
22. Parks, D.H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
23. Bapteste, E. *et al.* Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.* **5**, 33 (2005).
24. Tonini, J., Moore, A., Stern, D., Shcheglovitova, M. & Ortí, G. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLoS Curr.* https://doi.org/10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be (2015).
25. Hug, L.A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
26. Lang, J.M., Darling, A.E. & Eisen, J.A. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS One* **8**, e62510 (2013).
27. Dupont, C.L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012).
28. Wu, D., Jospin, G. & Eisen, J.A. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS One* **8**, e77033 (2013).
29. Giovannoni, S.J., Rappé, M.S., Vergin, K.L. & Adair, N.L. 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *Proc. Natl. Acad. Sci. USA* **93**, 7979–7984 (1996).
30. Dojka, M.A., Hugenholtz, P., Haack, S.K. & Pace, N.R. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.* **64**, 3869–3877 (1998).
31. Zwart, G. *et al.* Rapid screening for freshwater bacterial groups by using reverse line blot hybridization. *Appl. Environ. Microbiol.* **69**, 5875–5883 (2003).
32. Wolf, M., Müller, T., Dandekar, T. & Pollack, J.D. Phylogeny of *Firmicutes* with special reference to *Mycoplasma* (*Mollicutes*) as inferred from phosphoglycerate kinase amino acid sequence data. *Int. J. Syst. Evol. Microbiol.* **54**, 871–875 (2004).
33. Lonergan, D.J. *et al.* Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *J. Bacteriol.* **178**, 2402–2408 (1996).
34. Beiko, R.G. Telling the whole story in a 10,000-genome world. *Biol. Direct* **6**, 34 (2011).
35. Zhang, Y. & Sievert, S.M. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in *Epsilonproteobacteria*. *Front. Microbiol.* **5**, 110 (2014).
36. Hugenholtz, P., Pitulle, C., Hershberger, K.L. & Pace, N.R. Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* **180**, 366–376 (1998).

