LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-211879

Comparative Metagenomics of Microbial Communities

S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, E. Rubin

May 2, 2005

Science

Comparative Metagenomics of Microbial Communities

UCRL-JRNL-211879

Susannah Green Tringe[1,2†], Christian von Mering[3†], Arthur Kobayashi[1], Asaf A. Salamov[1], Kevin Chen[4], Hwai W. Chang[5], Mircea Podar[5], Jay M. Short[5], Eric J. Mathur[5], John C. Detter[1], Peer Bork[3], Philip Hugenholtz[1], Edward M. Rubin[1,2*]

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

[2]Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, CA 94720, USA

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

[4]University of California, Berkeley, Department of Electrical Engineering and Computer Science, Berkeley, CA 94720, USA

[5]Diversa Corporation, 4955 Directors Place, San Diego, CA 92121, USA

One-sentence summary: The predicted proteins encoded in unassembled DNA sequence from environmental microbial community samples reveal habitat-specific metabolic fingerprints.

*To whom correspondence should be addressed: emrubin@lbl.gov

†S.G.T and C.v.M contributed equally to this work

Assembled genomes are difficult to obtain from environmental samples, owing to the species complexity of microbial communities and challenges in culturing representative isolates. Here we characterize and compare the metabolic capabilities of terrestrial and marine microbial communities utilizing largely unassembled sequence data obtained via the shotgun sequencing of DNA isolated from the various environments. Quantitative gene content analysis revealed habitat-specific fingerprints that reflect known characteristics of the sampled environments. The identification of environment-specific genes through a gene-centric comparative analysis presents new opportunities for interpreting and diagnosing environments.

Despite their ubiquity, relatively little is known about the majority of environmental microorganisms largely because of their resistance to culture under standard laboratory conditions. A variety of environmental sequencing projects targeted at 16S ribosomal RNA (rRNA) (*1, 2*) has offered a glimpse into the phylogenetic diversity of uncultured organisms. The direct sequencing of environmental samples has provided further valuable insight into the lifestyles and metabolic capabilities of uncultured organisms occupying various environmental niches. These efforts include the sequencing of individual large-insert BAC clones as well as small-insert libraries made directly from environmental DNA (*3-7*). The application of high-throughput shotgun sequencing of environmental samples has recently provided global views of those communities not obtainable from 16S rRNA or BAC clone sequencing surveys (*6,*

7). The sequence data have also posed challenges to genome assembly, suggesting that complex communities will demand enormous sequencing expenditure for the assembly of even the most predominant members (*7*).

A practical question emerging from environmental sequencing projects is the extent to which the data are interpretable in the absence of significant individual genome assemblies. Most microbial communities are extremely complex and thus not amenable to genome assembly (*8*). This obstacle may in part be offset by the high gene density of prokaryotes (~1 open reading frame / 1000 base pairs) and currently attainable read lengths (700-750 base pairs) which result in most individual sequences containing a significant portion of at least one gene (*9*). Accordingly, while microbial as well as animal sequencing studies have typically targeted complete genomes, for metagenomic data this approach may not always be necessary or feasible. Determining the proteins encoded by a community, rather than the types of organisms producing them, suggests a means to distinguish samples based on the functions selected for by the local environment and reveal insights into features of that environment. In these studies, we took a gene-centric approach to environmental sequencing in our analysis of several disparate microbial communities.

The samples we characterized were derived from agricultural soil and from three isolated deep sea "whale fall" carcasses (*10*). In contrast to the nutrient-poor environments previously subjected to large-scale metagenomic sequencing (*6, 7*), each of these environments was nutrient-rich albeit with very different nutrient sources (plant material for soil, lipid-rich bone for deep sea whale fall samples).

