Quantitative phylogenetic assessment of microbial communities in diverse environments

von Mering C[1,4], Hugenholtz P[2], Raes J[1], Tringe SG[2], Doerks T[1], Jensen LJ[1], Ward N[3], Bork P[1].

[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

[2] DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

[3] The Institute for Genomic Research, Rockville, MD 20850, USA

[4] present address: University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

The taxonomic composition of environmental communities is an important indicator of their ecology and function. Here, we use a set of protein-coding marker genes, extracted from large-scale environmental shotgun sequencing data, to provide a more direct, quantitative and accurate picture of community composition than traditional rRNA-based approaches using polymerase chain reaction (PCR). By mapping marker genes from four diverse environmental data sets onto a reference species phylogeny, we show that certain communities evolve faster than others, determine preferred habitats for entire microbial clades, and provide evidence that such habitat preferences are often remarkably stable over time.

Microorganisms are estimated to make up more than a third of Earth's biomass (*1*). They play essential roles in the cycling of nutrients, interact intimately with animals and plants, and directly influence the Earth's climate. Yet, our molecular and physiological knowledge of microbes remains surprisingly fragmentary – largely because most naturally occurring microbes cannot be cultivated in the laboratory (*2*).

For characterizing this 'unseen majority' of cellular life, the first step is to provide a taxonomic census of microbes in their environments (*3-6*). This is usually achieved by cloning and sequencing their ribosomal RNA genes (most notably the 16S/18S small subunit rRNA). This approach has been extremely successful in revealing the overwhelming diversity of microbial life (*7*), but it also has some limitations due to quantitative errors: the PCR step introduces amplification bias, and it generates chimeric and otherwise erroneous molecules that hamper phylogenetic analysis ((*8*), see Supplementary Information for details).

Shotgun sequencing of community DNA ('metagenomics') provides a more direct and unbiased access to uncultured organisms (*9-12*): No PCR amplification step is involved, and since no specific primers or sequence anchors are needed, even very unusual organisms can be captured by this technique. While current metagenomics data are still not entirely free of quantitative distortions (mostly due to sample preparation), remaining biases are bound to diminish further with the optimization of yield and reproducibility of DNA extraction protocols (*13-15*).

In order to utilize metagenomics data for taxonomic profiling, we analyze 31 protein-coding marker genes that have been shown previously to provide sufficient information for phylogenetic analysis (they are universal, occur only once per genome and are rarely transferred horizontally (*16*)). We extract these marker genes from metagenomics sequence data (see Supplementary Information), align them to a set of hand-curated reference proteins, and use maximum likelihood to map each sequence to an externally provided phylogeny of completely sequenced organisms (*tree of life*; we use the tree from (*16*), although any reference tree can be used as long as the marker genes have been sequenced for all its taxa). Our procedure provides branch length information and confidence ranges for each placement ((*17*), Figure 1), allowing statements such as "*this unknown sequence evolves relatively fast, is from a proteobacterium (95% confidence), and more specifically, probably from a novel clade related to the Campylobacterales (65% confidence)*". Importantly, the procedure weighs the number of informative residues that are found on each sequence fragment, and adjusts the spread and confidence of its placement in the tree accordingly (after alignment, concatenation and gap removal, the number of remaining informative residues ranges from 80 to more than 3000, per sequence fragment, see Supplementary Information). We have implemented the entire phylogenetic assignment protocol as an automated software pipeline with a web-interface that allows submission of sequences online (http://MLtreemap.embl.de/).

Jack-knife validation of our method (i.e. leaving out various parts of the reference tree, and measuring the consequences on placement accuracy; see Supplement Methods) showed that the performance of our method depends on the completeness and balance of the reference tree: the larger the phylogenetic distance to any known relative of an environmental sequence, the less precise is its placement. Overall, the mapping precision is remarkably good, as long as each sequence has some relative from the same phylum among the reference genomes

(Figure S2). In contrast, BLAST-based assignments of taxonomy based on 'best hit', a frequently used method, are more error-prone: for example, more than 10% of the sequences change to a different domain of life (e.g. changing assignment from *Bacteria* to *Archaea*) upon removal of the phylum to which they originally mapped, compared to merely 0.19% using our method (Figure S2). Moreover, since the best BLAST match always assigns a single organism as the most likely phylogenetic neighbor, it does not specify the level of relatedness (e.g. class-, order-, or phylum-level), which is needed to trace organisms in their preferred habitats and through time.

In one of the recent, large-scale metagenomics sequencing projects (*11*), traditional PCR-based assessment of 16S rRNA molecules was executed in parallel to the shotgun sequencing. This enabled us to compare our approach to this currently most-widely used experimental method for phylogenetic profiling of environments. Overall, the relative abundances of phyla as reported by both methods were broadly similar, although the metagenomics approach appears quantitatively closer to the truth as can be measured by comparison to rRNAs that are contained directly in the PCR-independent shotgun reads (see Supplementary Information for a detailed analysis). The PCR-based approach presumably suffers from amplification biases and from copy-number variations among rRNA genes in bacteria (18), but benefits from an exhaustive coverage of phyla among known rRNA sequences. In contrast, the approach we present here requires far more resources in terms of sequencing and computation, but at least for phyla already represented among fully sequenced genomes, it is noticeably more quantitative. Our approach should essentially be seen as a by-product of metagenomics sequencing projects, which are usually conducted for functional purposes (see Supplementary Information for a detailed discussion of the strengths, weaknesses and complementarities of both approaches).

We applied our procedure to four large, heterogeneous datasets of microbial community sequences, derived from distinct and geographically separate environments (*10-12*). The consistent treatment of the data allowed us to quantitatively compare habitat preferences in the context of the tree of life (Figure 2, Figure S1, see also Figure S3 for robustness estimates).

Overall, we observed a remarkably un-even representation of previously sequenced genomes in naturally occurring communities. Some parts of the tree of life (such as the *Streptococci* or the *Enterobacteriales*) are well-covered by published genome sequencing projects, but they only represent a small part of naturally occurring microbes. Conversely, entire phyla such as the *Acidobacteria* or the *Chloroflexi* are poorly represented among the sequenced genomes, but widely abundant in natural communities.

As noted previously (*19*), we find *Proteobacteria* to be the most dominant phylum of microbial life in both marine and soil environments (Figure 2). However, as is the case with other phyla, marked differences within the *Proteobacteria* become apparent: relatives of the *Rickettsiales*, for example (including the marine genus *Pelagibacter* (*20*)), are mostly found in the surface water samples , whereas relatives of *Rhizobiales* or *Burkholderiales* are mostly found in the soil sample. We observed surprisingly few endospore-forming organisms in the community sequences: both *Bacilli* and *Clostridia* are quite rare, their largest combined abundance is a mere 1% (in soil). Similarly, *Actinobacteria* (many of which have a spore stage)

range from being virtually absent in the acidic mine drainage biofilm to only 6.2% in the soil sample. It is conceivable that spores are underrepresented in the data (they may withstand the DNA extraction protocols), but at least among the vegetative, actively growing cells, spore-formers appear to be a minority.

Quantitative analyses of relatively rare phyla, as for example in the case of the spore-formers mentioned above, can potentially suffer from limited sampling. While our approach uses 31 marker genes with a total of about 7,500 amino acid residues per genome, low-abundance organisms might be represented by only a few of these (the total number of sufficiently complete marker genes useable for our approach ranges from 247 for the smallest dataset, up to 15,741 for the largest dataset). We have quantified the potential under-sampling errors, using jackknife and bootstrap analysis (Figure S3). These tests show that, for the worst case of a low abundance clade in the smallest dataset, the quantitative error due to under-sampling is on the order of 50% (Figure S3). However, such errors are bound to decrease with the expected rise in sequencing depth, facilitated by technological advances. In addition, even for a low estimate such as the 1% abundance mentioned above for *Bacilli* and *Clostridia*, the current data support a 95% confidence interval of 0.995% - 2.153%, meaning that endospore-formers are indeed rare in soil, and not just under-sampled. Generally, none of the results reported here would change much if all datasets had as many as 15,000 marker genes sampled (in particular since we do not comment on diversity, and because we discuss entire clades, not individual species).

Almost all placements of environmental sequences occurred at relatively deep, internal nodes in the reference tree; only a few could be placed towards the tips as close relatives of the cultured and sequenced genomes. Indeed, the average sequence similarity of the 'best hits' of environmental sequences to sequenced genomes is usually less than 60% (for soil, the median identity is only 47%). This dissimilarity is reflected in the maximum likelihood branch lengths: on average, more than 0.3 substitutions per site have occurred since the branching from the reference tree. This corresponds roughly to the sequence divergence between beta- and gamma-proteobacteria, which has been tentatively dated at more than 500 million years ago (*21-23*), clearly enough time for functional capabilities and lifestyles to have changed. Thus, the closest sequenced relative of an environmental microbe should generally *not* be considered as a reliable guide for its phenotypes and functions.

The environments we analyzed contained a few sequences that were placed unusually deep in the tree, i.e. basal to the three known domains of life: *Archaea*, *Bacteria* and *Eukaryota*. Upon closer inspection, we determined that most of these deep placements in fact originated from lineages not yet represented among sequenced genomes (for example the *Cenarchaeales*, a deeply branching archaeal lineage, data not shown). Therefore, it is likely that the remaining deep placements will also find a home as soon as more lineages are included in the reference tree, rather than belonging to a hypothetical '4th domain' of life.

The maximum likelihood branch lengths, as measured by our method, provide detailed information on the community-wide distribution of evolutionary rates (that is, the rates at which mutations occur and are fixed). We therefore assessed, for each sequence fragment placed into the tree, the cumulative branch length from the tip of its branch down to the base of the

corresponding phylum, and compared these to the branch lengths of all known reference organisms in that same phylum, measured for the very gene families found on the fragment (Figure 3; very deeply placed fragments are compared to all phyla in their sister clade). Although not all 31 of the marker genes are present for each organism in the metagenomics data, the measurements of relative rates in each gene family revealed distinct branch length distributions for the four environmental communities tested. These indicate that organisms at the ocean surface evolve the fastest, whereas organisms in the soil evolve the slowest (Figure 3). Large-scale trends like this, involving entire communities, have been observed previously mainly for multicellular organisms (e.g. a dependency between latitudinal geographic location and mutation rates in plants (*24*)). In the case of microbes, fast-evolving species were previously known in the context of symbiotic or pathogenic settings, or in cases of extreme genome 'streamlining' (*20, 25*). The more subtle, global variations in mutation rates reported here may be caused by differences in population sizes, generation times, or by the abundance of external mutagens (such as the strong fluxes of ultraviolet light in ocean surface water). In the case of soil, the apparent evolutionary stability at the sequence level is also consistent with intermittent periods of dormancy (for example during winter and/or under desiccation).