37. Konstantinidis, K.T. & Tiedje, J.M. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187**, 6258–6264 (2005).
38. Wu, D., Doroud, L. & Eisen, J.A. TreeOTU: operational taxonomic unit classification based on phylogenetic trees. Preprint at https://arxiv.org/abs/1308.6333 (2013).
39. Maniloff, J. in *Molecular Biology and Pathogenicity of Mycoplasma* (eds. Razin, S. & Herrmann, R.) 31–43 (Springer, New York, 2002).
40. Kumar, S., Stecher, G., Suleski, M. & Hedges, S.B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
41. Marin, J., Battistuzzi, F.U., Brown, A.C. & Hedges, S.B. The timetree of prokaryotes: new insights into their evolution and speciation. *Mol. Biol. Evol.* **34**, 437–446 (2017).
42. Gadagkar, S.R., Rosenberg, M.S. & Kumar, S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zoolog. B Mol. Dev. Evol.* **304**, 64–74 (2005).
43. Balvočiūtė, M. & Huson, D.H. SILVA, RDP, Greengenes, NCBI and OTT: how do these taxonomies compare? *BMC Genomics* **18** (Suppl. 2), 114 (2017).
44. Whitman, W.B. Modest proposals to expand the type material for naming of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **66**, 2108–2112 (2016).
45. Konstantinidis, K.T., Rosselló-Móra, R. & Amann, R. Uncultivated microbes in need of their own taxonomy. *ISME J.* **11**, 2399–2406 (2017).
46. Comas, I., Homolka, S., Niemann, S. & Gagneux, S. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* **4**, e7815 (2009).
47. Martiny, J.B.H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
48. Trost, B., Haakensen, M., Pittet, V., Ziola, B. & Kusalik, A. Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera. *BMC Microbiol.* **10**, 258 (2010).
49. Beaz-Hidalgo, R., Hossain, M.J., Liles, M.R. & Figueras, M.J. Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for *aeromonas* genomes in the GenBank database. *PLoS One* **10**, e0115813 (2015).
50. Kook, J.K. *et al.* Genome-based reclassification of *Fusobacterium nucleatum* subspecies at the species level. *Curr. Microbiol.* **74**, 1137–1147 (2017).
51. Bobay, L.M. & Ochman, H. Biological species are universal across life's domains. *Genome Biol. Evol.* **9**, 491–501 (2017).
52. Galperin, M.Y., Brover, V., Tolstoy, I. & Yutin, N. Phylogenomic analysis of the family *Peptostreptococcaceae* (Clostridium cluster XI) and proposal for reclassification of *Clostridium litorale* (Fendrich *et al.* 1991) and *Eubacterium acidaminophilum* (Zindel *et al.* 1989) as *Peptoclostridium litorale* gen. nov. comb. nov. and *Peptoclostridium acidaminophilum* comb. nov. *Int. J. Syst. Evol. Microbiol.* **66**, 5506–5513 (2016).
53. Yarza, P. *et al.* The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* **31**, 241–250 (2008).
54. Sakamoto, M., Iino, T. & Ohkuma, M. *Faecalimonas umbilicata* gen. nov., sp. nov., isolated from human faeces, and reclassification of *Eubacterium contortum, Eubacterium fissicatena* and *Clostridium oroticum* as *Faecalicatena contorta* gen. nov., comb. nov., *Faecalicatena fissicatena* comb. nov. and *Faecalicatena orotica* comb. nov. *Int. J. Syst. Evol. Microbiol.* **67**, 1219–1227 (2017).
55. Hahnke, R.L. *et al.* Genome-based taxonomic classification of *Bacteroidetes. Front. Microbiol.* **7**, 2003 (2016).
56. Garrity, G.M., Bell, J.A. & Lilburn, T. in *Bergey's Manual of Systematic Bacteriology* (eds. Garrity, G. *et al.*) 575–922 (Springer, New York, 2005).
57. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
58. Waite, D.W. *et al.* Comparative genomic analysis of the class *Epsilonproteobacteria* and proposed reclassification to Epsilonbacteraeota (phyl. nov.). *Front. Microbiol.* **8**, 682 (2017).
59. Brown, D.R. in *Bergey's Manual of Systematic Bacteriology* (eds. Krieg, N.R. *et al.*) 567–724 (Springer, New York, 2010).
60. Skennerton, C.T. *et al.* Phylogenomic analysis of *Candidatus* 'Izimaplasma' species: free-living representatives from a *Tenericutes* clade found in methane seeps. *ISME J.* **10**, 2679–2692 (2016).
61. Munoz, R., Rosselló-Móra, R. & Amann, R. Revised phylogeny of Bacteroidetes and proposal of sixteen new taxa and two new combinations including Rhodothermaeota phyl. nov. *Syst. Appl. Microbiol.* **39**, 281–296 (2016).
62. Tanner, M.A., Everett, C.L., Coleman, W.J. & Yang, M.M. Complex microbial communities inhabiting sulfide-rich black mud from marine coastal environments. *Biotechnol. Alia* **8**, 1–16 (2000).
63. Yamada, T. *et al.* Characterization of filamentous bacteria, belonging to candidate phylum KSB3, that are associated with bulking in methanogenic granular sludges. *ISME J.* **1**, 246–255 (2007).
64. Sekiguchi, Y. *et al.* First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. *PeerJ* **3**, e740 (2015).
65. Chuvochina, M. *et al. Syst. Appl. Microbiol.* The importance of designating type material for uncultured taxa https://doi.org/10.1016/j.syapm.2018.07.003 (2018).

ONLINE METHODS

Genome data set. A data set of 87,106 bacterial genomes was obtained from RefSeq/GenBank[66] release 80 on 17 January 2017. An additional 11,603 MAGs obtained from Sequence Read Archive metagenomes[67] were added to this data set to improve coverage of uncultured lineages, most of which have been reported previously[22]. These genomes were dereplicated as described in Parks *et al.*[22] with the exception that dereplication was based on ANI values estimated on the basis of Mash distances[68] instead of pairwise AAI values calculated from the bac120 alignment. Specifically, a genome was assigned to a representative genome if one of the following criteria was met: (i) the Mash distance to the representative genome was ≤0.035 (~ANI of 96.5%), and the query genomes had no species assignment in the previous iteration of the GTDB taxonomy (release 78; http://gtdb.ecogenomic.org/downloads); (ii) the Mash distance was ≤0.05 (~ANI of 95%), and the query and representative genomes had the same species assignment in the previous iteration of the GTDB; or (iii) the Mash distance was ≤0.1 (~ANI of 90%), the query and representative genomes had the same species assignment in the previous iteration of the GTDB and under the NCBI taxonomy. After dereplication, genomes were excluded that had (i) amino acids in <50% of the columns within the bac120 alignment and/or (ii) an estimated quality <50, defined as completeness – 5× contamination and calculated with the default lineage-specific marker gene sets of CheckM[69].

