M. R. Waters, Eds. (Center for the Study of the First Americans, College Station, TX, 2005), pp. 113–132.

2. M. B. Collins, *Clovis Blade Technology* (Univ. of Texas Press, Austin, TX, 1999).

3. K. B. Tankersley, in *The Settlement of the American Continents: A Multidisciplinary Approach to Human Biogeography*, C. M. Barton, G. A. Clark, D. R. Yesner, G. A. Pearson, Eds. (Univ. of Arizona Press, Tucson, AZ, 2004), pp. 49–63.

4. T. W. Stafford Jr., P. E. Hare, L. Currie, A. J. T. Jull, D. J. Donahue, *J. Archaeol. Sci.* **18**, 35 (1991).

5. Materials and methods are available as supporting material on *Science* Online.

6. G. A. Pearson, in *The Settlement of the American Continents: A Multidisciplinary Approach to Human Biogeography*, C. M. Barton, G. A. Clark, D. R. Yesner, G. A. Pearson, Eds. (Univ. of Arizona Press, Tucson, AZ, 2004), pp. 85–102.

7. A. C. Roosevelt, J. Douglas, L. Brown, in *The First Americans: The Pleistocene Colonization of the New World,* no. 27 of *Memoirs of the California Academy of Sciences,* N. G. Jablonski, Ed. (California Academy of Sciences, San Francisco, CA, 2002), pp. 159–235.

8. C. V. Haynes Jr., in *Radiocarbon After Four Decades: An Interdisciplinary Perspective*, R. E. Taylor, A. Long, R. S. Kra, Eds. (Springer-Verlag, New York, 1992), pp. 355–374.

9. S. J. Fiedel, *Am. Antiq.* **64-1**, 95 (1999).

10. S. Bondevik, J. Mangerud, H. H. Birks, S. Gulliksen, P. Reimer, *Science* **312**, 1514 (2006).

11. B. Kromer *et al.*, *Radiocarbon* **46-3**, 1203 (2004).

12. R. G. Fairbanks *et al.*, *Quat. Sci. Rev.* **24**, 1781 (2005).

13. S. J. Fiedel, *J. Archaeol. Res.* **8-1**, 39 (2000).

14. C. A. S. Mandryk, H. Josenhans, D. W. Fedje, R. W. Mathewes, *Quat. Sci. Rev.* **20**, 301 (2001).

15. B. Bradley, D. Stanford, *World Archaeol.* **36**, 459 (2004).

16. T. D. Hamilton, T. Goebel, in *Ice Age Peoples of North America: Environments, Origins, and Adaptations of the First Americans*, R. Bonnichsen, K. L. Turnmire, Eds. (Oregon State Univ. Press, Corvallis, OR, 1999), pp. 156–199.

17. J. F. Hoffecker, W. R. Powers, T. Goebel, *Science* **259**, 46 (1993).

18. R. Bonnichsen, in *Clovis: Origins and Adaptations*, R. Bonnichsen, K. L. Turnmire, Eds. (Oregon State Univ. Press, Corvallis, OR, 1991), pp. 309–329.

19. G. C. Frison, Ed., *The Mill Iron Site* (Univ. of New Mexico Press, Albuquerque, NM, 1996).

20. J. R. Johnson, T. W. Stafford Jr., H. O. Ajie, D. P. Morris, in *Proceedings of the Fifth California Islands Symposium* (U.S. Department of the Interior, Minerals Management Service, Pacific Outer Continental Shelf Region, Washington, DC, 23 March to 1 April 1999), pp. 541–544.

21. T. Goebel, B. Hockett, K. Graf, D. Rhode, paper presented at the 30th Great Basin Anthropological Conference, Las Vegas, NV, 19 to 21 October 2006.

22. L. Miotti, M. C. Salemme, *Quat. Int.* **109-110**, 95 (2003).

23. M. Barbetti *et al.*, *Nucl. Instrum. Methods* **223-224B**, 366 (2004).

24. D. G. Anderson, J. C. Gillam, *Am. Antiq.* **65-1**, 43 (2000).

25. D. F. Overstreet, in *Paleoamerican Origins: Beyond Clovis*, R. Bonnichsen, B. T. Lepper, D. Stanford, M. R. Waters, Eds. (Center for the Study of the First Americans, College Station, TX, 2005), pp. 183–195.