Our tree-based mapping (with an implicit molecular clock) also allows us to trace the habitat preference of microbial organisms through time, and thus enables us to estimate how frequently lineages change their preferred environment. At short to intermediate evolutionary timescales, we observe a noticeable stability of habitats: many of the closer relatives in the tree show the same environmental preference, indicating that microbial lineages do not very often change (or specialize) their life-styles and habitats (Figure 2). Conversely, at longer timescales, we do observe significant changes of preferred habitats, for example within diverse lineages of at least two phyla, namely *Proteobacteria* and *Cyanobacteria;* this is consistent with the observed morphological and ecological variability of cultured isolates from most phyla. For example, in the case of *Cyanobacteria*, we identify relatives of the fast-evolving and widespread *Prochlorococci* in the ocean sample, whereas more basal, slower evolving *Cyanobacteria* such as *Gloeobacter* are mostly found in the soil sample.

Even though molecular methods tend to find most phyla ubiquitously, Baas-Becking and Beyerinck already postulated decades ago that microbial taxa have preferred environments: *"for microbial taxa, everything is everywhere — but the environment selects"* ((*26*) and references therein). The hypothesis posits that microorganisms are frequently dispersed globally, and that they are only subsequently selected by the environments based on their functional capacities. Existing communities would thus constantly be challenged by intruders from non-specialist phyla who may occasionally survive simply by chance, acquiring the necessary functionality through horizontal gene transfer (*27-29*). Our observations provide quantitative support for this hypothesis, showing strong environmental preference along lineages, but with a time-dependent decay. We confirmed and extended this finding, by also analyzing habitat information available for cultivated strains in culture collections, as well as the large body of publicly available rRNA sequence data. Both datasets provide information about hundreds of habitats, and allow an approximate ranking of lineage separation events in time: in the case of rRNA sequence data, branch length information can be analyzed using a

global phylogeny of small subunit RNA sequences, whereas in the case of cultivated strains, taxonomic assignments can be parsed for the last taxonomic rank still shared (for details, see Supplementary Information). Indeed, we observe a remarkable time-dependent stability of habitats and show that for any two microbial isolates, the similarity of their annotated habitat (as measured by automated keyword comparisons) is strongly correlated to their evolutionary relatedness (Figure 2, panels B & C). We observe such common habitat preferences surprisingly far back in time – even strains related only at the level of taxonomic *order* are still significantly more frequently found in the same environment than a random pair of isolates (Figure 2C). Thus, most microbial lineages remain associated with a certain environment for extended time periods, and successful competition in a new environment seems to be a rare event. The latter might require more than just the acquisition of a few essential functions; probably only a limited number of functionalities are self-sufficient enough, and provide sufficient advantage, to be pervasively transferred (*30*)). For most other adaptations, fine-tuned regulation and/or subtle changes in the majority of proteins may be needed. As this is difficult to achieve, well-adapted specialists might in fact rarely be challenged in their environment. This does not rule out the presence of a 'long tail' of rare, atypical organisms in each environment (*31*), but most microbial clades do seem to have a preferred habitat.

Taken together, our alternative approach of taxonomic profiling of complex communities has sufficient resolution to uncover differences in evolutionary rates of entire communities, as well as long lasting habitat preferences for bacterial clades. The latter raises the question of how many distinct environmental habitats there are on earth – a factor that might ultimately determine the true extent of microbial biodiversity.

FIGURE LEGENDS

Figure 1: Assessing community taxonomy from metagenomics sequence data
Schematic diagram depicting how a restricted set of marker genes can be used for phylogenetic characterization of community microbes from poorly assembled sequence data. Instances of the marker genes are sought in the sequences, and assessed relative to an external tree-of-life phylogeny using maximum likelihood scoring. A central step in the mapping procedure is the assignment of a confidence range for each placement, thereby avoiding to place sequence fragments too overly confident if they are short, or otherwise uninformative.

Figure 2: Habitat/Phylotype associations and their stability in time
A) Four microbial communities are mapped onto the same reference tree. Pie-charts represent the various environments in which a particular tree clade has been observed. If there is a clear preference, lines are colored accordingly, see Supplemental Methods. B) Comparison of rRNA sequences from public databases, indicating the similarity of habitats from which they were sampled. C) Comparison of cultured microbial strains, plotting habitat similarity against their level of relatedness in the NCBI taxonomy. For the taxonomic level of *order*, and all closer relations, the difference over random is highly significant ($p < 10^{-6}$).

Figure 3: Distinct evolutionary rates of environmental communities
Organisms found in the surface waters of the Sargasso Sea have accumulated, on average, the largest number of mutations (i.e. evolved fastest), those in the agricultural soil the fewest. For each dataset, the branch lengths of the placements are plotted as dots. Each branch length is expressed relative to the median of branch lengths of known genomes in the same phylum, or against all phyla in the sister clade in the case of very deep placements. The quantiles 5%, 25%, 50% (median), 75% and 95% are indicated. All datasets differ highly significantly (two-sided Kolmogorov-Smirnov tests, $p <= 10^{-5}$, except for the comparison of acidic mine drainage with whale bone: $p < 0.05$). The number of data points underlying each distribution is as follows: ocean surface water – 15,741 genes on 9,286 contigs, acidic mine drainage – 275 genes on 148 contigs, deep sea whale bones [three sub-samples pooled] – 630 genes on 362 contigs, and agricultural soil – 598 genes on 395 contigs.

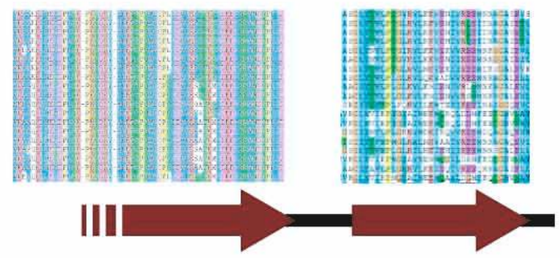
REFERENCES

1. W. B. Whitman, D. C. Coleman, W. J. Wiebe, *Proc Natl Acad Sci U S A* **95**, 6578 (Jun 9, 1998).
2. J. T. Staley, A. Konopka, *Annu Rev Microbiol* **39**, 321 (1985).
3. S. J. Giovannoni, T. B. Britschgi, C. L. Moyer, K. G. Field, *Nature* **345**, 60 (May 3, 1990).
4. D. M. Ward, R. Weller, M. M. Bateson, *Nature* **345**, 63 (May 3, 1990).
5. T. M. Schmidt, E. F. DeLong, N. R. Pace, *J Bacteriol* **173**, 4371 (Jul, 1991).
6. N. R. Pace, *Science* **276**, 734 (May 2, 1997).
7. P. Hugenholtz, B. M. Goebel, N. R. Pace, *J Bacteriol* **180**, 4765 (Sep, 1998).
8. F. von Wintzingerode, U. B. Gobel, E. Stackebrandt, *FEMS Microbiol Rev* **21**, 213 (Nov, 1997).
9. C. S. Riesenfeld, P. D. Schloss, J. Handelsman, *Annu Rev Genet* **38**, 525 (2004).
10. J. C. Venter *et al.*, *Science* **304**, 66 (Apr 2, 2004).
11. S. G. Tringe *et al.*, *Science* **308**, 554 (Apr 22, 2005).
12. G. W. Tyson *et al.*, *Nature* **428**, 37 (Mar 4, 2004).
13. G. M. Luna, A. Dell'Anno, R. Danovaro, *Environ Microbiol* **8**, 308 (Feb, 2006).
14. H. A. Barton, N. M. Taylor, B. R. Lubbers, A. C. Pemberton, *J Microbiol Methods* **66**, 21 (Jul, 2006).
15. I. M. Kauffmann, J. Schmitt, R. D. Schmid, *Appl Microbiol Biotechnol* **64**, 665 (Jun, 2004).
16. F. D. Ciccarelli *et al.*, *Science* **311**, 1283 (Mar 3, 2006).
17. K. Strimmer, A. Rambaut, *Proc Biol Sci* **269**, 137 (Jan 22, 2002).
18. J. A. Klappenbach, P. R. Saxman, J. R. Cole, T. M. Schmidt, *Nucleic Acids Res* **29**, 181 (Jan 1, 2001).
19. M. S. Rappe, S. J. Giovannoni, *Annu Rev Microbiol* **57**, 369 (2003).
20. S. J. Giovannoni *et al.*, *Science* **309**, 1242 (Aug 19, 2005).
21. F. U. Battistuzzi, A. Feijao, S. B. Hedges, *BMC Evol Biol* **4**, 44 (Nov 9, 2004).
22. D. F. Feng, G. Cho, R. F. Doolittle, *Proc Natl Acad Sci U S A* **94**, 13028 (Nov 25, 1997).
23. H. Ochman, S. Elwyn, N. A. Moran, *Proc Natl Acad Sci U S A* **96**, 12638 (Oct 26, 1999).
24. S. Wright, J. Keeling, L. Gillman, *Proc Natl Acad Sci U S A* (May 3, 2006).
25. A. Dufresne, L. Garczarek, F. Partensky, *Genome Biol* **6**, R14 (2005).
26. J. B. Martiny *et al.*, *Nat Rev Microbiol* **4**, 102 (Feb, 2006).
27. W. F. Doolittle, *Trends Cell Biol* **9**, M5 (Dec, 1999).
28. R. F. Doolittle, *Curr Opin Struct Biol* **15**, 248 (Jun, 2005).
29. I. Chen, P. J. Christie, D. Dubnau, *Science* **310**, 1456 (Dec 2, 2005).
30. N. U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, *Nature* **439**, 847 (Feb 16, 2006).
31. C. Pedros-Alio, *Trends Microbiol* **14**, 257 (Jun, 2006).