Metadata. The NCBI taxonomy[5] associated with the reference genomes was obtained from the NCBI Taxonomy FTP site on January 17, 2017. This taxonomy was standardized to seven ranks (domain to species) by removing nonstandard ranks and identifying missing standard ranks with rank prefixes. Standard ranks were also prefixed with rank identifiers as previously described[8]. For example, the full NCBI lineage for *'Nostoc azollae' 0708* (GCF_000196515.1) at the time of download was 'cellular organisms (no rank); Bacteria (superkingdom); Terrabacteria group (no rank); Cyanobacteria/Melainabacteria group (no rank); Cyanobacteria (phylum); Nostocales (order); Nostocaceae (family); Trichormus (genus); Trichormus azollae (species); 'Nostoc azollae' 0708 (strain)', which was standardized to 'd__Bacteria; p__Cyanobacteria; c__; o__Nostocales; f__Nostocaceae; g__Trichormus; s__Trichormus azollae'. Additional metadata from NCBI such as 'Isolate', 'Assembly level' and 'Genome representation' were also parsed from the assembly reports of all bacterial genomes (http://gtdb.ecogenomic.org/downloads) to provide information for manual tree curation. To complement the NCBI taxonomy and metadata, we identified 16S rRNA gene sequences from all NCBI genomes and MAGs by using HMMER[70] v3.1b1 and the most similar sequence within the Greengenes[8] 2013 and SILVA[6] v123.1 databases identified with BLASTN v2.2.30+[71].

Inference and annotation of the bac120 tree. A phylogenetic tree spanning the dereplicated genomes was inferred from the concatenation of 120 ubiquitous single-copy marker genes (bac120 marker set) identified within the Pfam[72] v27 and TIGRFAMs[73] v15.0 databases, which had previously been evaluated to be phylogenetically informative[22]. Gene calling was performed with Prodigal[74] v2.6.3, and the 120 marker proteins were identified and aligned with HMMER v3.1b1. The resulting MSA was trimmed by removal of columns represented by <50% of genomes and/or with an amino acid consensus <25%. In addition, genomes with amino acids in <50% of columns were removed before phylogenetic inference. The bac120 reference tree was inferred with FastTree[75] v2.1.7 under the WAG model[76] of protein evolution with gamma-distributed rate heterogeneity[77] (+GAMMA). Branch support was estimated by performing 100 nonparametric bootstrap replicates. Group names based on the standardized NCBI genome taxonomy were added to interior nodes of the bac120 tree with tax2tree[8].

Calculating relative evolutionary divergence and thresholds for taxonomic ranks. RED values were calculated from the annotated bac120 tree with PhyloRank (v0.0.27; https://github.com/dparks1134/PhyloRank/). PhyloRank performs a preorder tree traversal with the RED of the root defined to be zero and the RED of node n defined as $p + (d/u)(1 - p)$, where p is the RED of the parent node, d is the branch length to the parent node, and u is the average branch length from the parent node to all extant taxa descendant from n. Because the RED of taxa is influenced by root placement, and the rooting of the bacterial tree remains controversial[78], we took an operational approach and rooted trees at the midpoint of all branches leading to phyla with two or more classes. The RED of a taxon was then taken as the median RED over all tree rootings, excluding the tree in which the taxon was part of the outgroup. Median RED values for each taxonomic rank were determined from taxa with two or more immediately subordinate taxa (for example, phyla with two or more defined classes) and the RED rank intervals used to guide the GTDB taxonomy defined as ±0.1 from these median RED values.

Tree-based taxonomic curation. The annotated bac120 tree was manually curated in ARB[79] to (i) resolve polyphyletic groups, (ii) correct taxa falling outside of their RED distribution and (iii) add 16S rRNA-based group names to uncultured lineages. Branch lengths in the bac120 tree were replaced with their corresponding RED values to produce a 'scaled' tree as a visual aid in the taxonomic-rank normalization process (**Fig. 1c**). Polyphyletic groups were identified as part of the initial annotation of group names with numerical suffixes generated by tax2tree. Groups containing type material according to the List of Prokaryotic Names with Standing in Nomenclature[80] (LPSN) kept the original unsuffixed name to indicate the validity of name assignment, and other groups were renamed according to a set of nomenclatural rules (described below). Outlier-group names (±0.1 from the median RED values) were moved into their rank distributions in one of two ways: (i) the name was moved to another interior node in the bac120 tree, or (ii) the name was left on the original interior node but reclassified to a different rank. 16S rRNA taxonomy-based names were assigned to clades in the bac120 tree if one or more genomes spanning the clade had ≥95% identity over ≥500 bp to a reference 16S rRNA sequence with a given name. Robust interior nodes (bootstrap support >90%) were given preference for name assignments.