We first analyzed the microbial diversity in these samples through polymerase chain reaction (PCR) amplified small ribosomal RNA libraries generated for each sample using primers specific for Bacteria, Archaea, and Eukaryota. In the soil sample, a wide diversity of bacteria, very few archaeal species and some fungi and unicellular eukaryotes were found (Figure S2). We sequenced a total of 1700 clones from two independent libraries of PCR-amplified bacterial 16S rRNA sequences prepared from the soil DNA, and we identified at least 847 distinct ribotypes from more than a dozen phyla (Figure S2B). A rarefaction curve built from these data failed to reach saturation, and coverage estimators such as Chao1 (*11, 12*) predicted the total number of bacterial ribotypes in this sample to be more than 3000 (Figures 1, S1), reflecting the enormous diversity found in soil (*8*). The most common ribotype accounts for 112 (6.6%) of the clones (Figure S2D) when a 97% identity cutoff is used, and 81 (4.8%) when 98% identity is required. The whale fall samples are both less diverse and less evenly distributed than the soil cohort and are estimated to contain between 25 and 150 distinct ribotypes of which the most abundant accounts for 15-25% of the library (Figure 1, Supplemental Figure S3). The reduced species and phyla diversity of the whale fall microbial communities as compared to soil is consistent with the extreme and specialized nature of this deep ocean ecological niche.

We explored the genomic diversity of the communities by sequencing genomic small-insert libraries made from all four samples. In light of the organismal complexity seen in the soil sample, we generated 100 million base pairs (Mbp) of sequence from this sample and 25 Mbp for each whale fall library. Consistent

with the predicted high species diversity in the soil sample, attempts at sequence assembly were largely unsuccessful. Less than 1% of the nearly 150,000 reads generated from the soil library exhibited overlap with reads from independent clones. Based on our 16S rRNA data and the overlaps in the genomic sequence, we projected that somewhere between two and five billion base pairs of sequence would be necessary to obtain the eightfold coverage traditionally targeted for draft genome assemblies, even for the single most predominant genome in this complex community (*13*). For each whale fall library, we estimate that between 100 and 700 Mb of shotgun sequence data would be needed in order to generate a draft assembly for the most prevalent genome. Assembling genomes for low-abundance community members in any of these samples would clearly require significantly more sequence data.

Given these hurdles to the assembly of complete genomes from the samples, we investigated the genes present without attempting to place them in the context of an individual genome. In preliminary studies we compared gene predictions from assembled versus unassembled sequence using available metagenomic data (*13*). With our analysis supporting the validity of gene predictions on unassembled reads, we applied an automated annotation process to the sequence data from several different environmental samples. As our analysis relied primarily on the predicted genes on small DNA fragments, the majority of which were individual sequence reads, we termed each environmental sequence an Environmental Gene Tag (EGT), to distinguish them from the sequencing reads primarily used for the assembly of genomes. The gene contents of the partially assembled and

unassembled reads from soil and whale fall samples were compared to each other and to those of an acid mine drainage biofilm community (*6*) and each of three independent samples from Sargasso Sea surface waters (*7*). Putative genes were predicted on at least 90% of the EGTs from all samples, even when the sequence fragments were individual reads. More than a third of the EGTs contained two or more predicted open reading frames, raising the possibility of nearest-neighbor analysis (*14*).

Roughly half of the predicted proteins in each sample exhibited homology to orthologous groups in an expanded in-house COG database (*15, 16*). To test whether the orthologous groups observed in a limited sampling of each library were representative of the full range of groups in a community, we plotted the number of orthologous groups detected at increasing levels of sequencing depth. For all samples, saturation for frequently occurring orthologous groups is observed after a modest amount of sequencing while the general slope of the curve reveals information about community diversity (Figure 2). In the relatively simple acid mine drainage biofilm community, 90% of the orthologous groups were detected with just 25 Mbp raw sequence (~15 Mbp quality sequence) – a fraction of that needed to assemble genomes. Even in the considerably more complex soil community, the curve starts to flatten at 25 Mbp, suggesting that new orthologous groups detected at this point are found only in a minority of the community members. The Sargasso Sea communities, consistent with their species complexity, fell between acid mine drainage and soil; the whale falls, however, exhibited trajectories quite similar to soil. We observed qualitatively

similar curves when limiting the analysis to the 4873 COGs contained in the 2003 release or to the domain-oriented Pfam database (*17*)(Supplemental Figure S3), suggesting that this phenomenon is not an artifact of comparison to a particular database.