26. D. J. Joyce, *Quat. Int.* **142-143**, 44 (2006).

27. T. D. Dillehay, Ed., *Monte Verde: A Late Pleistocene Settlement in Chile: Volume 2: The Archaeological Context and Interpretation* (Smithsonian Institution Press, Washington, DC, 1997).

28. G. K. Ward, S. R. Wilson, *Archaeometry* **20**, 19 (1978).

29. We thank A. Hannus, C. V. Haynes, J. Gingerich, G. Frison, and A. Hemmings for providing samples for dating; M. Payn and the North Star Archaeological Research Program established by J. Cramer and R. Cramer for providing funding; P. Reimer and J. Southon for providing advice on ^{14}C calibration; D. Carlson, T. Goebel, J. Southon, S. Forman, and three anonymous reviewers for offering useful comments to improve this paper; and L. Lind, C. Pevny, J. Halligan, and P. Johnson for helping in the preparation of the text and illustrations.

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]†

The taxonomic composition of environmental communities is an important indicator of their ecology and function. We used 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 that provided by traditional ribosomal RNA–based approaches depending on the polymerase chain reaction. Mapping marker genes from four diverse environmental data sets onto a reference species phylogeny shows that certain communities evolve faster than others. The method also enables determination of preferred habitats for entire microbial clades and provides evidence that such habitat preferences are often remarkably stable over time.

Microorganisms are estimated to make up more than one-third of Earth's biomass (*1*). They play essential roles in the cycling of nutrients, interact intimately with animals and plants, and directly influence 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 (rRNA) genes (most notably the 16*S*/18*S* 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 polymerase chain reaction (PCR) step introduces amplification bias, and it generates chimeric and otherwise erroneous molecules that hamper phylogenetic analysis (*8*, *9*).

Shotgun sequencing of community DNA ("metagenomics") provides a more direct and unbiased access to uncultured organisms (*10–13*): No PCR amplification step is involved, and because no specific primers or sequence anchors are needed, even very unusual organisms can be captured by this technique. Although 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 (*14–16*).

To make use of metagenomics data for taxonomic profiling, we analyzed 31 protein-coding marker genes previously shown to provide sufficient information for phylogenetic analysis [they are universal, occur only once per genome, and are rarely transferred horizontally (*17*)]. We extracted these marker genes from metagenomics sequence data (*9*), aligned them to a set of hand-curated reference proteins, and used maximum likelihood to map each sequence to an externally provided phylogeny of completely sequenced organisms [tree of life; we used the tree from (*17*), 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 (*18*) (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)." The procedure weighs the number of informative residues that are found on each sequence fragment, then adjusts the spread and confidence of its placement

[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.

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

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

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 (9)]. 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).

Jackknife validation of our method [i.e., leaving out various parts of the reference tree and measuring the consequences on placement accuracy (9)] 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; with our method, such changes are reduced to 0.19% (fig. S2). Moreover, because 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 (12), 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 (9). The PCR-based approach presumably suffers from amplification biases and from copy number variations among rRNA genes in bacteria (19) 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 (9) for a discussion of the strengths, weaknesses, and complementarities of both approaches].

We applied our procedure to four large, heterogeneous data sets of microbial community se-

quences derived from distinct and geographically separate environments (11–13). 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 uneven 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 represent only a small fraction of naturally occurring microbes. Conversely, entire phyla such as the Acidobacteria or the Chloroflexi are poorly represented among the sequenced genomes but are widely abundant in natural communities.

