Sciencexpress Report

Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments

C. von Mering,[1]* P. Hugenholtz,[2] J. Raes,[1] S. G. Tringe,[2] T. Doerks,[1] L. J. Jensen,[1] N. Ward,[3] P. Bork[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]Institute for Genomic Research, Rockville, MD 20850, USA.

*Present address: University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.

†To whom correspondence should be addressed. E-mail: peer.bork@embl.de

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*) (Fig. 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 (fig. 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 (fig. 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 (Fig. 2 and fig. S1; see also fig. 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 (Fig. 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 (fig. 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% (fig. 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 (Fig. 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 (Fig. 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 (Fig. 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 (Fig. 2, B and 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 (Fig. 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.

References and Notes

1. W. B. Whitman, D. C. Coleman, W. J. Wiebe, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578 (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 (1990).
4. D. M. Ward, R. Weller, M. M. Bateson, *Nature* **345**, 63 (1990).
5. T. M. Schmidt, E. F. DeLong, N. R. Pace, *J. Bacteriol.* **173**, 4371 (1991).
6. N. R. Pace, *Science* **276**, 734 (1997).
7. P. Hugenholtz, B. M. Goebel, N. R. Pace, *J. Bacteriol.* **180**, 4765 (1998).
8. F. von Wintzingerode, U. B. Gobel, E. Stackebrandt, *FEMS Microbiol. Rev.* **21**, 213 (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 (2004); published online 4 March 2004 (10.1126/science.1093857).
11. S. G. Tringe *et al.*, *Science* **308**, 554 (2005).
12. G. W. Tyson *et al.*, *Nature* **428**, 37 (2004).
13. G. M. Luna, A. Dell'Anno, R. Danovaro, *Environ. Microbiol.* **8**, 308 (2006).
14. H. A. Barton, N. M. Taylor, B. R. Lubbers, A. C. Pemberton, *J. Microbiol. Methods* **66**, 21 (2006).
15. I. M. Kauffmann, J. Schmitt, R. D. Schmid, *Appl. Microbiol. Biotechnol.* **64**, 665 (2004).
16. F. D. Ciccarelli *et al.*, *Science* **311**, 1283 (2006).
17. K. Strimmer, A. Rambaut, *Proc. Biol. Sci.* **269**, 137 (2002).
18. J. A. Klappenbach, P. R. Saxman, J. R. Cole, T. M. Schmidt, *Nucleic Acids Res.* **29**, 181 (2001).
19. M. S. Rappe, S. J. Giovannoni, *Annu. Rev. Microbiol.* **57**, 369 (2003).
20. S. J. Giovannoni *et al.*, *Science* **309**, 1242 (2005).

21. F. U. Battistuzzi, A. Feijao, S. B. Hedges, *BMC Evol. Biol.* **4**, 44 (2004).
22. D. F. Feng, G. Cho, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13028 (1997).
23. H. Ochman, S. Elwyn, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12638 (1999).
24. S. Wright, J. Keeling, L. Gillman, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7718 (2006).
25. A. Dufresne, L. Garczarek, F. Partensky, *Genome Biol.* **6**, R14 (2005).
26. J. B. Martiny *et al.*, *Nat. Rev. Microbiol.* **4**, 102 (2006).
27. W. F. Doolittle, *Trends Cell Biol.* **9**, M5 (1999).
28. R. F. Doolittle, *Curr. Opin. Struct. Biol.* **15**, 248 (2005).
29. I. Chen, P. J. Christie, D. Dubnau, *Science* **310**, 1456 (2005).
30. N. U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, *Nature* **439**, 847 (2006).
31. C. Pedros-Alio, *Trends Microbiol.* **14**, 257 (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).