We next explored the relative proportion of the total protein sets devoted to particular functions in a sample, given evidence that not only message levels (*18*) but library representation (*19*) of genes coding for specialized enzymes can vary with sample source. We specifically explored whether independent samples from similar, though geographically separated, environments would exhibit functional profiles more similar to each other than to those from disparate environments. We binned predicted proteins into functional categories at four levels; first, individual genes (orthologous groups inferred from sequenced genomes), second, groups of genes frequently observed as neighbors in complete genomes ("operons," shown to correlate with metabolic pathways (*20*)), third, higher order cellular processes from the manually curated KEGG database (*21*), and fourth, broad functional categories from the COG database (*13*, *15*). Assembled contigs were weighted to account for the number of independent clones contributing to them.

A two-way clustering of samples and KEGG maps, in which over- and under-represented categories are indicated by red and blue blocks respectively, is displayed in Figure 3 (Figure S4 displays similar figures based on COGs and operons). Regardless of the functional binning employed, the independent Sargasso Sea samples clustered together, as did the whale fall samples. These

profiles clearly suggest that the predicted protein complement of a community is similar to that of other communities whose environments of origin pose similar metabolic demands. Our results further support the hypothesis that the "functional" profile of a community is influenced by its environment and that EGT data can be used to develop fingerprints for particular environments.

To assess the significance of these similarities and differences, and to identify functions of importance for communities existing in specific environments, we systematically examined the differences in gene content between samples (Figure 4). For this analysis, the three whale fall samples were pooled together, as were the three ocean samples. At each level, significant differences among the respective microbial communities were observed that suggested environment-specific variations in both biochemistry and phylogeny. The acid mine drainage was not included in this analysis because of its high dissimilarity from the other samples (Figures 3, S6) and low species diversity, both likely reflective of the very extreme nature of this environment.

At the individual gene level, quite a few orthologous groups are exclusive to a particular environment (Figure 4, upper left). For example, 73 putative orthologs of cellobiose phosphorylase, involved in degradation of plant material, are found in the ~100 Mb of soil sequence but none are found in the ~700 Mb of sequence examined from the Sargasso Sea. On the other hand, 466 distinct copies of the light-driven proton pump bacteriorhodopsin are found in the surface waters of the Sargasso Sea, while none are found in the deep sea whale falls or in soil.

The analysis of operons likewise reveals similarities and differences in functional systems (Figure 4, upper right) that suggest features of the environments. The most discriminating operons tend to be systems for the transport of ions and inorganic components, highlighting their importance for survival and adaptation. With respect to ionic and osmotic homeostasis, for example, the two maritime environments are very similar – both show a strong enrichment in operons that contain transporters for organic osmolites and sodium ion exporters coupled to oxidative phosphorylation. The soil sample, on the other hand, has a strong enrichment in operons responsible for active potassium channeling. These biases nicely reflect the relative abundance of these ions in the respective environments: while typical ocean water contains considerably more sodium ions than potassium, the soil sample examined here contained high potassium and low sodium concentrations (*13*).

Examination of higher order processes reveals known differences in energy production (e.g. photosynthesis in the oligotrophic waters of the Sargasso Sea, starch and sucrose metabolism in soil) (*7*) or population density and interspecies communication (overrepresentation of conjugation systems, plasmids, and antibiotic biosynthesis in soil; Figure 4, lower left) (*22*). The broad functional COG categories, on the other hand, primarily suggest differences in genome size and phylogenetic composition (*13*).

Notably, many uncharacterized genes and processes are among the most overrepresented categories in each sample. This hints at an abundance of previously unknown functional systems, specific to each environment, whose

occurrence patterns may offer useful guidance for further, more directed experimental and computational investigations. More extensive sampling in both time and space will reveal which features are broadly distributed within a given environment and which are unique to the places and times sampled here. Nonetheless, this analysis of genes and functional modules in environments reveals expected contrasts, hints at certain nutrition conditions, and points to novel genes and systems contributing to a particular "lifestyle" or environmental interaction.

The predicted metaproteome, based on fragmented sequence data, is sufficient to identify functional fingerprints that can provide insight into the environments from which microbial communities originate. Information derived from extension of the comparative metagenomic analyses performed here could be used to predict features of the sampled environments such as energy sources or even pollution levels, while the environment-specific distribution of unknown orthologous groups and operons offers exciting avenues for further investigation. Just as the incomplete but information-dense data represented by expressed sequence tags (ESTs) have provided useful insights into various organisms and cell types, EGT-based ecogenomic surveys represent a practical and uniquely informative means for understanding microbial communities and their environments.