In agreement with (20), we found 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 were apparent: relatives of the Rickettsiales, for example [including the marine genus *Pelagibacter* (21)], were mostly found in the surface-water sample, whereas relatives of Rhizobiales or Burkholderiales were mostly found in the soil sample. We observed surprisingly few endospore-forming organisms in the community sequences: Both Bacilli and Clostridia were quite rare; their largest combined abundance was a mere 1% (in soil). Similarly, Actinobacteria (many of which have a spore stage) ranged 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. Although our approach used 31 marker genes with a total of about 7500 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 usable for our approach ranged from 247 for the smallest data set to 15,741 for the largest data set). We quantified the potential undersampling errors by jackknife and bootstrap analysis (fig. S3). These tests showed that, for the worst case of a low-abundance clade in the smallest data set, the quantitative error due to undersampling was 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 to 2.153%, meaning that endospore-formers are indeed rare in soil and are not just undersampled. Generally, none of the results reported here would change much if all data sets had as many as 15,000 marker genes sampled (in particular because 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 toward 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 was usually less than 60% (for soil, the median identity was only 47%). This dissimilarity was
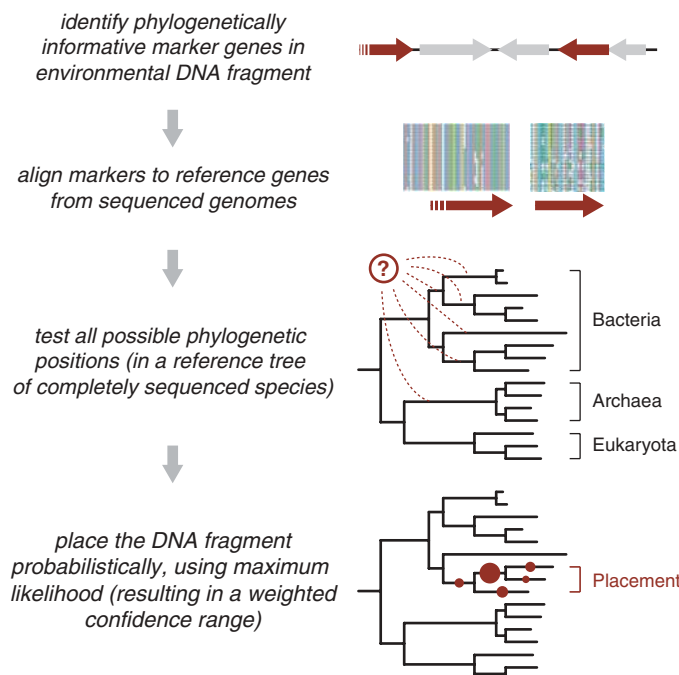

Fig. 1. Assessing community taxonomy from metagenomics sequence data. The diagram depicts how a restricted set of marker genes can be used for phylogenetic characterization of community microbes from poorly assembled sequence data. Instances of the marker genes are sought in the sequences and assessed relative to an external tree-of-life phylogeny with the use of maximum likelihood scoring. A central step in the mapping procedure is the assignment of a confidence range for each placement, thereby avoiding overconfident placement of sequence fragments that are short or otherwise uninformative.

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)

Bacteria

Archaea

Eukaryota

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

Placement

A

Legend:
- Agricultural Soil
- Ocean Surface Water
- Acidic Mine Drainage
- Deep Sea Whale Bone

γ-proteo-bacteria

β-proteo-bacteria

α-proteo-bacteria

ε-proteo-bacteria

δ-proteobacteria
Acidobacteria

Cyanobacteria
Deinococcales
Chloroflexi
Aquificae
Thermotogae

Fusobacteria   Actinobacteria

Nanoarchaeota
Crenarchaeota
Euryarchaeota
Diplomonadida
Kinetoplastida
Chromalveolata
Plantae
Amoebozoa
Fungi

Metazoa

Firmicutes

Fibrobacteres
Chlorobi
Bacteroidetes
Chlamydiae
Planctomycetes
Spirochaetes

B

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 similarity
(shared keywords in habitat description)

pairwise comparisons
- 15,350 cultured strains
- same data, but habitats shuffled

taxonomic divergence
(taxonomic rank of last common branching point in phylogeny)