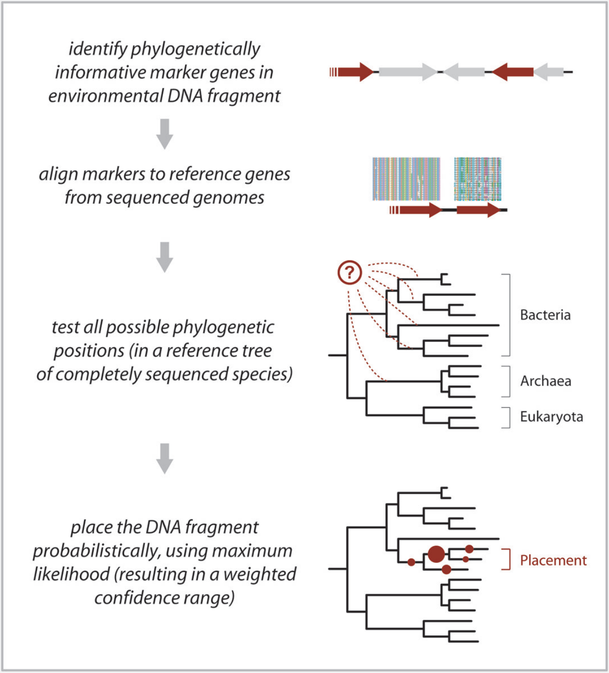
Supporting Online Material
www.sciencemag.org/cgi/content/full/1133420/DC1
SOM Text
Figs. S1 to S4
Tables S1 to S3
References

3 August 2006; accepted 19 January 2007
Published online 1 February 2007; 10.1126/science.1133420
Include this information when citing this paper.

Fig. 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.

Fig. 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 ($p < 10^{-6}$).

Fig. 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, $p \leq 10^{-5}$, 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.