References

1. E. F. DeLong, N. R. Pace, *Syst Biol* **50**, 470 (Aug, 2001).
2. P. Hugenholtz, *Genome Biol* **3**, REVIEWS0003 (2002).

3. M. R. Liles, B. F. Manske, S. B. Bintrim, J. Handelsman, R. M. Goodman, *Appl Environ Microbiol* **69**, 2684 (May, 2003).
4. O. Beja *et al.*, *Science* **289**, 1902 (Sep 15, 2000).
5. C. Schmeisser *et al.*, *Appl Environ Microbiol* **69**, 7298 (Dec, 2003).
6. G. W. Tyson *et al.*, *Nature* **428**, 37 (Mar 4, 2004).
7. J. C. Venter *et al.*, *Science* **304**, 66 (Apr 2, 2004).
8. V. Torsvik, L. Ovreas, T. F. Thingstad, *Science* **296**, 1064 (May 10, 2002).
9. Y. A. Goo *et al.*, *BMC Genomics* **5**, 3 (Jan 12, 2004).
10. C. R. Smith, H. Kukert, R. A. Wheatcroft, P. A. Jumars, J. W. Deming, *Nature* **341**, 27 (Sep 7, 1989).
11. J. B. Hughes, J. J. Hellmann, T. H. Ricketts, B. J. Bohannan, *Appl Environ Microbiol* **67**, 4399 (Oct, 2001).
12. R. K. Colwell. (1994-2004). EstimateS: Statistical estimation of species richness and shared species from samples.
13. Supplementary Online Material.
14. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, N. Maltsev, *Proc Natl Acad Sci U S A* **96**, 2896 (Mar 16, 1999).
15. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (Sep 11, 2003).
16. C. von Mering *et al.*, *Nucleic Acids Res* **33 Database Issue**, D433 (Jan 1, 2005).
17. A. Bateman *et al.*, *Nucleic Acids Res* **32 Database issue**, D138 (Jan 1, 2004).
18. S. K. Rhee *et al.*, *Appl Environ Microbiol* **70**, 4303 (Jul, 2004).
19. D. E. Robertson *et al.*, *Appl Environ Microbiol* **70**, 2429 (Apr, 2004).
20. C. von Mering *et al.*, *Proc Natl Acad Sci U S A* **100**, 15428 (Dec 23, 2003).
21. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, *Nucleic Acids Res* **32 Database issue**, D277 (Jan 1, 2004).
22. R. Daniel, *Curr Opin Biotechnol* **15**, 199 (Jun, 2004).
22. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36 and SGT was supported by Grant No. THL007279F, an NIH NRSA Training and Fellowship grant to ER. Sequencing of the environmental libraries was performed under a license agreement to US patent #6455254. We gratefully acknowledge the efforts of Cynthia Baptista, Leif Christoffersen, Joe Garcia, Ke Li, Jason Ritter, Patrick Sammon, Steve Wells, Denise Whitney, Jonathan Eads, Toby Richardson, Mick Noordewier, and Lisa Bibbs. We wish to thank Craig Smith for providing the whale falls samples; Karin Remington for providing Sargasso Sea sample information; Natalia Ivanova, Nikos Kyrpides, and members of the Rubin lab for helpful comments on the manuscript; and Jarrod Chapman, Igor Grigoriev, Ernest Szeto, Jan Korbel, Tobias Doerks, Konrad

Foerstner, Eoghan Harrington and Marcus Krupp for assistance with data processing and analysis. These Whole Genome Shotgun projects have been deposited at DDBJ/EMBL/GenBank under the project accessions AAFX00000000 (soil), AAFY00000000, AAFZ00000000, and AAGA00000000. For each project, the version described in this paper is the first version, AAFX01000000, AAFY01000000, AAFZ01000000 and AAGA01000000. The metagenomic data will also be incorporated into the JGI Integrated Microbial Genomes (IMG) system (http://www.jgi.doe.gov/), to facilitate detailed comparative analysis of the data in the context of all publicly available complete microbial genomes.