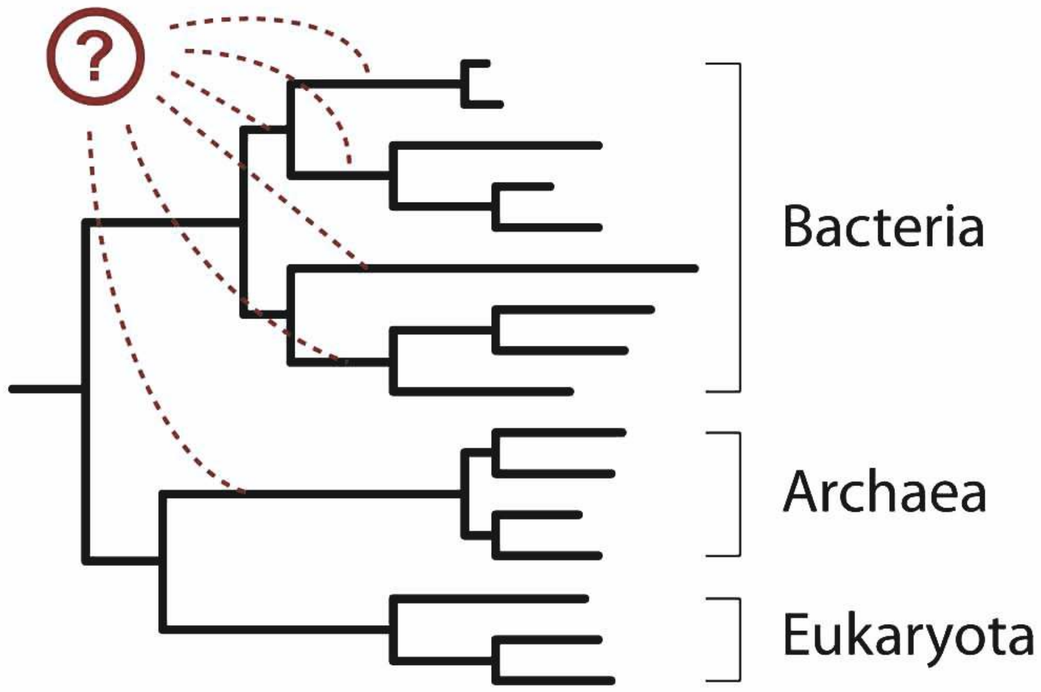
32. The authors wish to thank Peter Dawyndt for providing an early version of his integrated strain database, and members of the Bork team for insightful discussions. This work has been supported by the European Union through its BioSapiens and GeneFun networks, and by the German Federal Government through its National Genome Research Network (NGFN).

identify phylogenetically informative marker genes in environmental DNA fragment

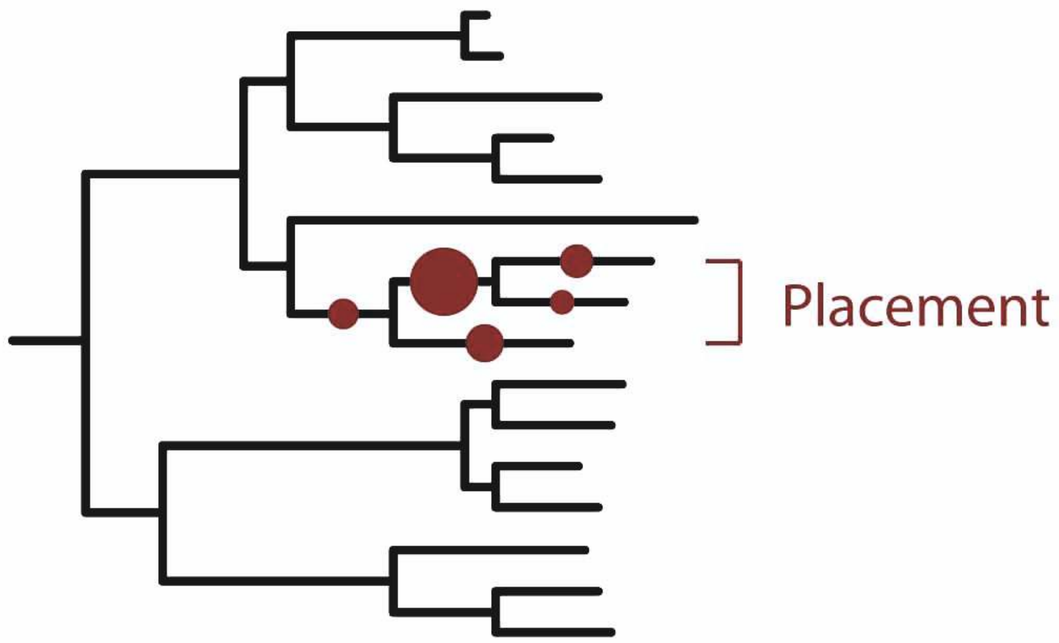
align markers to reference genes from sequenced genomes

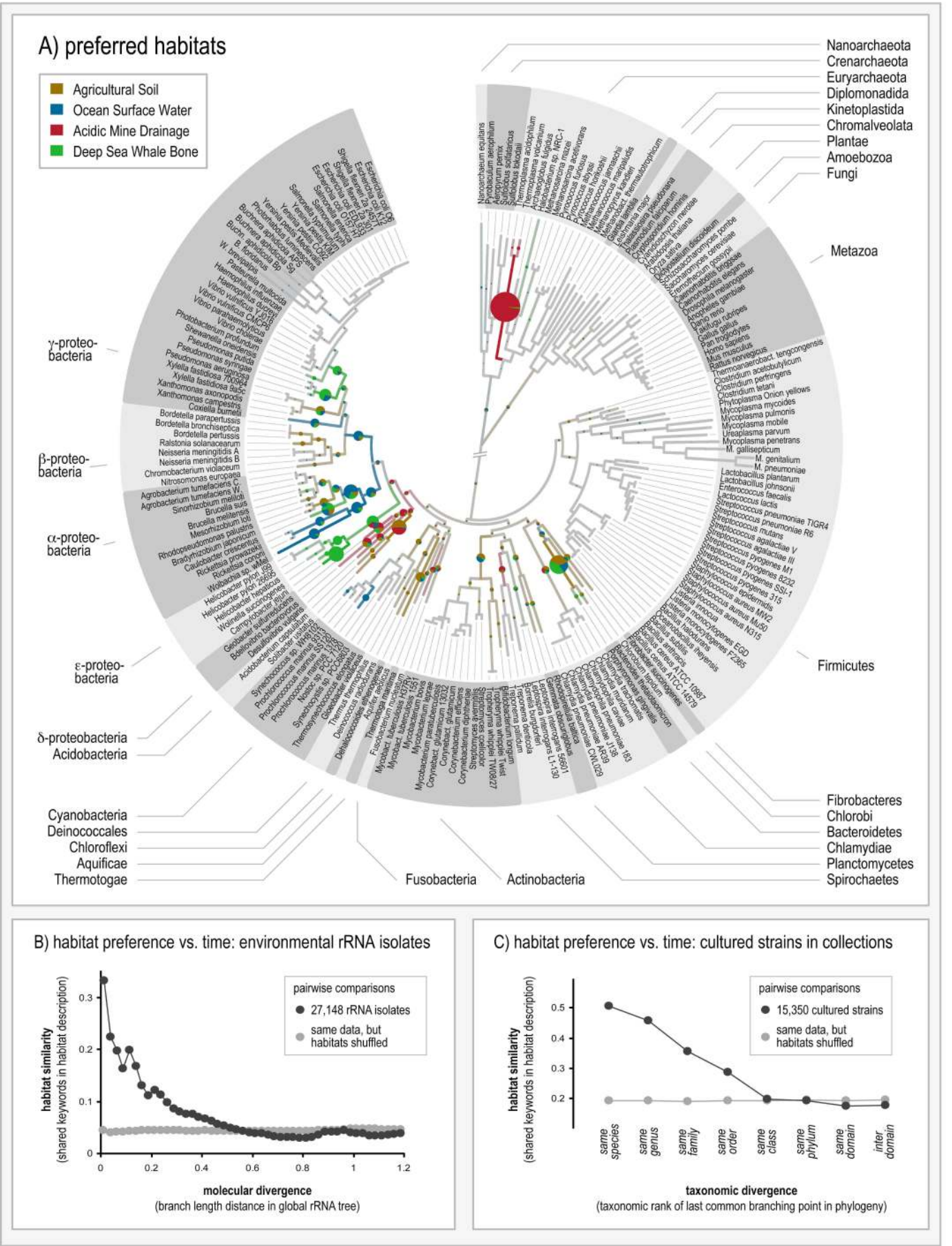


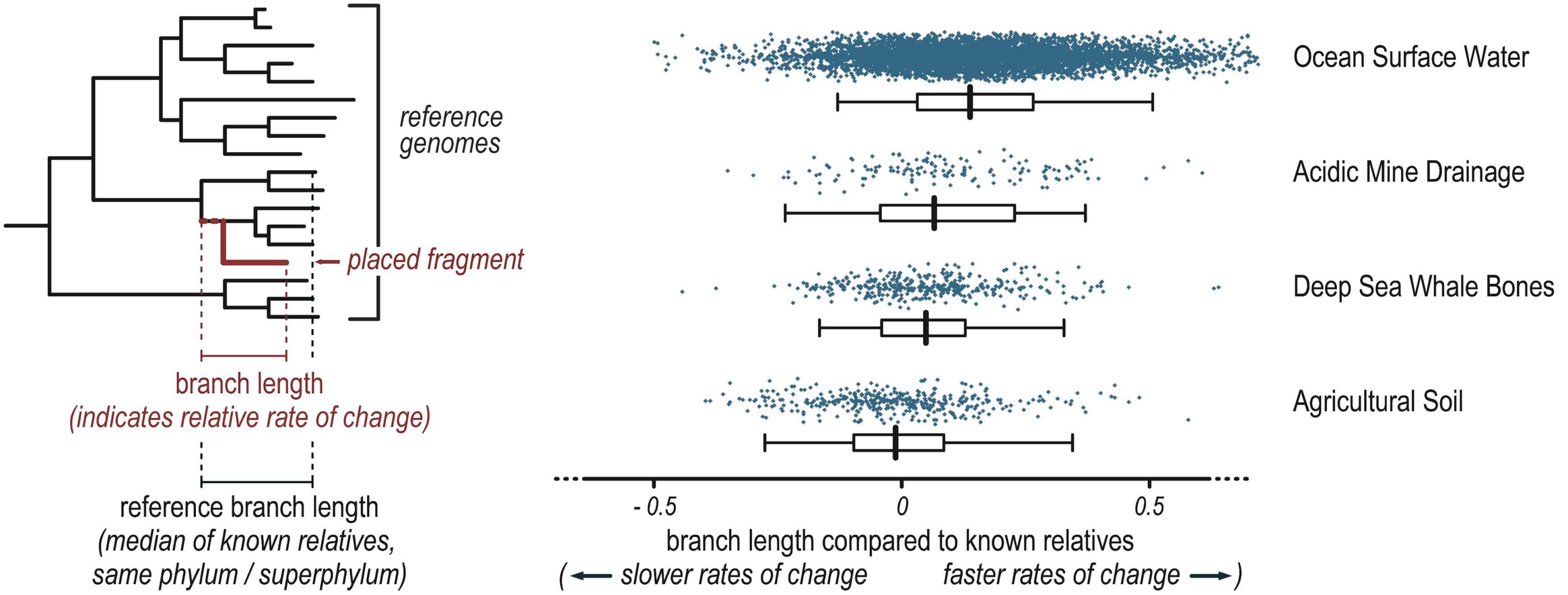
test all possible phylogenetic positions (in a reference tree of completely sequenced species)



place the DNA fragment probabilistically, using maximum likelihood (resulting in a weighted confidence range)