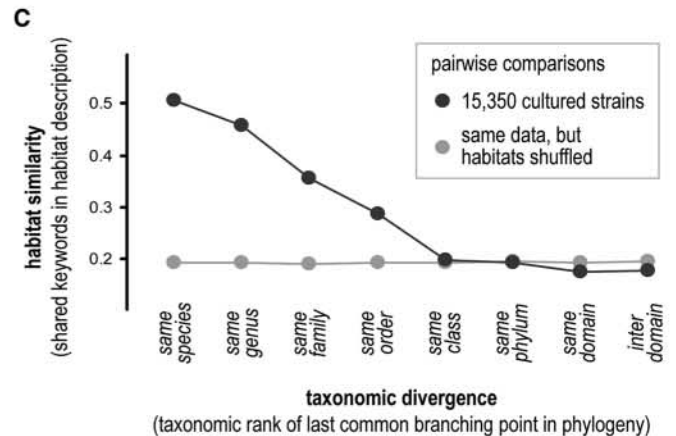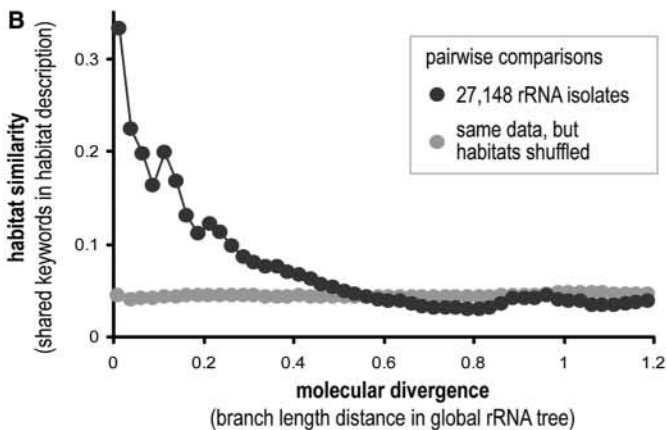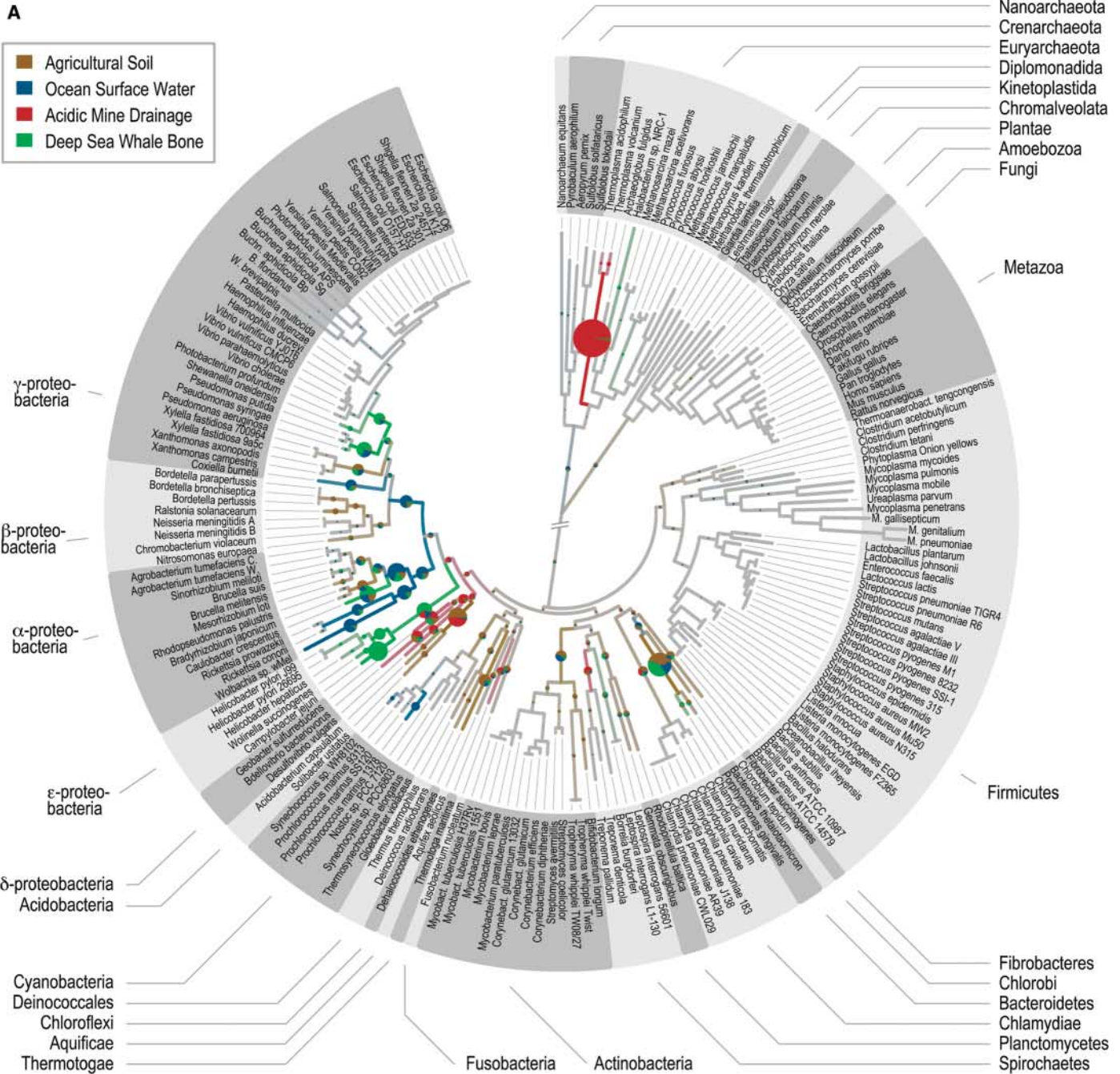same species, same genus, same family, same order, same class, same phylum, same domain, inter domain