Generation of final GTDB taxonomy. The GTDB taxonomy was extracted from the curated bac120 tree (Newick input format) by concatenating group names from the relevant interior nodes for each genome and exporting them as a flat text file for validation. Validation included checks for correct number and order of taxonomic ranks; presence of multiple parents (polyphyly); orthographic and semantic errors; and consistency of order (-ales) and family (-aceae) rank suffixes. Because names can be applied only to groups of two or more taxa in ARB, 'singleton' genomes often have incomplete taxonomic lineages in the exported flat text file. These were autocompleted to at least the level of genus on the basis of the nomenclatural rules outlined below. The consistency of filled ranks between releases was tracked with additional scripts, and the completed taxonomy was validated once more.

Nomenclatural rules for standard names. Nomenclatural changes of validly published names were made according to the International Code of Nomenclature of Prokaryotes[81]. In the event of the nomenclature type being excluded from or not present in the group, a new type was designated on the basis of priority in the literature, and provisional higher-rank names were established with the addition of corresponding rank suffixes to the stem of the generic name, including the recently proposed standard suffix -aeota for the rank of phylum[82]. Priority was established for all other taxa names, namely those without standing in nomenclature and *Candidatus* taxa, on the basis of the earliest published taxon in the group, and ranks with missing annotations derived their name from the corresponding generic name of the earliest named taxon.

The term *Candidatus* was removed from GTDB taxon names to standardize the taxonomy but is easily tracked via the NCBI Organism Name in the genome metadata (http://gtdb.ecogenomic.org/). In cases in which new names were not proposed to resolve polyphyly, notably for the rank of genus, alphabetical suffixes were added to the standard name (for example, *Bacillus*_A, *Bacillus*_B and so forth). Species-level groups with nonstandard or ambiguous names were designated as 'genus name' sp1, 'genus name' sp2 and so forth. The official naming hierarchy from lower to higher ranks was followed, with the exception of some provisional phylum names lacking named species (notably CPR phyla), which were retained after rank normalization with appropriate rank suffix changes, for example, o__Levybacterales from '*Candidatus* Levybacteria'[20].

doi:10.1038/nbt.4229

Nomenclatural rules for nonstandard placeholder names. Nonstandard placeholder names were given to groups lacking standardly named representatives. Several sources were used to derive nonstandard names; (i) 16S rRNA environmental clone names grafted onto the bac120 tree (described above; **Supplementary Table 2**), (ii) isolate strain names, for example, g__Mor1 from the genome 'Acidobacteria bacterium Mor1' (GCA_001664505.1), (iii) MAG names, for example, g__UBA4820 from 'SRA genome UBA4820' (GCA_002402325.1), and (iv) genome assembly identifiers for groups exclusively comprising complex symbiont names, for example, g__GCF-001602625 for '*Sodalis*-like endosymbiont of *Proechinophthirus fluctus*' (GCF_001602625.1). Nonstandard names longer than 15 characters were trimmed for brevity and to minimize spelling errors, for example, g__2-02-FULL-67-57 from the name 'Acidobacteria bacterium RIFCSPHIGHO2_02_FULL_67_57' (GCA_001766975.1). In the rare event of identical placeholder names for two or more phylogenetically distinct groups resulting from automated name trimming, or rank filling, we appended hyphenated alphabetical suffixes (-A, -B and so forth) to distinguish them. As with standard binomial names, species-level groups were defined as 'genus name' sp1, 'genus name' sp2 and so forth. When necessary, nonstandard names were propagated to higher ranks differentiated only by rank prefix, for example, d__Bacteria; p__Acidobacteria; c__UBA4820; o__UBA4820; f__UBA4820; g__UBA4820.

