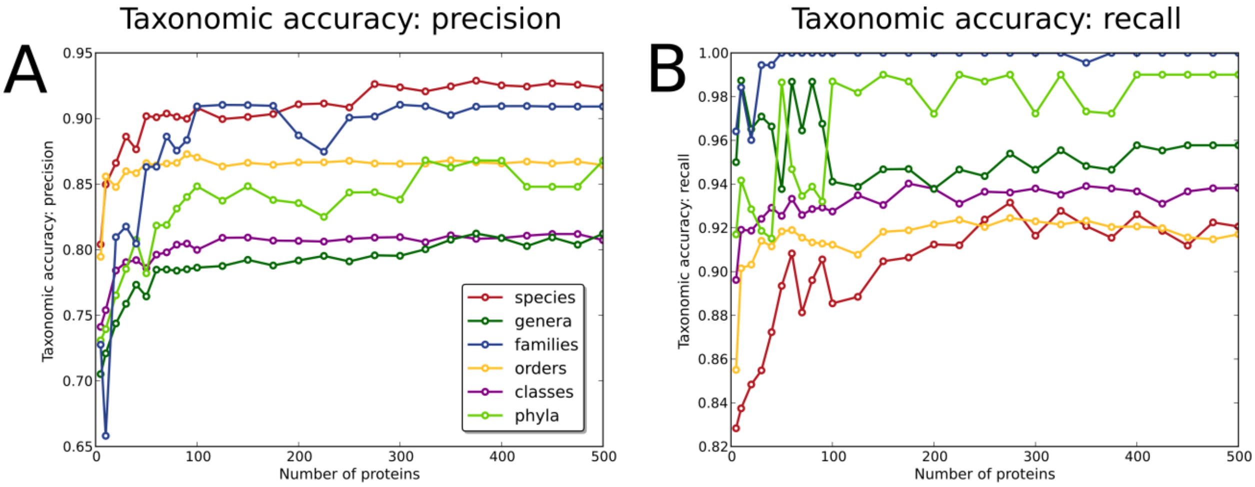
PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes

SUPPLEMENTARY INFORMATION

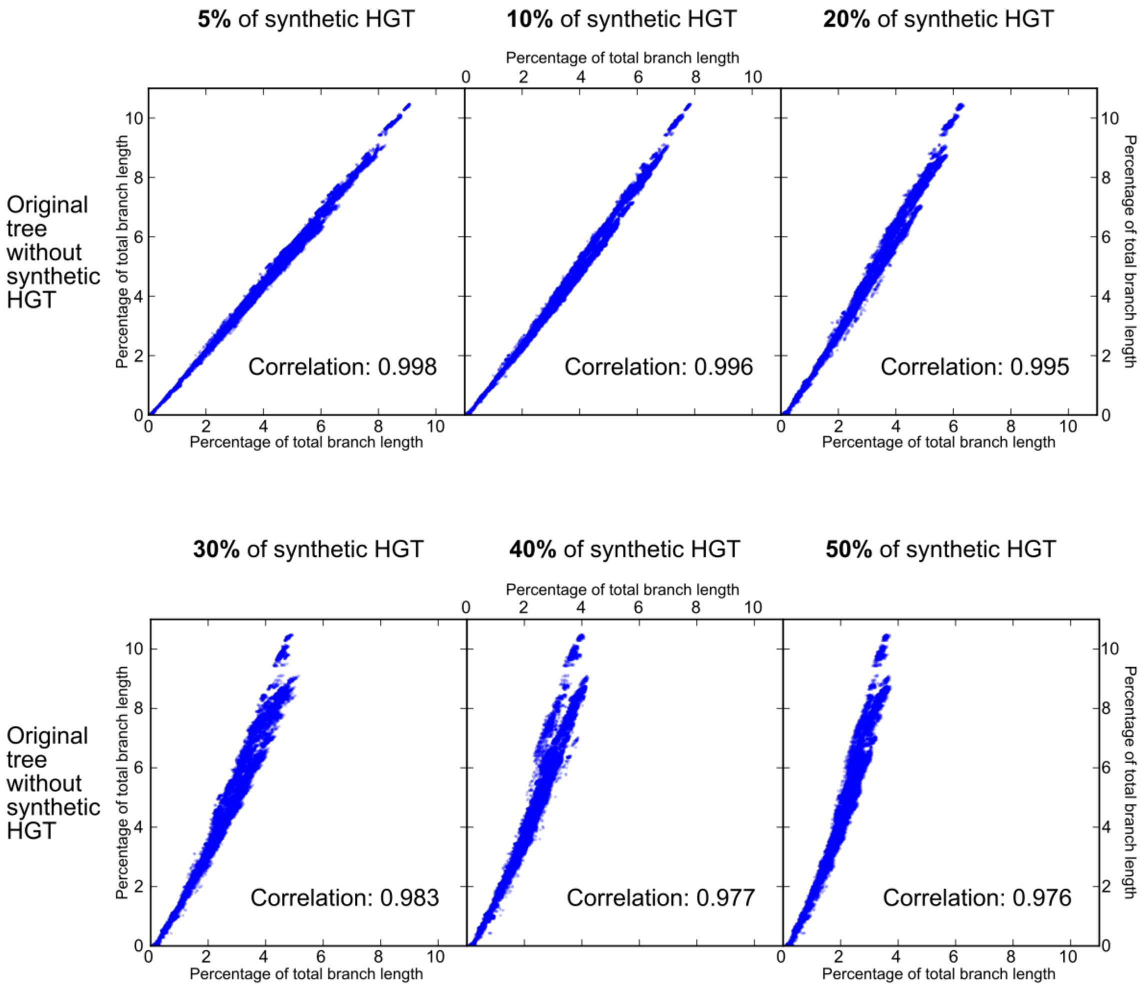
Nicola Segata[1,2], Daniela Börnigen[1,3], Xochitl C. Morgan[1,3], Curtis Huttenhower[1,3]

[1] Biostatistics Department, Harvard School of Public Health, 655 Huntington Avenue, 02115, Boston, MA
[2] Current address: Centre for Integrative Biology, University of Trento, Via Sommarive 14, 38123, Trento, Italy
[3] Broad Institute of Harvard and MIT, 301 Binney Street, 02142 Cambridge, MA
Corresponding author: Curtis Huttenhower, chuttenh@hsph.harvard.edu