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 clearly preferred habitat, lines are colored accordingly (*9*). (**B** and **C**) Habitat preference over time. (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 is highly significant ($P < 10^{-6}$). The tree-drawing algorithm is implemented for public use at (*34*).

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 β- and γ-proteobacteria, which has been tentatively dated at more than 500 million years ago (*22–24*), 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. 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 "fourth 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
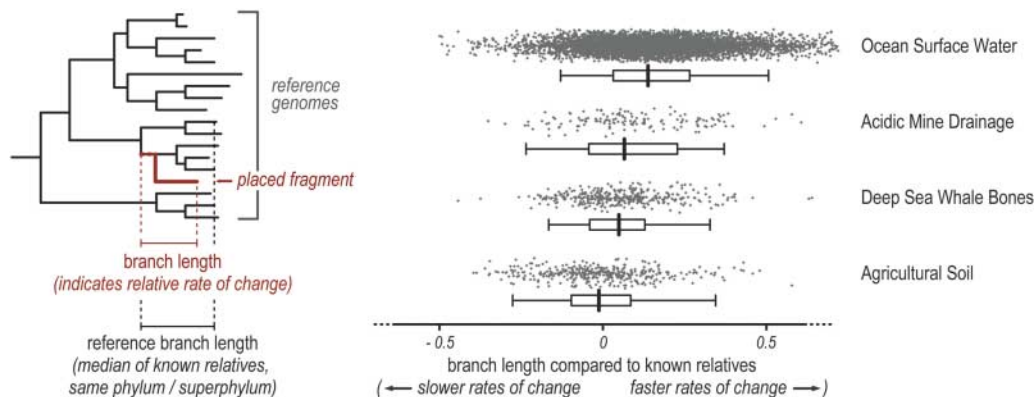
assessed, for each sequence fragment placed into the tree, the cumulative branch length from the tip of its branch down to the base of the corresponding phylum, and compared these to the branch lengths of all known reference organisms in that same phylum, measured for the very gene families found on 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 were 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, were previously observed mainly for multicellular organisms [e.g., a dependency between latitudinal geographic location and mutation rates in plants (*25*)]. 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" (*21*, *26*). The more subtle, global variations in mutation rates reported here may be caused by differences in population sizes or generation times, or by the abundance of external mutagens (such as the strong fluxes of ultraviolet light in ocean surface water). Notably, the ocean surface community is not only evolving the fastest, it is also the one with the smallest genomes (*27*). 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 allowed us to trace the habitat preference of microbial organisms through time, and thus enabled us to estimate how frequently lineages change their preferred environment. At short to intermediate evolutionary time scales, we observed a noticeable stability of habitats: Many of the closer relatives in the tree showed the same environmental preference, indi-