Inference of trees used to validate the GTDB taxonomy. The stability of the GTDB taxonomy was evaluated with trees inferred in a manner analogous to that described for the bac120 tree. Briefly, proteins were called with Prodigal; marker genes were identified and aligned with HMMER with Pfam and TIGRfam HMMs; MSAs were trimmed according to consistency and ubiquity; genomes with poor representation in the alignment were removed before phylogenetic inference; and trees were inferred with FastTree under the WAG+GAMMA models unless otherwise specified.

Alternative inference methods. A neighbor-joining tree was inferred with NINJA[83] v1.2.2 with default parameters. Maximum-likelihood trees were inferred with ExaML[84] v3.0.20 under the JTT+PSR (-m PSR and -D flags) models and IQ-TREE[85] v1.5.5 under the WAG+GAMMA models. ExaML requires a starting tree, and RaxML[86] v8.1.11 was used to create a parsimony tree for this purpose. To reduce computational requirements, the trees were inferred over a reduced set of 4,985 or 10,462 genomes dereplicated to one genome per GTDB genus or species, respectively. This dereplication was performed by preferentially selecting type strains, genomes with good assembly statistics and genomes estimated to be highly complete with minimal contamination from the 21,943 genomes used to define the GTDB. The original MSA of 34,744 columns was also subsampled to 5,038 columns by evenly sampling columns with ≤10% gaps and ≤95% identical amino acids from each of the 120 bacterial marker genes. Branch support for the neighbor-joining and IQ-TREE trees were determined with 100 nonparametric bootstrap replicates, whereas the ExaML tree was limited to 30 replicates, owing to the high computational requirements of this method.

Alternative marker sets. A ribosomal protein tree (rp1) was inferred from the concatenation of 16 ribosomal proteins[20,22] and consisted of 1,949 aligned columns after trimming of 101 columns represented by <50% of the genomes and 11 columns with an amino acid consensus <25%. The rp1 tree spanned 21,444 genomes after removal of 1,967 genomes with amino acids in <50% of the filtered columns. Subsampled marker trees were inferred by random selection of 60 genes from the bac120 marker set. A total of 100 replicate trees were generated in this fashion to assess the effects of using different subsets of the bac120 marker set. The filtered MSAs ranged in length from 15,010 to 20,061 amino acids, and all trees spanned 21,943 genomes, because no additional genome filtering was performed. Individual gene trees were also inferred for each of the genes composing the bac120 marker set. The filtered alignments for these trees ranged in length from 43 to 1,069 amino acids and spanned 16,932 to 21,050 genomes.

Alternative genome sets. Trees were inferred from sets of extent taxa subsampled to one genome per genus in the GTDB taxonomy. Representative genomes for each genus were randomly selected from the 21,943 dereplicated genomes composing the bac120 tree. The alignments used for the full bac120 tree were used to infer the 100 subsampled trees.

Alternative models. A bac120 tree under the LG model[87] was inferred with FastTree v2.1.7 compiled for double precision to avoid numerically unstable issues. This tree was inferred with the same MSA used for the bac120 tree under the WAG+GAMMA models, and it spans the same set of genomes.

Inference of 16S rRNA gene tree. A 16S rRNA gene tree was inferred from genes >1,200 bp identified within the 21,943 dereplicated and quality-controlled genomes. The 16S rRNA genes were identified with HMMER and domain-specific SSU/LSU HMM models, as implemented in the 'ssu-finder' method of CheckM, and the longest gene was retained for genomes with multiple copies of the 16S rRNA gene. The 12,712 identified 16S rRNA genes were filtered to remove sequences potentially representing contamination with a reciprocal BLAST protocol. Genes were searched against each other with BLASTN, and a gene was removed from consideration if its closest match belonged to a genome classified in a different taxonomic order, as defined by the GTDB, the gene had an alignment length ≥800 bp, and the gene had a percentage identity ≥82%. The percentage-identity criterion was based on the thresholds proposed by Yarza *et al.*[11]. This procedure resulted in 277 sequences being removed from consideration (**Supplementary Table 12**). The remaining 12,435 16S rRNA genes were aligned with ssu-align[88] v0.1, and trailing or leading columns represented by ≤70% of the sequences were trimmed, thus resulting in an alignment of 1,409 bp. The gene tree was inferred with FastTree v2.1.7 under the GTR[89] and GAMMA models.