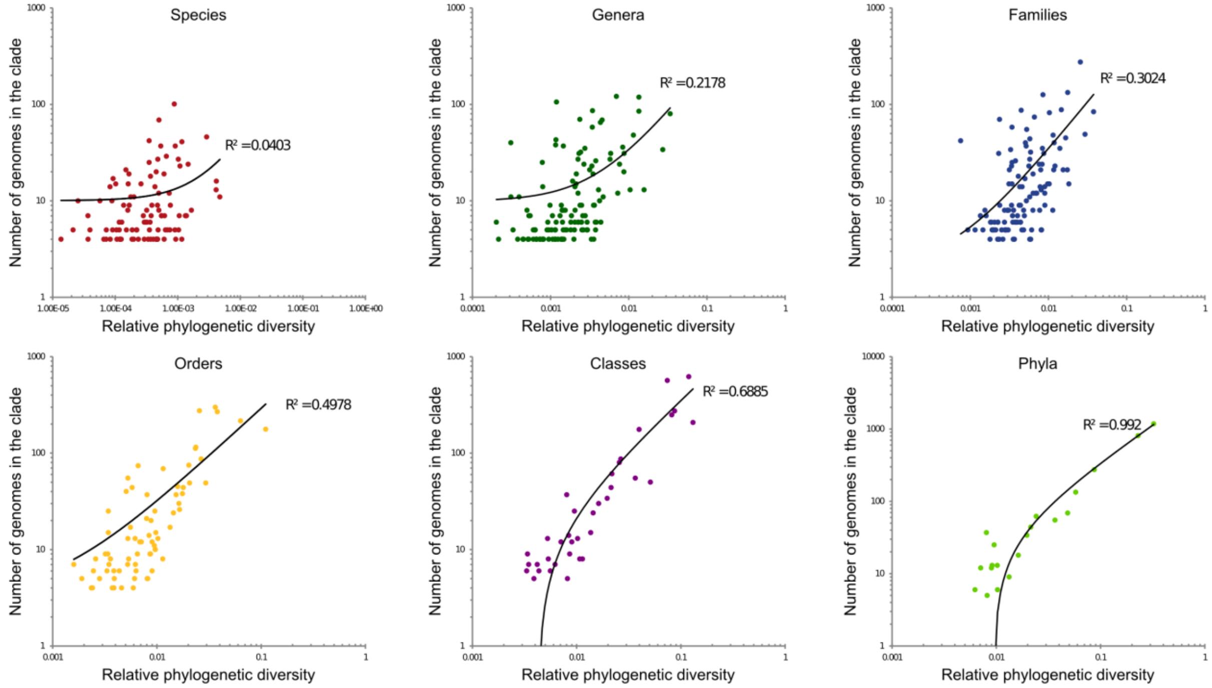
Supplementary Figures



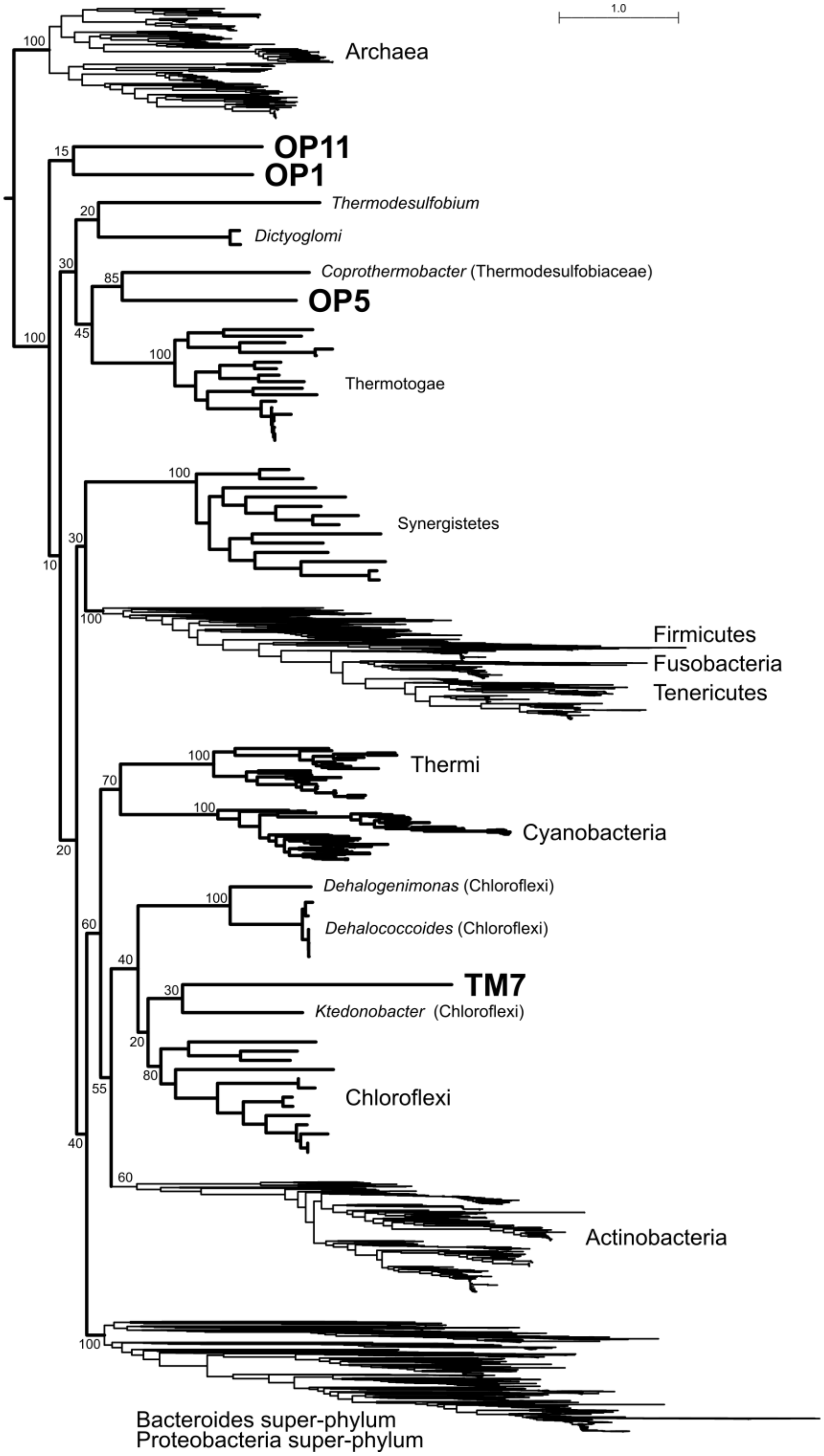
Supplementary Figure S1: Taxonomic precision and recall achieved by PhyloPhlAn using increasing numbers of universal proteins across all taxonomic levels. As compared to a gold standard derived from the IMG taxonomy, both precision (A) and recall (B) of inferred phylogenies increase at all taxonomic levels as up to the 500 most conserved proteins (values are averaged across all clades at each level). PhyloPhlAn outperforms alternative methods at all levels compared above genera (see Figure 2).