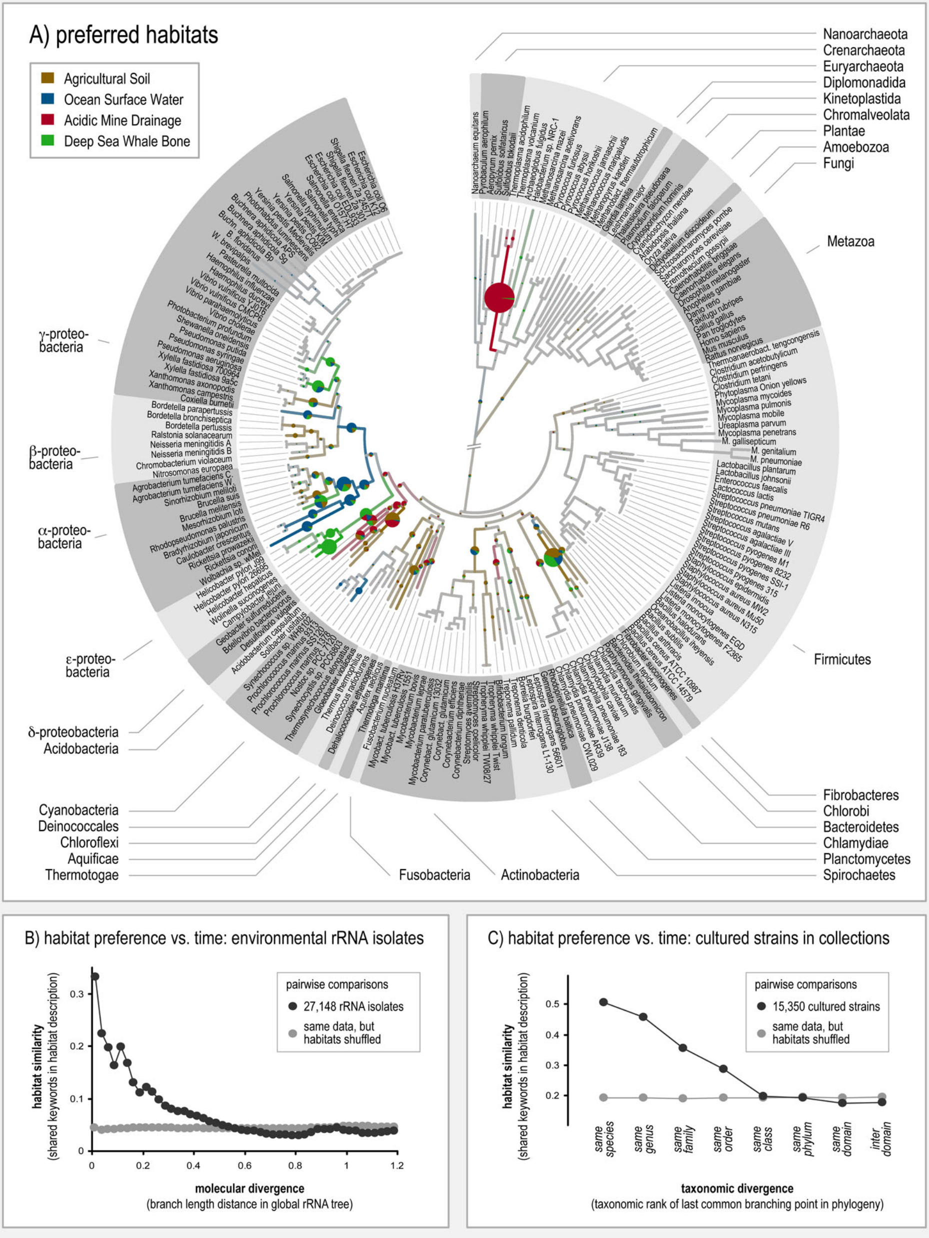


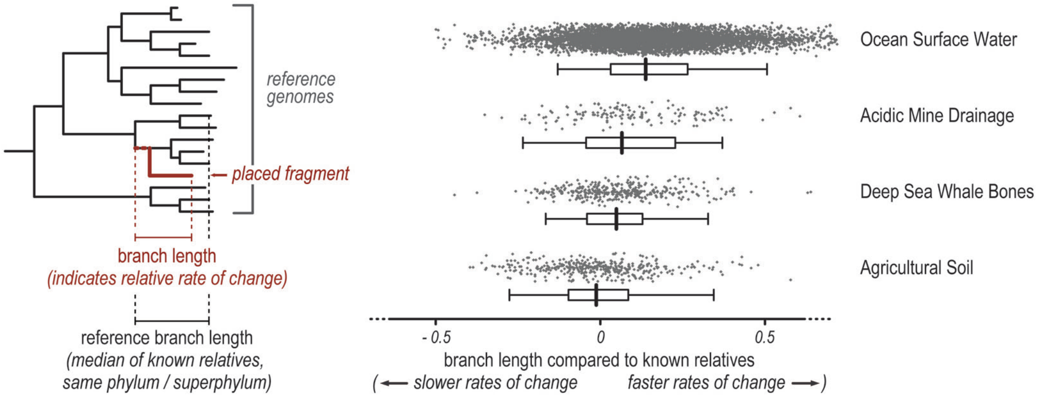
identify phylogenetically informative marker genes in environmental DNA fragment