cating that microbial lineages do not very often change (or specialize) their lifestyles and habitats (Fig. 2). Conversely, at longer time scales, we did observe notable 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. In the case of Cyanobacteria, we identified relatives of the fast-evolving and widespread Prochlorococci in the ocean sample, whereas more basal, slower-evolving Cyanobacteria such as *Gloeobacter* were mostly found in the soil sample.

Even though molecular methods tend to find most phyla ubiquitously, Baas-Becking and Beyerinck postulated decades ago that microbial taxa have preferred environments: "for microbial taxa, everything is everywhere—but the environment selects" [(*28*) and references therein]. The hypothesis posits that microorganisms are frequently dispersed globally, and that they are only subsequently selected by the environments on the basis of their functional capacities. Existing communities would thus constantly be challenged by intruders from nonspecialist phyla that may occasionally survive simply by chance, acquiring the necessary functionality through horizontal gene transfer (*29–31*). 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 data sets 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 (*9*). Indeed, we observe a remarkable time-dependent stability of

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 data set, 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 data sets differ highly significantly (two-sided Kolmogorov-Smirnov tests, $P \le 10^{-5}$, 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 subsamples pooled), 630 genes on 362 contigs; agricultural soil, 598 genes on 395 contigs.

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 (*32*). For most other adaptations, fine-tuned regulation and/or subtle changes in the majority of proteins may be needed. Because 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 (*33*), 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. See supporting material on *Science* Online.
10. C. S. Riesenfeld, P. D. Schloss, J. Handelsman, *Annu. Rev. Genet.* **38**, 525 (2004).
11. J. C. Venter *et al.*, *Science* **304**, 66 (2004); published online 4 March 2004 (10.1126/science.1093857).
12. S. G. Tringe *et al.*, *Science* **308**, 554 (2005).
13. G. W. Tyson *et al.*, *Nature* **428**, 37 (2004).
14. G. M. Luna, A. Dell'Anno, R. Danovaro, *Environ. Microbiol.* **8**, 308 (2006).
15. H. A. Barton, N. M. Taylor, B. R. Lubbers, A. C. Pemberton, *J. Microbiol. Methods* **66**, 21 (2006).
16. I. M. Kauffmann, J. Schmitt, R. D. Schmid, *Appl. Microbiol. Biotechnol.* **64**, 665 (2004).
17. F. D. Ciccarelli *et al.*, *Science* **311**, 1283 (2006).
18. K. Strimmer, A. Rambaut, *Proc. Biol. Sci.* **269**, 137 (2002).
19. J. A. Klappenbach, P. R. Saxman, J. R. Cole, T. M. Schmidt, *Nucleic Acids Res.* **29**, 181 (2001).
20. M. S. Rappe, S. J. Giovannoni, *Annu. Rev. Microbiol.* **57**, 369 (2003).
21. S. J. Giovannoni *et al.*, *Science* **309**, 1242 (2005).
22. F. U. Battistuzzi, A. Feijao, S. B. Hedges, *BMC Evol. Biol.* **4**, 44 (2004).
23. D. F. Feng, G. Cho, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13028 (1997).
24. H. Ochman, S. Elwyn, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12638 (1999).
25. S. Wright, J. Keeling, L. Gillman, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7718 (2006).
26. A. Dufresne, L. Garczarek, F. Partensky, *Genome Biol.* **6**, R14 (2005).
27. J. Raes *et al.*, *Genome Biol.* **8**, R10 (2007).
28. J. B. Martiny *et al.*, *Nat. Rev. Microbiol.* **4**, 102 (2006).
29. W. F. Doolittle, *Trends Cell Biol.* **9**, M5 (1999).
30. R. F. Doolittle, *Curr. Opin. Struct. Biol.* **15**, 248 (2005).
31. I. Chen, P. J. Christie, D. Dubnau, *Science* **310**, 1456 (2005).
32. N. U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, *Nature* **439**, 847 (2006).
33. C. Pedros-Alio, *Trends Microbiol.* **14**, 257 (2006).
34. I. Letunic *et al.*, *Bioinformatics* **23**, 127 (2007).
35. We thank P. Dawyndt for providing an early version of his integrated strain database (www.straininfo.net), and members of the Bork team for insightful discussions. Supported by the European Union through its BioSapiens and GeneFun networks, by the University of Zurich through its Research Priority Program Systems Biology and Functional Genomics, by the NSF Assembling the Tree of Life and Microbial Genome Sequencing programs, and by the German Federal Government through its National Genome Research Network.