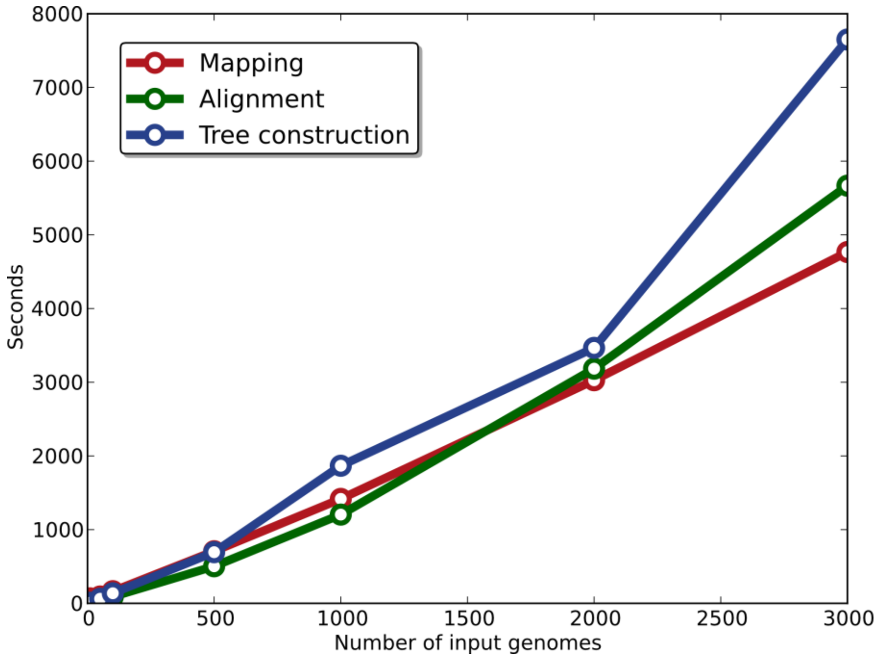
Supplementary Figure S2: PhyloPhlAn is robust to horizontal gene transfer as assessed by HGT simulations. Although the improved taxonomic consistency of our method already suggested a limited impact of horizontal gene transfer (HGT) as a confounding factor, we further assessed this potential issue by means of systematic gene transfer simulation. Specifically, each scatterplot contrasts the patristic distances among leaf genomes within the Actinobacteria phylogenetic tree as reconstructed by PhyloPhlAn using the unmodified input genomes and the genomes at increasing rates of synthetic HGT (5 to 50% of genes from one genome randomly moved to another; see Supplementary Methods). Correlation remains high even for artificially high levels of HGT, indicating correct placement of genomes due to the repeated measures offered by many conserved proteins. Overall branch length grows shorter as expected, however, due to the (again artificially) increased level of overall similarity among genomes induced by synthetic gene swapping. This was dependent on the amount of synthetic HGT with an overall reduction of ~20% for 10% synthetic HGT and 55% for the biologically unlikely case of 50% synthetic HGT.