A) preferred habitats
Agricultural Soil
Ocean Surface Water
Acidic Mine Drainage
Deep Sea Whale Bone
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Diplomonadida
Kinetoplastida
Chromalveolata
Plantae
Amoebozoa
Fungi
Metazoa
γ-proteo-bacteria
β-proteo-bacteria
α-proteo-bacteria
ε-proteo-bacteria
δ-proteobacteria
Acidobacteria
Cyanobacteria
Deinococcales
Chloroflexi
Aquificae
Thermotogae
Fusobacteria
Actinobacteria
Firmicutes
Fibrobacteres
Chlorobi
Bacteroidetes
Chlamydiae
Planctomycetes
Spirochaetes
B) habitat preference vs. time: environmental rRNA isolates
habitat similarity
(shared keywords in habitat description)
pairwise comparisons
27,148 rRNA isolates
same data, but habitats shuffled
molecular divergence
(branch length distance in global rRNA tree)
C) habitat preference vs. time: cultured strains in collections
habitat similarity
(shared keywords in habitat description)
pairwise comparisons
15,350 cultured strains
same data, but habitats shuffled
same species
same genus
same family
same order
same class
same phylum
same domain
inter domain
taxonomic divergence
(taxonomic rank of last common branching point in phylogeny)


reference genomes
placed fragment
branch length
(indicates relative rate of change)
reference branch length
(median of known relatives, same phylum / superphylum)
Ocean Surface Water
Acidic Mine Drainage
Deep Sea Whale Bones
Agricultural Soil
- 0.5
0
0.5
branch length compared to known relatives
(← slower rates of change faster rates of change →)

– SUPPLEMENTARY INFORMATION –

Quantitative phylogenetic assessment of microbial communities in diverse environments.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P.

Contents:

Methods and Procedures

Phylogenetic markers and their detection

A set of 31 protein-coding, universally occuring marker genes was used to phylogenetically assess environmental sequencing data. This set of genes has been described previously (Ciccarelli et al. 2006), and has been chosen based on systematic searches of fully sequenced genomes: the genes were required to be universally present in all genomes known to date (including *Archaea* and *Eukaryotes*), and were selected such that the average number of paralogous copies in each genome was as low as possible. The rationale behind this choice is that such genes are apparently under strong selection against both gene loss and copy number variations. This should make them least likely to tolerate horizontal gene transfer (since horizontal transfer presumably entails episodes of either gene-absence or multiple gene copies); such genes should therefore be most likely to represent species phylogeny. Some remaining cases of horizontal transfer have been detected manually; these have been neutralized by artificially pruning marker genes from the affected organisms (such that in these organisms, the genes are considered 'missing data' in subsequent analyses (Ciccarelli et al. 2006)). Likewise, paralogs and additional gene copies derived from organelles were removed, until each gene family was represented by no more than a single, full-length sequence in each reference organism (Ciccarelli et al. 2006).

The set of marker gene families described above (mainly genes related to protein translation) is available here: http://MLtreemap.embl.de/treemap_html/marker_genes.txt. The list refers to proteins grouped into 'COGs' (clusters of orthologous groups); COGs were originally created by Tatusov and Koonin (Tatusov et al. 1997). We used an extended version of the COG database, maintained at the STRING website (http://string.embl.de/), which covers more organisms (von Mering et al. 2005). The environmental metagenomics data sets used were exactly those described in detail previously (Tringe et al. 2005).

Marker genes were detected among the environmental sequences using BLAST (searching predicted genes from the various data sets against the extended COG database maintained at the STRING server (Tatusov et al. 2003; von Mering et al. 2005)). COG-matches were called for any gene whose first hit was a protein assigned to a COG in STRING, as long as the BLAST score was better than 60 bits (multiple COG-mappings were allowed for single proteins, unless they were overlapping by more than 50% of their length). Each open reading frame found to map to one of the marker gene COGs was then re-aligned to all known members of that COG using HMMALIGN (Durbin et al. 1998). In cases where a single DNA fragment (read, contig or scaffold) contained more than one marker gene, these alignments were concatenated. Finally, gaps in the alignment were removed using GBLOCKS, with the following settings: *Maximum Number Of Contiguous Nonconserved Positions*: 15; *Minimum Length Of A Block*: 3; *Allowed Gap Positions*: with half; *Minimum Number Of Sequences For A Flank Position*: 55% of the Sequences. Depending on the degree of genome assembly that has been performed for a dataset, the number of informative columns contained in the final concatenated alignment of each fragment varies (ranging from an average of 180 residues in

the case of the Soil data, to 250 residues in the case of the Acid Mine Drainage data). It should be noted that this number of residues is not sufficient to build a phylogenetic tree *de novo*, but should contain enough information to place a single sequence into an existing tree (see also below).

We used all relevant marker proteins that were annotated in a read or contig, concatenated as a single amino acid sequence, in order to make use of all the available information. In addition, we assembled some contigs into larger 'scaffolds' prior to maximum likelihood scoring, when we found a) contigs to be connected by mate-pair information and b) both contigs to have marker genes annotated. This was done in a conservative manner: large contigs were joined only when three or more independent mate-pairs suggested a linkage; smaller contigs required fewer mate-pairs (two or more mate pairs in cases where maximally seven clones contributed sequences to the contigs, and one or more mate pairs in cases where maximally five clones contributed sequences). Apart from this joining procedure based on mate-pair information, no other grouping of contigs into scaffolds was performed beyond what had already been done in the original publications of the data. In the case of eukaryotes, this means that most of the sequence fragments contain only a single marker gene (because eukaryotes generally do not cluster translation-related genes in their genome). This potentially results in a lower precision of the placements of eukaryotes, especially for datasets where assembly of reads into contigs was possible and thus several marker genes can be found in prokaryotic contigs, but not in eukaryotic contigs. However, eukaryotes make up only a very small fraction of the datasets studied here (see Figure 2). In addition, they are often better amenable to alternative forms of taxonomic classification, due to their greater phenotypic variability, and are thus less dependent on sequence characters in general.

The actual distribution of detected marker genes among the environmental DNA fragments is available for detailed inspection and download, at
http://MLtreemap.embl.de/treemap_html/marker_gene_overview.html

Maximum likelihood scoring

After the above step, each environmental DNA fragment with at least one marker gene is represented by a multiple sequence alignment (this alignment contains the concatenated sequences of the relevant marker gene family/families, including the novel sequence(s) to be tested). For all the known sequences in the alignment, their phylogenetic relations are assumed to be that of an externally provided reference phylogeny of complete genomes (Ciccarelli et al. 2006). The novel sequence (the 'query') could in principle be branching anywhere in that tree. The possible branching positions effectively define an *ensemble* of trees, which are all identical except for the position of the query sequence. We analyzed these ensembles using TREE-PUZZLE (Schmidt et al. 2002), in the context of the above alignment, employing the same maximum likelihood model (and settings) as were used to generate the reference phylogeny itself: *substitution model: JTT; model of rate heterogeneity: Gamma distributed rates (4 categories).* Settings specific to TreePuzzle were: *parameter estimation: approximate (using quartet sampling + neighbor-joining tree).* This procedure resulted in a

maximum likelihood score for each tree in the ensemble, and the most likely tree then defined the most probable placement of the query sequence.

Often, however, more than one placement in the reference tree is possible, and these can be almost equally likely – especially in the case of short (or partial) query sequences, which may not contain enough phylogenetic information. We employed two measures to avoid unjustified precision when assigning such sequences: firstly, we required a minimum length of informative sequence in each query – this cutoff was set at 80 columns of blocked alignment (shorter queries were not considered; we balanced here the need for precision, with the requirement of having enough query sequences, even in unassembled, single-read metagenomics data). Secondly, we assigned queries to more than one position in the reference tree if necessary (giving them a fractional weight at each position). To do this, we used the 'expected likelihood weight distribution' defined by (Strimmer et al. 2002); this distribution takes into account any differences in likelihood, as well as possible mis-specifications of the substitution model or the reference tree (in order to increase reproducibility, we set the number of samplings in the expected likelihood weight algorithm to 100,000).

The final result of the above step is a likely placement of each query sequence in the reference tree (broken down into a weighted distribution of placements if necessary). Note that the branching pattern of the reference phylogeny itself is never altered – only the novel sequences are assessed, relative to the fixed reference phylogeny. This limits the amount of computation that is necessary: each relevant DNA fragment can be assigned within a few hours on a 64bit CPU (maximally a few days, for cases when the alignment is very long).