Pairwise comparison of inferred trees. The similarity of inferred trees was determined with the normalized Robinson–Foulds[90] distance, which yields values between 0 (identical tree topologies) and 1 (trees with no splits in common). We also considered the proportion of supported splits (branches) in common between two trees. A split was considered supported if it had a nonparametric bootstrap support value ≥70%. The similarity of two trees was determined by dividing the weight of all supported splits in common between the two trees by the total weight of all supported splits. This procedure yielded a value between 0 (no supported splits in common) and 1 (all supported splits in common). Pairwise distances were visualized as an Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical-cluster tree.

Assessing stability of the NCBI and GTDB taxonomy on inferred trees. The congruency of the NCBI and GTDB taxonomies on different trees was assessed by placing each taxon on the node with the highest resulting *F* measure. The *F* measure is the harmonic mean of precision and recall, and it has been previously proposed for decorating trees with a donor taxonomy[8]. Taxonomic stability was assessed as both the percentage of taxa identified as being monophyletic, operationally monophyletic or polyphyletic within a tree and the percentage of genomes in the tree with identical, unresolved or conflicting taxonomic assignments relative to the NCBI or GTDB taxonomies. Because a few incongruent genomes (resulting from phylogenetic incongruence, phylogenetic instability, chimeric artifacts or erroneous NCBI taxonomic assignments) are sufficient to cause a large number of polyphyletic taxa, we classified taxa with an *F* measure ≥0.95 as operationally monophyletic. The results were restricted to taxa containing two or more genomes, because taxa represented by a single genome are guaranteed to be monophyletic in a tree. Genomes in a tree with incongruent taxonomic assignments were classified as either (i) conflicting if the genome was assigned to a different taxon or (ii) unresolved if the taxon had no taxonomic label or multiple taxonomic labels at a specific taxonomic rank, one of which was the expected label (for example, a polyphyletic lineage spanning two or more genera, one of which is the expected genus for the genome).

Comparison of GTDB to the NCBI and SILVA taxonomies. GTDB and NCBI taxonomic assignments were compared across the 84,634 bacterial genomes comprising RefSeq/GenBank release 80. Assigned names at each taxonomic rank were classified as 'unchanged', 'passive' change or 'active' change. A taxon was classified as unchanged if its name was identical in both taxonomies; passively changed if the GTDB taxonomy provided name information absent in the NCBI taxonomy; or actively changed if the name was different between the two taxonomies. Because the GTDB taxonomy does not qualify taxon names as '*Candidatus*' or 'candidate division', differences due solely to these designations were classified as unchanged. Taxa modified with alphabetical suffices to resolve identified polyphyly were treated as active changes, because such taxa should eventually be assigned new designations.

Comparison to SILVA was performed in a similar fashion. The 12,435 16S rRNA genes identified within the 21,943 dereplicated genomes were associated

with the SILVA v128 taxonomy according to sequence similarity with BLASTN. To ensure an accurate comparison, only the 11,178 16S rRNA genes with a match in SILVA of >99% identity and >95% alignment length were considered. SILVA does not specify a seven-rank taxonomy for all genes, and the 399 genes with fewer than seven specified taxa were removed from consideration to ensure that taxa at corresponding ranks were being compared. Differences due solely to a '*Candidatus*', 'candidate division', 'uncultivated candidate division', '*sensu stricto*', 'clade', 'cluster', 'sp.' or 'marine group' designations were classified as unchanged. Taxa containing any of the following designations were treated as missing taxonomic information and classified as a passive change: unknown, unidentified, uncultured, bacterium, metagenome, lineage, class, order, family, genus, subgroup, group, subsection, surface, env or *incertae sedis*. Taxa in SILVA with numerical suffixes used to designate polyphyly were considered unchanged if they matched a corresponding GTDB taxon with a different suffix also indicating polyphyly (for example, 'Spirochaeta 2' was considered unchanged with 'Spirochaeta_A').

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

Code availability. RED values were calculated with PhyloRank v0.0.27 (https://github.com/dparks1134/PhyloRank/), which is freely available under the GNU General Public License v3.0.