Supplementary Figure S3: A comparison of PhyloPhlAn's relative phylogenetic diversity and the number of genomes at all taxonomic levels. Each panel reports all clades at a particular taxonomic level containing at least four genomes, comparing the assigned relative phylogenetic diversity (diameter) and the number of genomes on a logarithmic scale. Linear fits (distorted by log scaling) and R^2 values are computed as a measure of correlation, generally low at sufficiently detailed taxonomic levels.



Supplementary Figure S4: The PhyloPhlAn microbial tree of life allows the phylogenetic placement of candidate deep branching phylum-level clades such as OP1, Caldiserica (OP5), OP11, and TM7. The comprehensive PhyloPhlAn tree of life represented is collapsed here at variable resolution generally approximating the phylum level. The OP1 and Caldiserica divisions, each represented by one genome, and the similar divisions OP11 and TM7 (also represented by one genome) are automatically placed in agreement with existing manually curated evidence.



Supplementary Figure S5. Computational performance of the three main PhyloPhlAn steps as the number of input genomes increases. PhyloPhlAn is executed using 16 threads on an eight-processor six-core Intel Xeon E7540, with the number of input genomes ranging from 10 to 3,000. The mapping step was performed using USEARCH 5.0 and the alignment operation with MUSCLE 3.8. Both steps are dominated by the linear-scale constant even up to several thousand genomes, and further scale linearly with the number of threads available (up to the number of genomes for mapping and the number of universal proteins for alignment, typically 400). Phylogeny reconstruction (using FastTree 2.1) is slightly super-linear in practice and in the theoretical worst case O(N^1.5 log(N) L) for N sequences of length L (ref. 33).

Supplementary Tables

Supplementary Table S1. Number of newly sequenced genomes (not already present in the PhyloPhlAn DB) that lie in the LCA of the corresponding taxonomic clade at decreasing percentages of proteins considered from each genome. Two new genomes each from a total of 5 clades were considered. The set of proteins selected at different percentages was chosen randomly at each iteration.

Percentage of genes considered	Clades from which 2 newly sequenced genomes not present in the PhyloPhlAn DB were available				
	E. coli	**Mycoplasma**	**S. aureus**	**S. pneumoniae**	**B. subtilis**
	IMG ID1: 2516143071	IMG ID1: 2519899697	IMG ID1: 2513237206	IMG ID1: 2519899814	IMG ID1: 2510917016
	IMG ID2: 2519899564	IMG ID2: 2521172699	IMG ID2: 2521172729	IMG ID2: 2512564060	IMG ID2: 2518645535
100%	2	2	2	2	2
80%	2	2	2	2	2
60%	2	2	2	2	2
40%	2	2	2	2	2
20%	2	2	0	2	2
10%	2	2	0	1	2
5%	2	2	2	0	2
3%	1	1	2	2	2
1%	1	2	1	1	2

Supplementary Table S2: Number of proteins originally contained (i.e. 100% row above) in each genome