Aggregation and visualization

We computed the maximum likelihood mapping for all those DNA fragments of a sample which were found to contain one or more of the marker genes. All these fragments were then weighted by their assembly depth – such that the final measurement corresponds to the distribution of organisms in the sample (deeply assembled fragments represent more organisms, and are therefore given more weight; in the case of data sets for which assembly information was not available we approximated it by mapping the raw sequence reads against the assembly using BLAT). We also normalized fragments by the length of their marker gene alignment (each organism should contain roughly one complement of marker genes, so a longer alignment corresponds to a larger part of an organism). All placements were then added up, divided by the total, and visualized in the context of the reference tree, using in-house tree drawing software (see Figures 2 and S1). Note that while TREE-PUZZLE requires rooted trees as input, the visualization software shows unrooted trees; however, the root received very little placements, and these were omitted for the visualization. For Figure S1, the placements were additionally 'projected' onto the reference taxa (as bar-charts, merely for illustration): each placement was distributed among the reference taxa which were descendents of the placements' branching position, dividing the weight evenly at each bifurcation in the tree while proceeding from the actual placement up to the tips of the tree.

We always aggregated all relevant sequences of an environmental sample, with one exception: in the case of the Sargasso Sea data, a small number of environmental sequences were removed from consideration because of a reported contamination (DeLong 2005): we removed all contigs which were predominantly derived from Sargasso sample no. 1, and which in addition had been placed by our procedure into the clades *Burkholderiales* or *Shewanellaceae*; these two clades are the presumed contaminants (DeLong 2005). This filter removed about 0.9% of the relevant Sargasso Sea sequences.

In Figure 2, we visualize the taxonomic distributions of four distinct environments onto the same tree image. Color codes are used to distinguish the environments. In addition, we use the same color codes to visualize which environment is the 'preferred habitat' for a particular section (i.e. line) in the tree of life. These colors are shown with various intensities, in order to visualize the extent of preference. No color is shown when the clade in question is found equally frequent in all four environments, or not at all; maximum color is shown when it is found exclusively in one single environment. In order to avoid spurious signals in the case of rarely detected clades, pseudocounts were added to each of the four environment counts before determining line color (1 % of the total placement were added to each count). The line color saturation is then proportional to $\Sigma(\mathrm{abs}(\log((\mathrm{count}_i/\mathrm{total\_count})*n)))$, summing over all environments i, where n is the total number of environments.

Validation, Comparison with BLAST

For validation, we repeated the above mapping procedure with altered reference trees (i.e. some clades in the reference tree were intentionally omitted – all the placements they originally received were expected to re-appear elsewhere: namely, basal at the next available sister clade). We compared the placements for each altered reference tree to the original placements in the unaltered reference tree, and recorded the average deviation from the expected, 'ideal' behavior – this deviation can be expressed quantitatively in units of branch lengths, and provides a measure by which to rank the relative performance of various methods or parameter choices (see Figure S2). We contrasted the performance of our method to two frequently used BLAST-based mapping techniques; these were simple 'best-hit' approaches wherein each environmental fragment is placed directly at the reference taxon to which it has the best BLAST score: An open reading frame is simply declared to be originating from the same phylum (or class, order, …) as the taxon in the database to which is has the highest BLAST score. For metagenomics contigs, the added complication is that a single DNA fragment can have several open reading frames. We chose to consider all predicted proteins that were mapped to any known COG (in order to exclude spurious ORFs), and searched those against the same set of reference genomes as used in the maximum likelihood placement (alternatively, we restricted the analysis to the same set of 31 COGs as used in our marker gene set). In the case of multiple valid proteins per contig, we simply added up all the scores of a contig, for each species in the database, and chose the species with the highest cumulative score as the 'best hit'.

Variations in apparent evolutionary speed

Each maximum likelihood placement provides not only a most likely branching position for the query sequence, but also branch length information – indicating the approximate amount of sequence changes that have accumulated since the query sequence branched from the reference tree. This enables estimates about the evolutionary rate in an environment. We have assessed branch lengths for each query sequence (using the 'expected likelihood weights' to weigh multiple alternative placements if necessary). Branch lengths were expressed as distances to the root of the tree (the root was determined by mid-point rooting), and were found to be significantly different between environments (data not shown). However, the exact position of the root in the reference tree is uncertain, and can have a large effect on the result when one of the samples is dominated, for example, by *Archaea*, while the other is dominated by *Bacteria.* Therefore, we sought a more objective baseline for branch length measurements. We decided to compare branch lengths only within phyla (i.e. branch lengths were measured from the tips to the base of the phylum) and to use sequenced relatives from the reference tree for comparison: each query sequence was compared to all sequenced relatives of the same phylum. For query sequences that were placed basal to existing phyla (presumably from phyla not yet sequenced) the comparison was done to all sequenced genomes in the immediate sister clade. It should be noted that the above procedure is not influenced by differences in the underlying evolutionary rate of the 31 marker gene families: firstly, each environmental sequence fragment is tested exclusively in the context of its alignment to other genes of the same family (e.g. SecY genes are always compared to other SecY genes), and secondly, the overall occurrence of the 31 marker genes is purely stochastic (they all occur once per genome, and are sampled randomly by the shotgun procedure) so any differential rates among the families should affect each environment in the same way.

Robustness estimation

A set of marker genes is advantageous over a single marker, since a larger fraction of the sequences will be informative. Nevertheless, because the markers are not enriched, the amount of raw sequence data needed is fairly high. We estimated the robustness of our placements, with respect to potential under-sampling, using both jackknife and bootstrap approaches (Figure S3). Jackknife analysis of the smallest dataset (the whale bone sample) revealed that the overall placement pattern remains stable, even when only a random subset of 50% of the data is used. Nevertheless, small variations in the lower-abundance placements became visible. In order to quantify these, we performed one hundred bootstrap tests for each of the datasets. These revealed that in the worst case (low abundance items in the smallest dataset) the average quantitative error in the placement is about 50% (Figure S3). Higher abundance items (like those discussed in the text) have correspondingly lower errors. The bootstrap tests also provide confidence intervals for all quantitative statements in the paper, such as for the statement "roughly 1% endospore formers (bacilli and clostridia) in the soil"; the 95% confidence interval for this statement is "0.995% to 2.153% endospore formers in the soil". Similarly, for the amount of Actinobacteria it is "4.2% to 8.2%", with a median of 6.3%.

Habitat stability

16S rRNA sequences from the 'Greengenes' database (Desantis et al. 2006), as well as collective strain information from public microbial culture collections (Dawyndt et al. 2005), were parsed for their descriptions of sampling sites (habitats). We compared strains (or rRNA isolates) pair wise, and first determined their level of relatedness, as follows. In the case of cultivated strains, we parsed their assigned position in the NCBI taxonomy, and assessed the taxonomic level at which they were related (this level is defined by the last term they share in their lineage descriptions). In the case of rRNA sequences, we measured the node-to-node branch length distance in a published global phylogeny of small subunit rRNA sequences (Desantis et al. 2006)). This global phylogeny of rRNA sequences has been initially built from a core of 500 taxa, and has been subsequently extended by the insertion of thousands of additional sequences using the 'ARB parsimony insertion tool' (this tool inserts additional sequences without changing the topology of the initial tree, and estimates branch lengths so as to roughly reflect the degree of sequence divergence). From this tree, we derived a 'phylogenetic distance' for any pair of rRNA isolates by traversing between their two positions in the tree, through their last common ancestor, and then summing up all branch lengths encountered in between.

After having determined the relatedness of any two given isolates, we then assessed the pair wise similarity of their assigned habitats by automated detection of shared keywords in the habitat descriptions (in the case of strains, keywords were generalized using an ontology relating terms to higher-level habitat categories, see Table S2). Keywords were weighted according to their frequency among habitat annotations (rare keywords gave more signal then ubiquitous keywords), thus habitat similarity was expressed in terms of shared keywords as follows: $\mathbf{s = 1 - \Pi(f_{keyword})}$, where $f_{keyword}$ is the fraction of habitats that contain this keyword. If not a single keyword was shared, the similarity was set to zero. For rRNA sequences, we only compared clones between experiments (not within experiments, so as to avoid researcher method/annotation biases within individual experiments).

Comparison to PCR-based approaches

Traditionally, environmental microbes are assessed taxonomically by cloning and sequencing their ribosomal RNA genes (most notably the 16S/18S rRNA). In early studies, reverse transcription or hybridization was used to enrich for rRNA sequences (Ward et al. 1990; Schmidt et al. 1991), but polymerase chain reaction (PCR) quickly became the method of choice. PCR has the advantage of requiring little starting material; it circumvents labor-intensive clone selection protocols, and it is independent of the expression level of rRNA genes (since genomic DNA is the template for amplification). Many PCR-based studies have since been undertaken, using samples from natural and man-made environments (Cole et al. 2005; Giovannoni et al. 2005; Robertson et al. 2005). Collectively, these studies have shown that microbial diversity is far greater than previously appreciated, revealing the existence of

more than 50 phylum-level bacterial lineages (many of which have no cultivated representatives to date) (Hugenholtz et al. 1998).

However, PCR amplification is also a well-known source of qualitative and quantitative error (von Wintzingerode et al. 1997). Firstly, the result of PCR amplifications cannot generally be assumed to be a proportional representation of the starting material – especially not in the case of a complex mixture of rRNA genes. Primers will bind with various strengths, and the processivity and efficiency of polymerization will depend on the primary sequence and its GC-content. PCR-products may also re-anneal with potential templates in the next cycle, a process leading to the progressive inhibition of the more frequent genotypes. This 'non-linearity' of PCR amplification for rRNA genes has been quantified, and shown to be dependent on the primers and protocols used (Suzuki et al. 1996).

Secondly, a notorious problem of PCR is the unintended priming of reactions by incomplete (truncated) products of previous PCR cycles. In the case of mixed templates, this mechanism can lead to a substantial fraction of amplified molecules being chimeric, i.e. the N-terminal part may have a different origin than the C-terminal part. The fraction of chimeras can exceed 30% (Wang et al. 1997); this represents a serious problem when propagated into databases, because chimeras may be erroneously annotated as 'novel' clades or phyla (Hugenholtz et al. 2003; Ashelford et al. 2005).

Lastly, the biology of rRNA genes also holds some problems for quantitative taxonomic surveys: rRNA genes are known to occur with widely varying copy numbers in genomes (ranging from a single copy to as many as fifteen (Rainey et al. 1996; Klappenbach et al. 2001)). In the case of uncultivated organisms, the copy number status is often not known – making quantitative inferences about the number of individual cells of a particular taxon very imprecise (Farrelly et al. 1995). In addition, rRNA genes may even exhibit some phylogenetic instability: Some bacterial genomes have been observed to contain divergent 16S genes with as much as 5% sequence differences; and occasional discrepancies between the phylogenies of 16S genes and other genes in the same genome have been reported (Dennis et al. 1998; Yap et al. 1999; Badger et al. 2005).