Data availability. Data files for the GTDB taxonomy are available at http://gtdb.ecogenomic.org/ and include: (i) flat file with the GTDB taxonomy defined for 94,759 genomes; (ii) bootstrapped bac120 tree in Newick format spanning the 21,943 dereplicated genomes and annotated with the GTDB taxonomy; (iii) FASTA files for each marker gene and the trimmed concatenated alignment; (iv) metadata for all genomes including NCBI, SILVA and Greengenes taxonomies, completeness and contamination estimates, assembly statistics (for example, N50) and genomic properties (for example, GC content and genome size); (v) FASTA file of 16S rRNA gene sequences identified within the 21,943 dereplicated genomes; and (vi) ARB database containing the bac120 tree. The 3,087 MAGs introduced in this study are available under BioProject PRJNA417962 and on the GTDB website.

66. Haft, D.H. *et al*. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
67. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
68. Ondov, B.D. *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
69. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
70. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
71. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
72. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
73. Haft, D.H., Selengut, J.D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
74. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
75. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
76. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
77. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
78. Williams, T.A. *et al.* New substitution models for rooting phylogenetic trees. *Phil. Trans. R. Soc. Lond. B* **370**, 20140336 (2015).
79. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
80. Euzéby, J.P. List of bacterial names with standing in nomenclature: a folder available on the internet. *Int. J. Syst. Bacteriol.* **47**, 590–592 (1997).
81. Parker, C.T., Tindall, B.J. & Garrity, G.M. International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* https://doi.org/10.1099/ijsem.0.000778 (2015).
82. Oren, A. *et al.* Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* **65**, 4284–4287 (2015).
83. Wheeler, T.J. in *Proceedings of the 9th Workshop on Algorithms in Bioinformatics* (eds. Salzberg, S.L. & Warnow, T.) 375–389 (Springer, Berlin, 2009).
84. Kozlov, A.M., Aberer, A.J. & Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577–2579 (2015).
85. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
86. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
87. Le, S.Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
88. Nawrocki, E.P. *Structural RNA Homology Search and Alignment Using Covariance Models* PhD thesis,Washington Univ. in Saint Louis, (2009).
89. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math Life Sci.* **17**, 57–86 (1986).
90. Kupczok, A., Schmidt, H.A. & von Haeseler, A. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms Mol. Biol.* **5**, 37 (2010).

doi:10.1038/nbt.4229

nature research

Corresponding author(s): Philip Hugenholtz

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Calculation of relative evolutionary divergence values was performed with custom software released under GPL v3.0 and made freely available as PhyloRank v0.0.27 at GitHub.
Data analysis	The following software was used: Prodigal v2.6.3, HMMER v3.1b1, FastTree v2.1.7, NINJA v1.2.2, ExaML v3.0.20, IQ-TREE v1.5.5, RaxML v8.1.11, CheckM v1.0.7, BLASTN v2.2.30+, ssu-align v0.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data files for the GTDB taxonomy are available at http://gtdb.ecogenomic.org and include: i) flat file with the GTDB taxonomy defined for 94,759 genomes; ii)

Nature Biotechnology: doi:10.1038/nbt.4229

bootstrapped bac120 tree in Newick format spanning the 21,943 dereplicated genomes and annotated with the GTDB taxonomy; iii) FASTA files for each marker gene and the trimmed concatenated alignment; iv) metadata for all genomes including NCBI, SILVA, and Greengenes taxonomies, completeness and contamination estimates, assembly statistics (e.g., N50), and genomic properties (e.g., GC-content, genomes size); v) FASTA file of 16S rRNA gene sequences identified within the 21,943 dereplicated genomes; and vi) an ARB database containing the bac120 tree. The 3,087 MAGs introduced in this study are available under BioProject PRJNA417962 and on the GTDB website.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We propose a standardized bacterial taxonomy based on a concatenated protein phylogeny that conservatively removes polyphyletic groups and normalizes taxonomic ranks based on relative evolutionary divergence.
Research sample	Reference genomes from NCBI along with additional metagenome-assembled genomes (MAGs) recovered as part of this study.
Sampling strategy	Genomes were screen for quality using standard estimates of completeness and contamination, and the dereplicated based on an estimate of average nucleotide identity and described species affiliation.
Data collection	Not applicable.
Timing and spatial scale	Not applicable.
Data exclusions	Poor quality genomes were excluded using a predefined set of criteria. Such genomes were excluded as they may negatively impact the quality of the inferred phylogeny.
Reproducibility	The robustness of the approach used to generate the proposed taxonomy was evaluated by varying tree inference software, evolutionary models, marker sets, and genome datasets.
Randomization	Not applicable.
Blinding	Not applicable.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Nature Biotechnology: doi:10.1038/nbt.4229