	Total		Conserved targets only	
	Genome #1	**Genome #2**	**Genome #1**	**Genome #2**
E. coli	5123	5618	368	367
Mycoplasma	724	613	124	123
S. aureus	2623	2919	325	327
S. pneumoniae	2148	2252	287	284
B. subtilis	4198	4034	373	367

Supplementary Methods

Ranking proteins for ubiquitous conservation and covered diversity

Each of the proteins identified in >1,000 genomes as described in the Methods was ranked based on the total alpha-diversity they spanned on the 16S rRNA species-level tree of life. The 16S rRNA microbial tree was built by selecting one representative for each of the species considered here from the Greengenes repository[11], aligning the representatives with MUSCLE[32] version 3.8 and constructing the phylogenetic tree with FastTree[33] version 2.1. Alpha-diversity was estimated using the total branch length of the subtree containing all species with the protein in at least one of the available genomes, normalized by the total branch length of the tree. Selecting only one 16S sequence per species is not expected to influence the universality-based ranking procedure, due to the high ratio of between-species to within-species 16S gene sequence variability and the extensive conservation of proteins selected for ranking. Incremental sets of proteins were then selected by maximizing the sum of the normalized average diversity and the normalized total coverage of the genomes considered in each set. The resulting sets of proteins represented the most ubiquitous and universal microbial proteins, up to a total of the 500 most universal proteins.

Building a high resolution tree of life

The homologs of each universal protein were aligned independently using MUSCLE[32]. The resulting alignments were edited before tree building by removing amino acid positions with gaps in more than 10% of the genomes or with the same amino acid in more than 95% of the genomes. This emphasized regions both universally conserved and phylogenetically discriminating. Furthermore, we selected up to 30 amino acids from each alignment with the lowest score and thus highest entropy, as estimated by MUSCLE. For each protein, the number of selected amino acid positions (minimum 4) was inversely linearly proportional to the rank, with 30 positions selected for the most conserved proteins and down to 4 randomly subsampled (occurring at rank ~284 down to 500).

The number of amino acid residues considered was reduced relative to the full-length alignment both for computational reasons and to increase the relative contribution of the most conserved residues in the final tree. The thresholds selected here could be further optimized by an a posteriori evaluation of tree topologies using the taxonomic consistency measures, but this would likely introduce a bias towards placement of the currently sequenced genomes at the cost of generalizability for new genomes. The number of selected positions was occasionally smaller for alignments with insufficiently many discriminative amino acids. We also retained the full sequence concatenation for up to the first 100 proteins.

For tree building, we concatenated each protein alignment's subsequence with those from previous proteins, introducing gaps for genomes missing a specific protein. As this was performed for up to 500 total proteins, the total size of the concatenated alignments ranged from 142 amino acids for the top 5 protein set to 5,208 for the top 500. Phylogenetic trees were then generated using the "minimum-evolution principle" with heuristic neighbor-joining[61], minimum-evolution interchanges and subtree-pruning-regrafting[62], and approximated maximum likelihood joining[63] applying FastTree[33] on the concatenated alignments (default

JTT+CAT model), shown to be a more accurate alternative to consensus trees, and the only strategy with sufficient performance to scale to almost 3,000 peptide sequences of several thousand amino acids in length[64].

The evaluation of tree precision and recall (defined as reported below) at increasing numbers of universal proteins (from 5 to 500) suggested that the optimal number of proteins to concatenate fell between 300 and 500 (Figure 2), and we thus used the 400 most conserved peptide sequences to provide a final recommended tree. This value was chosen to balance the minimum optimum observed here (~300) against robustness (suggesting more proteins) and overfitting to this genome set (suggesting fewer); it could be further optimized with the addition of more genomes in the future. A high-quality version of this tree was generated using RAxML[28] using bootstrapping (20 trees) with maximum likelihood maximization (gamma model of rate heterogeneity) in addition to FastTree's faster but somewhat more approximate approach.