For one of the datasets studied here (the soil data), the traditional PCR-based assessment was executed in parallel to the shotgun sequencing. This enabled us to compare our assignments to the 16S/PCR-based results. Overall, the relative abundances of phyla reported by the two approaches were correlated, but not very strongly (Table S1, the R^2 value is only 0.40). This is partly due to the fact that both approaches missed some phyla entirely: 15% of the PCR derived assignments were to phyla not yet represented among sequenced genomes; conversely, 11% of the maximum likelihood placements were absent in the PCR data (for example, the primers used in this PCR setup were not applicable to *Archaea* and *Eukaryotes*). In this comparison, it appears that the metagenomics data are closer to the truth: the rRNA sequences that are contained in the shotgun sequences themselves (i.e. were obtained without PCR) show a distribution among phyla that agrees much better with our maximum likelihood placements (Table S1, R^2-value is 0.73, $p<0.05$), which indicates the extent of the PCR bias.

However, the method we present here has still some shortcomings, as well. The most important of these relates to the limited availability of completely sequenced genomes, especially from non-cultivatable free-living organisms. For phyla that are not yet represented at all among sequenced genomes, the placement of environmental fragments is naturally rather imprecise (although the distinction between Bacteria, Archaea and Eukaryotes is almost always made correctly). However, once the phylum of a sequence fragment is represented among the reference genomes at least once, the fragment is usually placed correctly (see Figure S2). This is even the case for unassembled single reads (such as in the soil data), as long as these have been derived using Sanger sequencing, with read lengths approaching 1000 nucleotides.

The use of shotgun sequencing followed by maximum likelihood analysis, to assess the phylogenetic distribution of organisms in a habitat, is also far more resource-intensive than a standard PCR-based analysis. It requires more sequencing, and also a substantial amount of computation time (up to several CPU hours per fragment). This is partially offset, however, by the fact that the information can be directly extracted from the metagenomics datasets (which are usually derived for a different purpose). As not only phylogenetically informative genes are being sampled, but all gene classes, functional insights into the community can be coupled to the phylogenetic assessment. In the long run, this will allow a quantification of the coupling of cellular processes to organisms and lineages. Overall, shotgun sequencing potentially offers tremendous advantages over the sequencing of a single phylogenetic marker, especially given the projected growth in completely sequenced reference genomes, and given the expected maturation of algorithms to assemble, annotate and interpret the data.

	PCR-based analysis (16S/18S rRNA genes)	Metagenomics-based analysis (31 protein-coding marker genes)
Advantages	- established procedure - many reference sequences - cost effective - can be performed in small lab	- no amplification biases - no primer dependency - functional genes sampled in parallel - more residues per genome tested - very little copy number variation
Disadvantages	- PCR amplification biases - chimera formation during PCR - non-universality of primers - rDNA copy number variations - no functional genes sampled in parallel	- less reference sequences (genomes) - requires deeper sequencing - computationally intensive - better with long reads, and at least some partial assembly - high throughput sequencing usually requires a sequencing center

Comparison: A short summary of advantages and disadvantage of the two methods for sequence-based phylogenetic assessment of natural microbial communities.

Supplementary References

Ashelford, K. E., N. A. Chuzhanova, et al. (2005). "At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies." Appl Environ Microbiol **71**(12): 7724-36.

Badger, J. H., J. A. Eisen, et al. (2005). "Genomic analysis of Hyphomonas neptunium contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and Caulobacterales." Int J Syst Evol Microbiol **55**(Pt 3): 1021-6.

Ciccarelli, F. D., T. Doerks, et al. (2006). "Toward automatic reconstruction of a highly resolved tree of life." Science **311**(5765): 1283-7.

Cole, J. R., B. Chai, et al. (2005). "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis." Nucleic Acids Res **33**(Database issue): D294-6.

Dawyndt, P., M. Vancanneyt, et al. (2005). "Knowledge Accumulation and Resolution of Data Inconsistencies during the Integration of Microbial Information Sources." IEEE Transactions on Knowledge and Data Engineering **17**: 1111-1126.

DeLong, E. F. (2005). "Microbial community genomics in the ocean." Nat Rev Microbiol **3**(6): 459-69.

Dennis, P. P., S. Ziesche, et al. (1998). "Transcription analysis of two disparate rRNA operons in the halophilic archaeon Haloarcula marismortui." J Bacteriol **180**(18): 4804-13.

Desantis, T. Z., P. Hugenholtz, et al. (2006). "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB." Appl Environ Microbiol **72**(7): 5069-72.

Durbin, R., S. Eddy, et al. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK, Cambridge University Press.

Farrelly, V., F. A. Rainey, et al. (1995). "Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species." Appl Environ Microbiol **61**(7): 2798-801.

Giovannoni, S. J. and U. Stingl (2005). "Molecular diversity and ecology of microbial plankton." Nature **437**(7057): 343-8.

Hugenholtz, P., B. M. Goebel, et al. (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." J Bacteriol **180**(18): 4765-74.

Hugenholtz, P. and T. Huber (2003). "Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases." Int J Syst Evol Microbiol **53**(Pt 1): 289-93.

Klappenbach, J. A., P. R. Saxman, et al. (2001). "rrndb: the Ribosomal RNA Operon Copy Number Database." Nucleic Acids Res **29**(1): 181-4.

Rainey, F. A., N. L. Ward-Rainey, et al. (1996). "Clostridium paradoxum DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences." Microbiology **142 (Pt 8)**: 2087-95.

Robertson, C. E., J. K. Harris, et al. (2005). "Phylogenetic diversity and ecology of environmental Archaea." Curr Opin Microbiol **8**(6): 638-42.

Schmidt, H. A., K. Strimmer, et al. (2002). "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics **18**(3): 502-4.

Schmidt, T. M., E. F. DeLong, et al. (1991). "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing." J Bacteriol **173**(14): 4371-8.

Strimmer, K. and A. Rambaut (2002). "Inferring confidence sets of possibly misspecified gene trees." Proc Biol Sci **269**(1487): 137-42.

Suzuki, M. T. and S. J. Giovannoni (1996). "Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR." Appl Environ Microbiol **62**(2): 625-30.

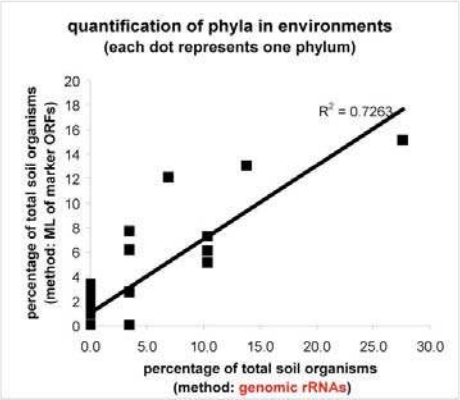
Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

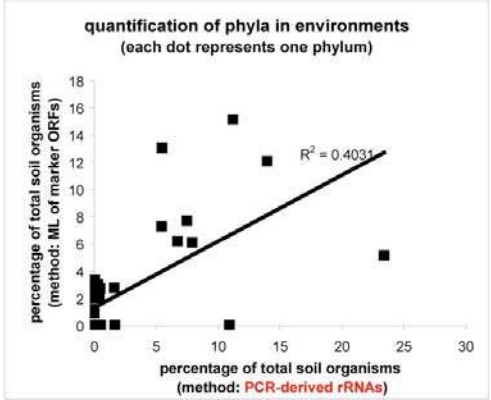
Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-7.

Tringe, S. G., C. von Mering, et al. (2005). "Comparative metagenomics of microbial communities." Science **308**(5721): 554-7.
von Mering, C., L. J. Jensen, et al. (2005). "STRING: known and predicted protein-protein associations, integrated and transferred across organisms." Nucleic Acids Res **33**(Database issue): D433-7.
von Wintzingerode, F., U. B. Gobel, et al. (1997). "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis." FEMS Microbiol Rev **21**(3): 213-29.
Wang, G. C. and Y. Wang (1997). "Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes." Appl Environ Microbiol **63**(12): 4645-50.
Ward, D. M., R. Weller, et al. (1990). "16S rRNA sequences reveal numerous uncultured microorganisms in a natural community." Nature **345**(6270): 63-5.
Yap, W. H., Z. Zhang, et al. (1999). "Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon." J Bacteriol **181**(17): 5201-9.

Minnesota Farm Soil - three ways to quantify taxa in the sample

phylum	phyla counts in genomic 16S	fraction genomic 16S (%)	clone counts in PCR-derived 16S	fraction in PCR 16S (%)	fraction in ML placement (%)
Acidobacteria	2	6.9	293	14.0	12.084
Actinobacteria	1	3.4	141	6.7	6.184
Alphaproteobacteria	8	27.6	235	11.2	15.149
Aquificae	0	0.0	0	0.0	0.906
Archaea	0	0.0	0	0.0	1.211
Bacteroidetes	4	13.8	115	5.5	13.054
Betaproteobacteria	1	3.4	157	7.5	7.704
BRC1	0	0.0	2	0.1	0
Chlamydiae	0	0.0	0	0.0	2.134
Chlorobi	0	0.0	9	0.4	2.064
Chloroflexi	3	10.3	491	23.4	5.131
CM11b	0	0.0	1	0.0	0
Cyanobacteria	1	3.4	10	0.5	2.73
Deinococci	0	0.0	1	0.0	3.386
Deltaproteobacteria	3	10.3	166	7.9	6.091
Endomicrobia	0	0.0	4	0.2	0
Epsilonproteobacteria	0	0.0	0	0.0	0.93
Eukaryota	0	0.0	0	0.0	1.214
Fibrobacteres	0	0.0	1	0.0	2.031
Firmicutes	0	0.0	6	0.3	3.022
Fusobacteria	0	0.0	0	0.0	1.175
Gammaproteobacteria	3	10.3	114	5.4	7.278
Gemmatimonadetes	1	3.4	229	10.9	0
Nitrospirae	0	0.0	11	0.5	0
NKB19	0	0.0	3	0.1	0
OD2	0	0.0	6	0.3	0
OP10	0	0.0	8	0.4	0
OP11	0	0.0	1	0.0	0
OP11-5	0	0.0	10	0.5	0
OP3	0	0.0	1	0.0	0
Planctomycetes	1	3.4	34	1.6	2.787
SC3	0	0.0	2	0.1	0
SPAM	0	0.0	2	0.1	0
Spirochaetes	0	0.0	0	0.0	2.606
Thermotogae	0	0.0	0	0.0	1.118
TM7	0	0.0	4	0.2	0
UB12	0	0.0	3	0.1	0
Verrucomicrobia	1	3.4	35	1.7	0
WS3	0	0.0	4	0.2	0
TOTAL	29	100.0	2099	100.0	99.989





orange fields: not applicable (for technical reasons, but still included in graphs [as zero counts])

the difference in R^2-values is statistically significant ($p < 0.05$; performing 10,000 resamplings)