Staphylococcus aureus Panton-Valentine Leukocidin Causes Necrotizing Pneumonia

Maria Labandeira-Rey,[1] Florence Couzon,[2–5] Sandrine Boisset,[2–5] Eric L. Brown,[1]* Michele Bes,[2–5] Yvonne Benito,[2–5] Elena M. Barbu,[1] Vanessa Vazquez,[1] Magnus Höök,[1] Jerome Etienne,[2–5] François Vandenesch,[2–5]†‡ M. Gabriela Bowden[1]†‡

The *Staphylococcus aureus* Panton-Valentine leukocidin (PVL) is a pore-forming toxin secreted by strains epidemiologically associated with the current outbreak of community-associated methicillin-resistant *Staphylococcus aureus* (CA-MRSA) and with the often-lethal necrotizing pneumonia. To investigate the role of PVL in pulmonary disease, we tested the pathogenicity of clinical isolates, isogenic PVL-negative and PVL-positive *S. aureus* strains, as well as purified PVL, in a mouse acute pneumonia model. Here we show that PVL is sufficient to cause pneumonia and that the expression of this leukotoxin induces global changes in transcriptional levels of genes encoding secreted and cell wall–anchored staphylococcal proteins, including the lung inflammatory factor staphylococcal protein A (Spa).

The combined actions of many virulence factors enable *Staphylococcus aureus* to cause disease (*1, 2*). Depending on these factors and on the immune status of the host, staphylococci can cause diseases ranging from superficial skin infections to deep-seated and systemic conditions such as osteomyelitis, septic shock, and necrotizing pneumonia. Staphylococcal necrotizing pneumonia can affect young, immunocompetent patients. This disease, characterized by leukopenia, hemoptysis, and extensive necrosis of the lung tissue, is caused by *S. aureus* strains that produce Panton-Valentine leukocidin (PVL) (*3*). *S. aureus* PVL-positive strains are often methicillin-resistant (MRSA) and, in the USA, they are the predominant cause of community-associated infections (*4*).

PVL is a bi-component, pore-forming exotoxin (*5*) that targets cells of the immune system such as polymorphonuclear neutrophils (PMNs). The active form of PVL requires the assembly of two polypeptides, LukS-PV and LukF-PV, into a heterooligomeric pore. Although PVL has potent cytolytic and inflammatory activities in vitro (*6, 7*), its role in necrotizing pneumonia has not been demonstrated. To analyze the molecular pathogenesis of PVL-expressing *S. aureus* strains, we have established a murine model of acute primary pneumonia.

We infected mice with strains isolated from necrotizing (PVL-positive) or nonnecrotizing

[1]Center for Extracellular Matrix Biology, Institute of Biosciences and Technology, The Texas A&M University System Health Science Center, Houston, TX 77030, USA. [2]Université de Lyon, Lyon, F-69008, France. [3]Université de Lyon 1, Faculté Laennec, Lyon, F-69008, France. [4]INSERM E0230, Centre National de référence des staphylocoques, Lyon, F-69008, France. [5]Hospices Civils de Lyon, Hôpital Edouard Herriot, Lyon, F-69003, France.

*Present address: University of Texas School of Public Health, Houston, TX 77030, USA.
†These authors contributed equally to this work.
‡To whom correspondence should be addressed. E-mail: gbowden@ibt.tamhsc.edu and denesch@univ-lyon1.fr