Adding new genomes and reconstructing phylogenetic trees using the PhyloPhlAn pipeline

The 400 universal proteins are stored in a database including all variants identified in the original set of 2,887 genomes, which is used as reference for identifying them in new genomes via protein mapping. Variants are removed at 90% sequence similarity using UCLUST[46] to select representative seeds. This is robust to the addition of new genomes, as conservation of these 400 markers is computed with respect to spanned diversity rather than number of genomes. Identification of these markers in a newly sequenced genome is performed by mapping using USEARCH[46] with best hit policy and evalue threshold at 1e-40. After this mapping, the procedures described above for multiple sequence alignment, concatenation and tree construction are again performed in order to place the genome (existing MUSCLE alignments are also stored, allowing this program to perform an efficient partial insertion). The reduced redundancy protein database includes 292,370 peptide sequences available for download (PhyloPhlAn website[30], with a copy in Supplementary Software).

Computationally, PhyloPhlAn performs these operations by parallelizing the computation in up to 400 processes, each of them specific for a single universal protein (for MUSCLE alignment) or for a single genome (for USEARCH mapping). The only exception is the operation of final tree reconstruction, for which parallelization is also a future option[28]. Moreover, again with the exception of the final reconstruction, the software pipeline for mapping, aligning, and concatenating the highly conserved universal PhyloPhAn proteins scales almost linearly with respect to the number of input genomes making it suitable for the rapidly increasing number of available genomes (Supplementary Figure S5).

Measures of tree accuracy: taxonomic precision, recall, and relative phylogenetic diversity

We developed a set of measures to evaluate the accuracy of a microbial phylogeny with respect to a curated taxonomy. Although this is in part a tree comparison problem, it is inappropriate to directly compare the structure or topology of a bifurcating phylogeny, including evolutionary branch lengths, to a multifurcating taxonomy without meaningful branch lengths. This precluded the application of a direct tree distance metric like the Robinson–Foulds distance[65].

Instead, we first wished to measure the taxonomic consistency of clades in a phylogenetic tree, i.e. the precision with which phylogenetically related organisms were labeled taxonomically. Ideally, a taxonomic clade with *n* organisms should form a monophyletic subtree in the phylogenetic tree comprising all and only the *n* organisms. When this is the case, the taxonomic precision is 1.0; otherwise, our index reflects the total diversity spanned by the *n* organisms normalized by the diversity spanned by their lowest common ancestor:

$$Cons_P(c) = \frac{\alpha(L(c))}{\alpha(LCA(L(c))} \qquad \text{(S1)}$$

where *c* is the clade under consideration, *LCA* is the lowest common ancestor, *L* returns the set of leaves (genomes) within a clade, and α is the total branch length of a clade (if applied on an internal node as in the case of the result of the LCA function) or of the tree spanning a set of organisms (if applied on a set of leaf nodes as in the case of the genomes returned by *L*). Note that this precision must be measured on a per-clade basis, e.g. for the species *Escherichia coli* or the phylum Firmicutes. When appropriate, we also provide average scores across a taxonomic level (e.g. averages across all species or all phyla).

We also introduced a recall measure indicating whether taxonomically similar organisms within a clade were grouped phylogenetically. This identified the fraction of taxa distant from the largest taxonomically consistent subtree (LTCS) in the phylogeny for a clade *c*. Cons$_R$ thus identifies the taxa within *k* times the diameter of the LTCS:

$$Cons_R(c) = \frac{\left|x \in L(c)\ s.t.\ dist(x, \mathrm{LTCS}\left(\mathrm{L(c)}\right) < k \times \delta(\mathrm{LTCS}\left(\mathrm{L(c)}\right))\right|}{|L(c)|} \qquad \text{(S2)}$$

where δ is the diameter (i.e. the branch length separating the most distant leaves), *dist* gives the branch length distance between nodes (possibly leaves), and *k* is a tunable parameter. For this study, *k* was typically fixed at 1.0 unless otherwise noted. Like precision, this measure of recall also operates on a per-clade basis and was averaged across all clades within a taxonomic level when appropriate.