Table S1:
Correlation between rRNA(16S)-based and metagenomics-based taxonomic classifications, for the soil sample. The placements we propose here (maximum likelihood) agree better with the genomic rRNAs than with the PCR-derived rRNAs.

category	Keywords
terrestrial	soil, sediment, garden, forest, grass, rhizosphere, sediments
aquatic	water, spring, sea, pond, lake, river, seawater, marine, brackish, lagoon, freshwater, ocean
extreme	acid, hydrothermal, thermophilic
internal	blood, human, feces, faeces, urine, sputum, tract, intestine, mouth, vagina, lung, stool, dental, fecal, rumen, intestine, host, manure, dung, spinal, throat, serological, gastric, vaginal, bronchial, mucosa
foodstuff	cheese, milk, beer, fruit, vinegar, brewery, meat, food, apple, yoghurt, sausage, mushroom, sake, cheddar
sewage_and_others	mud, sludge, sewage, waste, silage, wastewater

Table S2:
Keywords used in assessing the habitats of strains, together with their high-level categorization. The categories were used for the actual habitat similarity measure, in order to recognize ‘seawater’ and ‘marine’ as broadly similar habitats.

acid	14
acid.mine	12
activated.sludge	21
anaerobic.sludge	5
aquifer	5
bog	5
cecum	5
contaminated.soil	9
deep.sea	20
distilled.water	6
drainage	13
EBPR	5
EBPR.sludge	5
feces	22
field	12
forest	5
freshwater	12
geothermal	6
goat	8
goat.rumen	8
gold.mine	5
granular.sludge	7
grassland	5
grassland.soil	5
ground	5
groundwater	11
gut	25
gut.homogenate	6
hindgut	6
host	7
hot.spring	12

human	24
human.mouth	6
hydrothermal	30
hydrothermal.vent	18
hypersaline	5
ice	9
intestine	14
Lake	32
lake.water	7
landfill	5
marine	36
marine.sediment	20
mine	23
mine.drainage	13
mining	10
mouth	8
mud	8
mud.volcano	5
Ocean	10
oil	12
paddy.soil	5
peat	5
piglet	6
plant	10
pond	7
reactor	23
rhizosphere	7
rice	11
River	11
rock	6
rumen	16

Sargasso	5
Sargasso.Sea	5
Sea	45
sea.hydrothermal	10
sea.sediment	7
Seamount	5
seawater	9
sediment	96
seep.sediment	5
sludge	54
soil	84
sponge	5
spring	19
subsurface	10
termite	19
termite.gut	12
terrestrial	5
treatment.plant	5
Trough	6
uranium.mill	6
uranium.mining	10
vent	19
volcano	6
waste	17
wastewater	19
wastewater.treatment	6
water	63
waterbath	6

Table S3:
Keywords used in assessing the habitats of rRNA clones. The keywords were derived by a manual scan of all words that were found to be used in the annotation of at least five rRNA isolation experiments. Of those words, terms were kept that denote lifestyle or habitat (but not uninformative words such as ‘from’, ‘inside’ or ‘surface’).

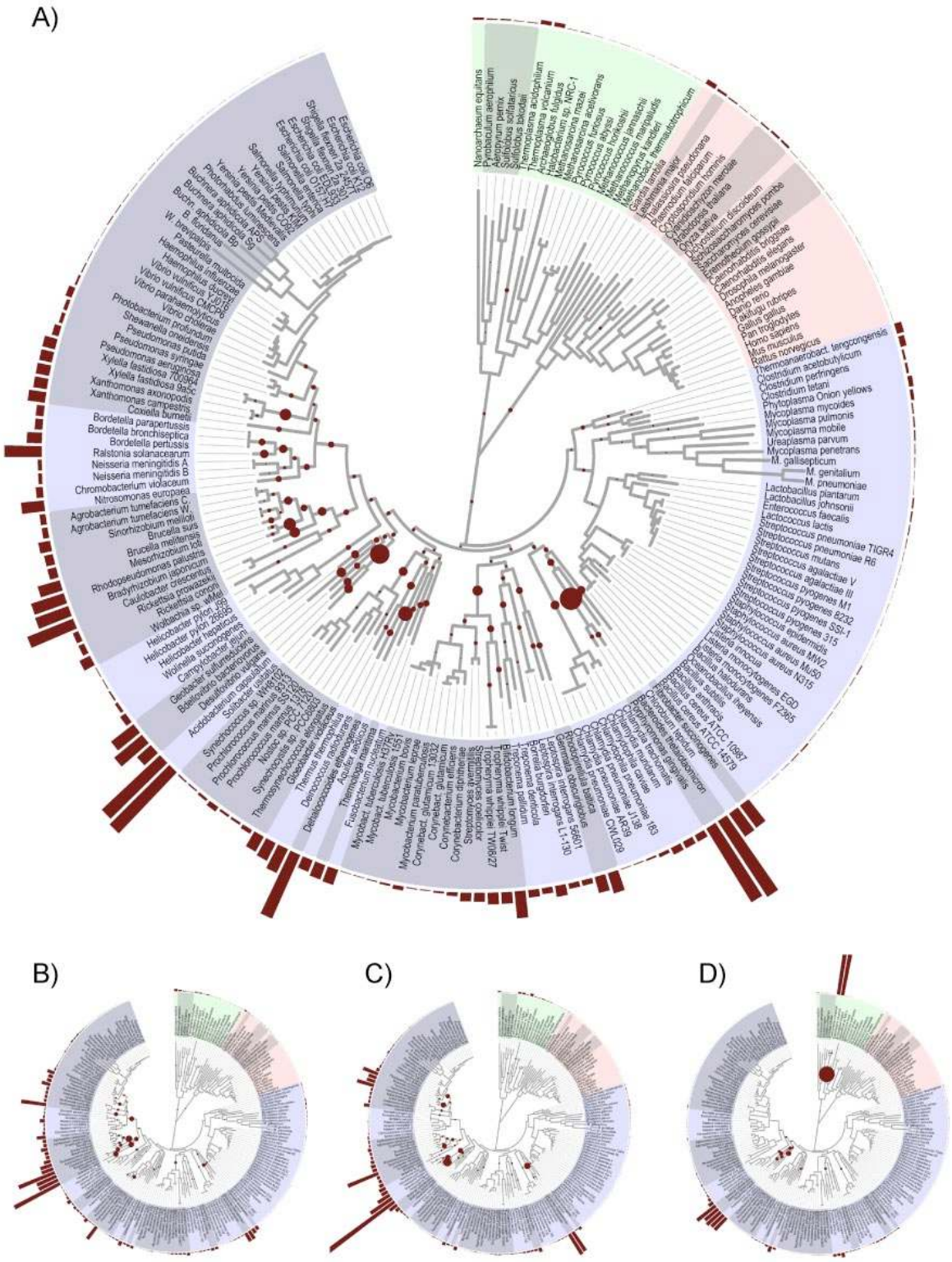
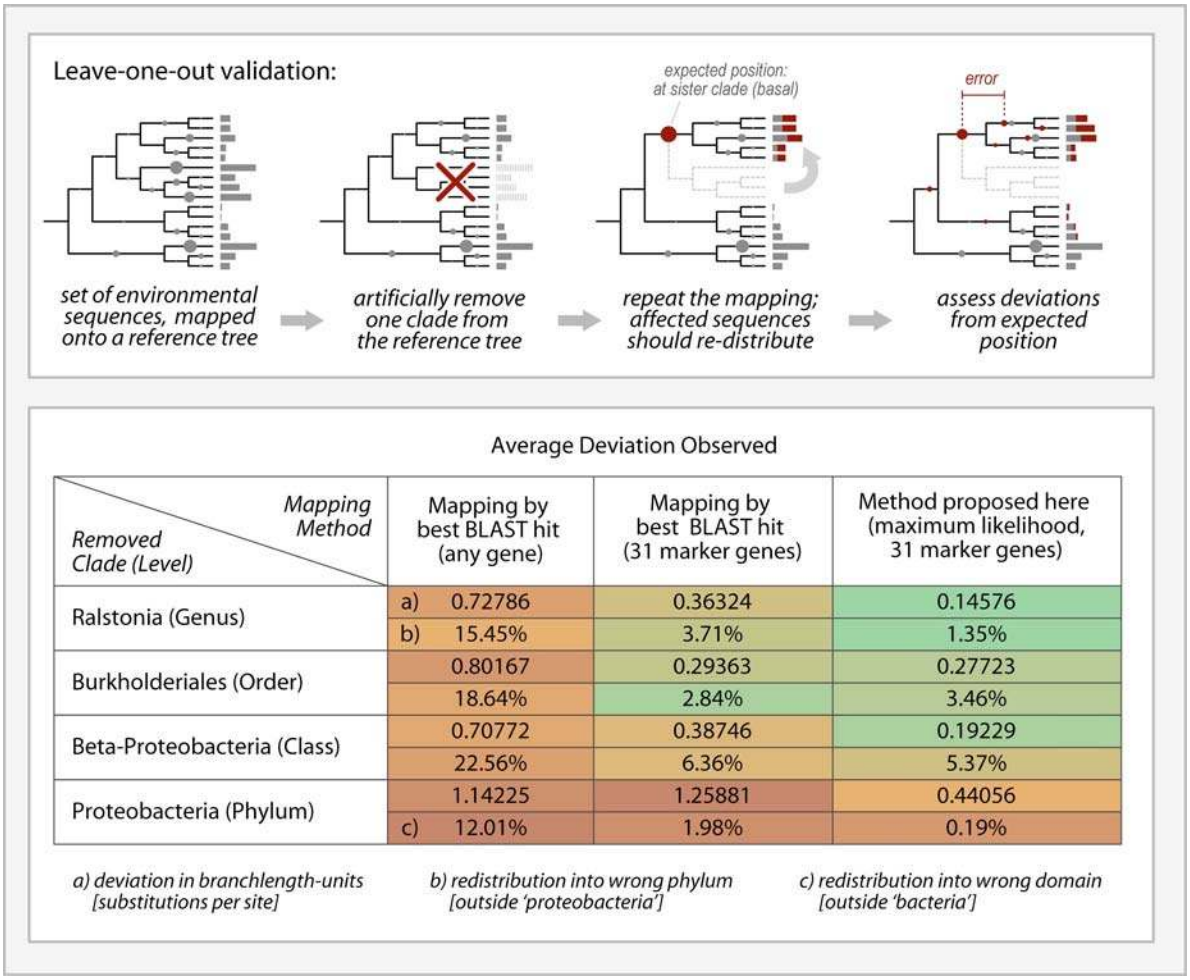


Figure S1:
Phylogenetic distribution of communities, separately for each environment. A) Agricultural Soil. B) Surface Ocean Water. C) Deep Sea Whale Bone. D) Acidic Mine Drainage. Higher-resolution versions of these images can be reached via the Online Supplement.