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
Firmicutes
Fibrobacteres
Chlorobi
Bacteroidetes
Chlamydiae
Planctomycetes
Spirochaetes
Actinobacteria
Fusobacteria
Thermotogae
Aquificae
Chloroflexi
Deinococcales
Cyanobacteria
Acidobacteria
δ-proteobacteria
ε-proteo-bacteria
α-proteo-bacteria
β-proteo-bacteria
γ-proteo-bacteria
Nanoarchaeum equitans
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus solfataricus
Sulfolobus tokodaii
Thermoplasma acidophilum
Thermoplasma volcanium
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Methanosarcina mazei
Methanosarcina acetivorans
Pyrococcus furiosus
Pyrococcus abyssi
Pyrococcus horikoshii
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanobact. thermautotrophicum
Giardia lamblia
Leishmania major
Thalassiosira pseudonana
Plasmodium falciparum
Cryptosporidium hominis
Cyanidioschyzon merolae
Arabidopsis thaliana
Oryza sativa
Dictyostelium discoideum
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Eremothecium gossypii
Caenorhabditis briggsae
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Danio rerio
Takifugu rubripes
Gallus gallus
Pan troglodytes
Homo sapiens
Mus musculus
Rattus norvegicus
Thermoanaerobact. tengcongensis
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Phytoplasma Onion yellows
Mycoplasma mycoides
Mycoplasma pulmonis
Mycoplasma mobile
Ureaplasma parvum
Mycoplasma penetrans
M. gallisepticum
M. genitalium
M. pneumoniae
Lactobacillus plantarum
Lactobacillus johnsonii
Enterococcus faecalis
Lactococcus lactis
Streptococcus pneumoniae TIGR4
Streptococcus pneumoniae R6
Streptococcus mutans
Streptococcus agalactiae V
Streptococcus agalactiae III
Streptococcus pyogenes M1
Streptococcus pyogenes 8232
Streptococcus pyogenes SSI-1
Streptococcus pyogenes 315
Staphylococcus epidermidis
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Listeria innocua
Listeria monocytogenes EGD
Listeria monocytogenes F2365
Oceanobacillus iheyensis
Bacillus halodurans
Bacillus subtilis
Bacillus anthracis
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579
Fibrobacter succinogenes
Chlorobium tepidum
Bacteroides thetaiotaomicron
Porphyromonas gingivalis
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila caviae
Chlamydophila pneumoniae 183
Chlamydophila pneumoniae J138
Chlamydophila pneumoniae AR39
Chlamydophila pneumoniae CWL029
Rhodopirellula baltica
Gemmata obscuriglobus
Leptospira interrogans 56601
Leptospira interrogans L1-130
Borrelia burgdorferi
Treponema denticola
Treponema pallidum
Bifidobacterium longum
Tropheryma whipplei Twist
Tropheryma whipplei TW08/27
Streptomyces avermitilis
Streptomyces coelicolor
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebact. glutamicum
Corynebact. glutamicum 13032
Mycobacterium paratuberculosis
Mycobacterium leprae
Mycobacterium bovis
Mycobact. tuberculosis 1551
Mycobact. tuberculosis H37Rv
Fusobacterium nucleatum
Thermotoga maritima
Aquifex aeolicus
Dehalococcoides ethenogenes
Deinococcus radiodurans
Thermus thermophilus
Gloeobacter violaceus
Thermosynechococcus elongatus
Synechocystis sp. PCC6803
Nostoc sp. PCC 7120
Prochlorococcus marinus 1375
Prochlorococcus marinus 9313
Prochlorococcus marinus SS120
Synechococcus sp. WH8102
Solibacter usitatus
Acidobacterium capsulatum
Desulfovibrio vulgaris
Bdellovibrio bacteriovorus
Geobacter sulfurreducens
Campylobacter jejuni
Wolinella succinogenes
Helicobacter hepaticus
Helicobacter pylori 26695
Helicobacter pylori J99
Wolbachia sp. wMel
Rickettsia conorii
Rickettsia prowazekii
Caulobacter crescentus
Bradyrhizobium japonicum
Rhodopseudomonas palustris
Mesorhizobium loti
Brucella melitensis
Brucella suis
Sinorhizobium meliloti
Agrobacterium tumefaciens W.
Agrobacterium tumefaciens C.
Nitrosomonas europaea
Chromobacterium violaceum
Neisseria meningitidis B
Neisseria meningitidis A
Ralstonia solanacearum
Bordetella pertussis
Bordetella bronchiseptica
Bordetella parapertussis
Coxiella burnetii
Xanthomonas campestris
Xanthomonas axonopodis
Xylella fastidiosa 9a5c
Xylella fastidiosa 700964
Pseudomonas aeruginosa
Pseudomonas syringae
Pseudomonas putida
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Haemophilus ducreyi
Haemophilus influenzae
Pasteurella multocida
W. brevipalpis
B. floridanus
Buchn. aphidicola Bp
Buchnera aphidicola Sg
Buchnera aphidicola APS
Photorhabdus luminescens
Yersinia pestis Medievalis
Yersinia pestis CO92
Yersinia pestis KIM
Salmonella typhimurium
Salmonella typhi
Salmonella enterica
Escherichia coli O157:H7
Escherichia coli EDL933
Shigella flexneri 2a 2457T
Shigella flexneri 2a 301
Escherichia coli K12
Escherichia coli O6
B) habitat preference vs. time: environmental rRNA isolates
habitat similarity
(shared keywords in habitat description)
0
0.1
0.2
0.3
0
0.2
0.4
0.6
0.8
1
1.2
molecular divergence
(branch length distance in global rRNA tree)
pairwise comparisons
27,148 rRNA isolates
same data, but habitats shuffled
C) habitat preference vs. time: cultured strains in collections
habitat similarity
(shared keywords in habitat description)
0.2
0.3
0.4
0.5
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)
pairwise comparisons
15,350 cultured strains
same data, but habitats shuffled


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 →)