Finally, we introduced a measure of relative phylogenetic diversity, which estimates the diversity spanned by a given taxonomic clade within a phylogenetic tree (typically the whole phylogeny), normalized by the latter's total diversity:

$$Res(c, c_0) = \frac{\alpha(L(c))}{\alpha(c_0)} \quad \text{with } c_0 \text{ an ancestor of c (S3)}$$

While precision and recall effectively assess a phylogeny using a taxonomy as a gold standard, relative phylogenetic diversity instead describes a taxonomy in terms of its phylogenetic breadth. As such, below we use precision and recall to evaluate a range of reconstructed phylogenies, whereas we use relative phylogenetic diversity to determine what the final "best" phylogeny implies about the evolutionary breadth of the current microbial taxonomy.

It is worth noting that Cons$_P$ and, to some extent, Cons$_R$, are unlikely to produce a value of identically 1.0 for all clades even in an ideal phylogenetic tree, as it is very likely that some errors are present in the microbial taxonomy and that some sequenced organisms are misannotated. It is, again, difficult to quantitatively evaluate any phylogenetic tree due to the absence of a single ground truth. Although evolutionary relationships as captured by phylogeny are not guaranteed to follow curated taxonomic groupings, relative increases in precision cannot occur randomly, as indicated by the high accuracy of curated ribosomal proteins or increasing numbers of conserved proteins as considered by our method and others. Thus if a phylogenetic tree A shows an overall taxonomic consistency higher than another phylogenetic tree B, A should be safely considered more accurate than B as a higher consistency with phenotypic data is very unlikely to be achieved by chance.

Building phylogenetic trees using 16S and ribosomal proteins

Since 16S rRNA genes are not consistently annotated in incomplete or draft genomes, we identified these genes from each of the >2,800 genomes with a nucleotide blast (e-value 1e-50) of the 1,221 16S rRNA sequences retrieved from Greengenes[11] as species representatives against the CDSs of the genomes. The CDS with the highest sequence similarity score to a 16S species representative was selected for each genome with multiple 16S genes. 56 of the 2,881 genomes (46 draft) did not possess a 16S rRNA gene and were thus excluded from the 16S tree generation. Sequences were aligned with MUSCLE and a phylogeny created with FastTree. We also reconstructed the microbial tree of life using the 31 proteins corresponding to single COG families[66] as described by the original works[15,16]. These 31 proteins were identified in each genome from IMG annotation[31], disambiguating multi-copy proteins based on best blast e-value. The resulting sequences were also aligned individually using MUSCLE, concatenated by adding gaps for missing proteins, and the resulting alignment processed using FastTree.

Horizontal gene transfer simulation

A synthetic HGT simulation was implemented by selecting genes uniformly at random with increasing probabilities *p* from each genome in the input set (taken for this simulation as the phylum Actinobacteria), removing them from the source genome, and including them in a randomly chosen target genome. The probability *p* was set at incrementally increasing values in the interval [0.05,0.5] to simulate systematic HGT affecting from 5 to 50% of genes. We evaluated the differences and biases compared to the original tree without synthetic HGT by contrasting the patristic distances induced by the trees[67] and computing the correlation among them as reported in Supplementary Figure S2.

Supplementary references

61 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406 (1987).

62 Hordijk, W. & Gascuel, O. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* **21**, 4338-4347 (2005).

63 Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-376 (1981).

64 Gadagkar, S. R., Rosenberg, M. S. & Kumar, S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool* **304**, 64-74 (2005).

65 Robinson, D. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131-147 (1981).

66 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science (New York, N.Y.)* **278**, 631-637 (1997).

67 Fourment, M. & Gibbs, M. J. PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. *BMC Evolutionary Biology* **6**, 1 (2006).