Average Deviation Observed

Removed Clade (Level) \ Mapping Method	Mapping by best BLAST hit (any gene)	Mapping by best BLAST hit (31 marker genes)	Method proposed here (maximum likelihood, 31 marker genes)
Ralstonia (Genus)	a) 0.72786	0.36324	0.14576
	b) 15.45%	3.71%	1.35%
Burkholderiales (Order)	0.80167	0.29363	0.27723
	18.64%	2.84%	3.46%
Beta-Proteobacteria (Class)	0.70772	0.38746	0.19229
	22.56%	6.36%	5.37%
Proteobacteria (Phylum)	1.14225	1.25881	0.44056
	c) 12.01%	1.98%	0.19%

a) deviation in branchlength-units [substitutions per site]

b) redistribution into wrong phylum [outside 'proteobacteria']

c) redistribution into wrong domain [outside 'bacteria']

Figure S2:

Validation, and comparison to BLAST-based mappings. Leave-one-out consistency checks show the performance and reliability of the mapping. Arbitrary parts (sub-clades) of the reference tree are removed – and any fragment previously placed there should now re-appear at the sister clade, in a basal position. The actual results will deviate from this ideal behavior, and the average distance to the ideal position on the tree provides an estimate on method error. The genus *Ralstonia* and its relatives were arbitrarily chosen as a test clade; the data set used for this test is the Minnesota farm soil sequence data. BLAST-based methods perform poor by comparison, even when restricted to the very same set of marker genes (for BLAST, placements are always at the tip of the tree – at the taxon showing the best BLAST score).

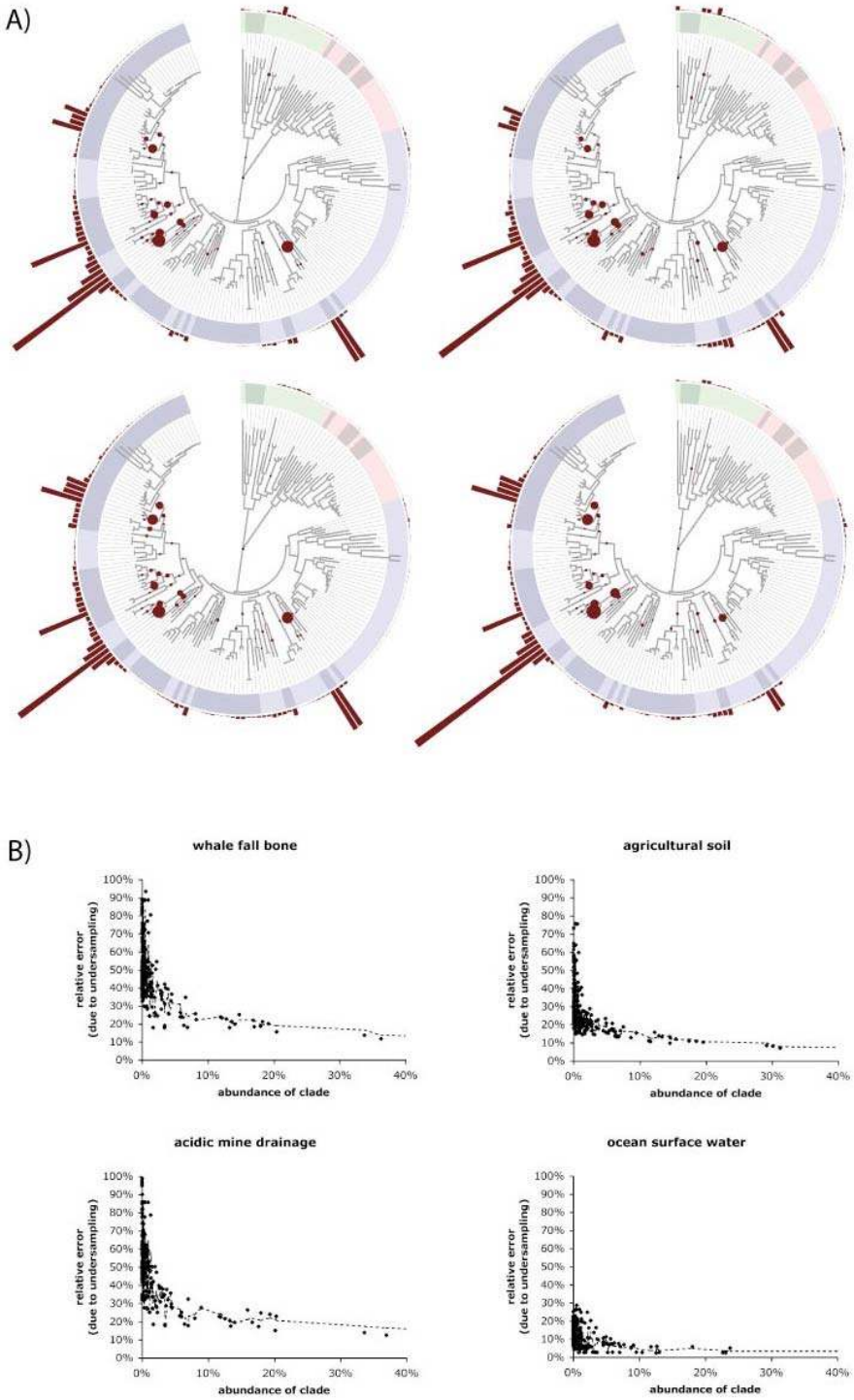


Figure S3:
Robustness estimation. A) jackknife testing of the smallest dataset (whale fall sample '#agzo'): four independent replicates using only 50% of the data were performed, showing good overall agreement but variation in the low abundance placements due to undersampling. B) bootstrap analysis. All four datasets used in Figure 2 were bootstrapped one hundred times (genes were sampled randomly, with replacement). For the smallest sample, this revealed an average relative error of 50.8% for low abundance items (abundance less then 1%), and 30.3% for all other items. For the largest sample (ocean water), average errors are as low as 12.5% and 7.4%, respectively. Each dot describes one possible clade in the tree, and the broken line is a running average of three.

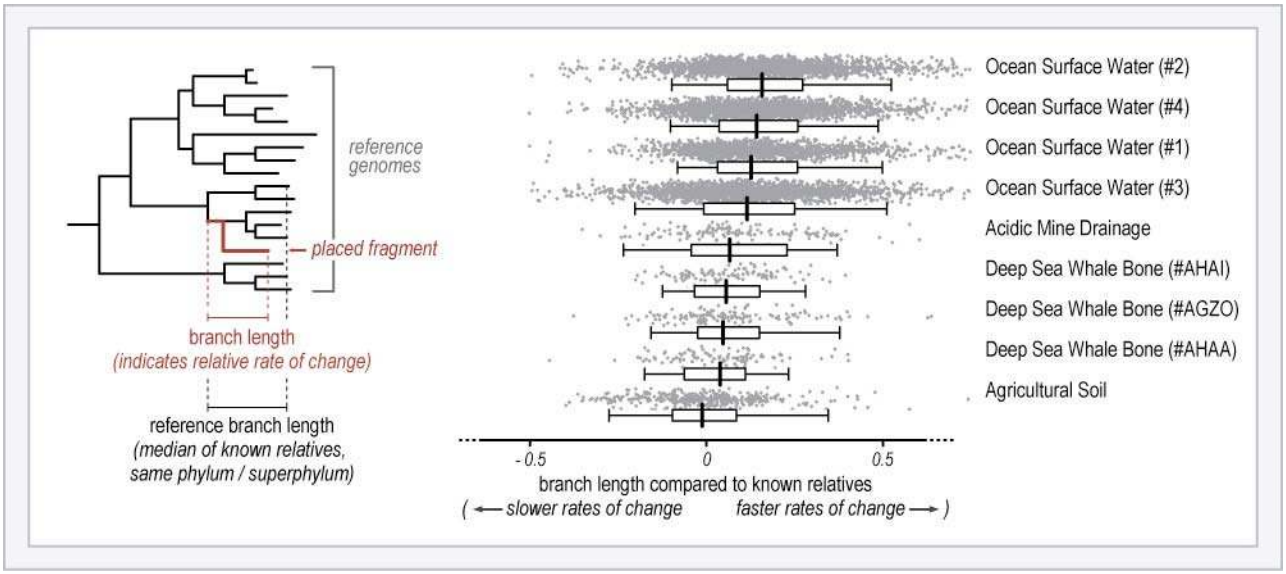


Figure S4:
Differences in evolutionary rates. This figure is equivalent to Figure 3 of the main text, but sub-samples are shown separately here. No sub-sampling is available for agricultural soil and acid mine drainage. For the ocean surface sample, assembled contigs were assigned to one of the four samples based on the source that contributed the majority of reads to the contig (the original assembly was done by pooling sequence reads from all four samples). As elsewhere in the paper, putative contaminants in the ocean water sample #1 were removed prior to analysis.