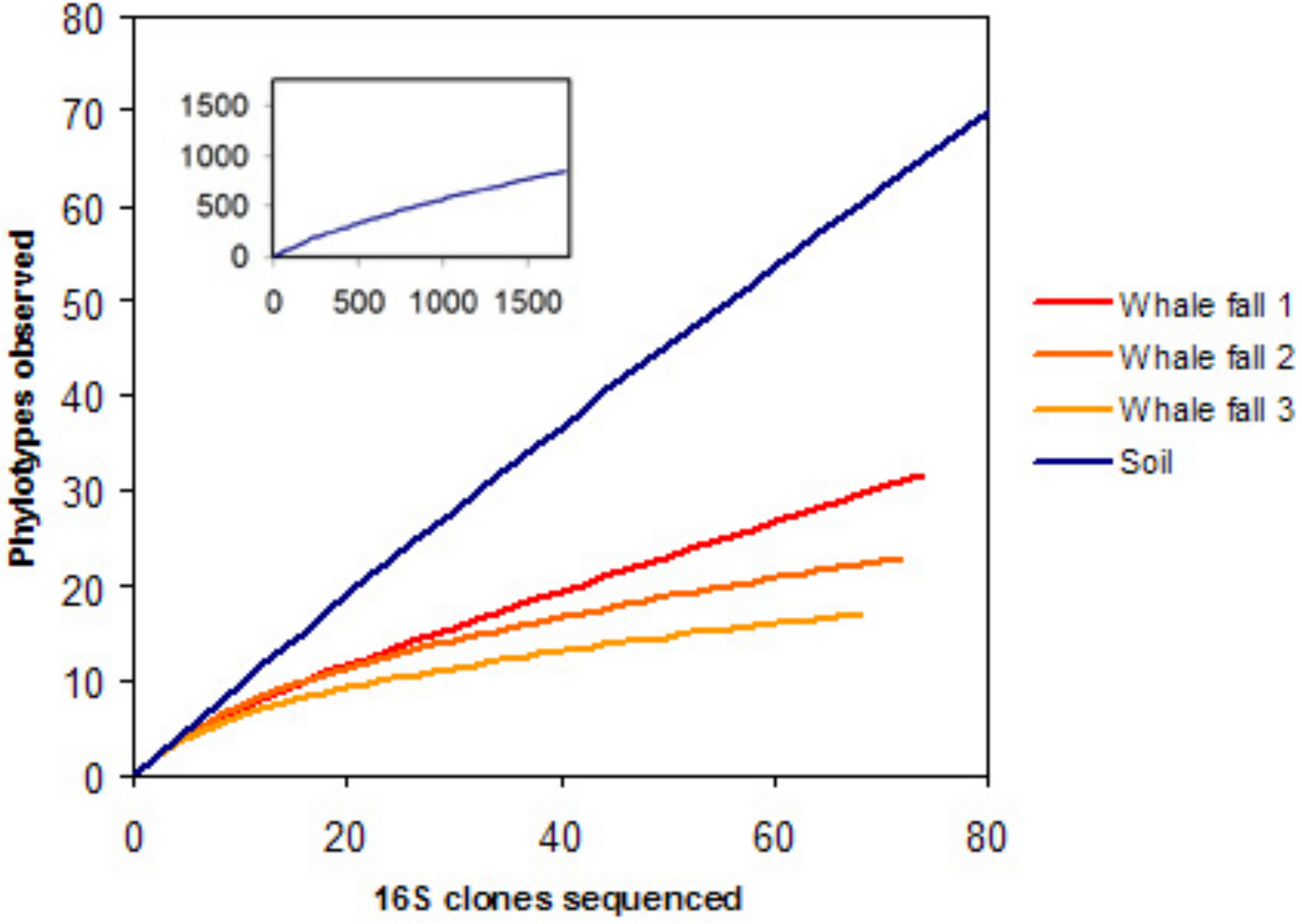
Figure 1: **Species complexity.** Rarefaction curves of bacterial 16S rRNA clone sequences for soil and whale fall samples. Inset: Rarefaction curve for all 1700 soil clones. The three whale falls are: 1, Santa Cruz Basin bone; 2, Santa Cruz Basin microbial mat; and 3, Antarctic bone.

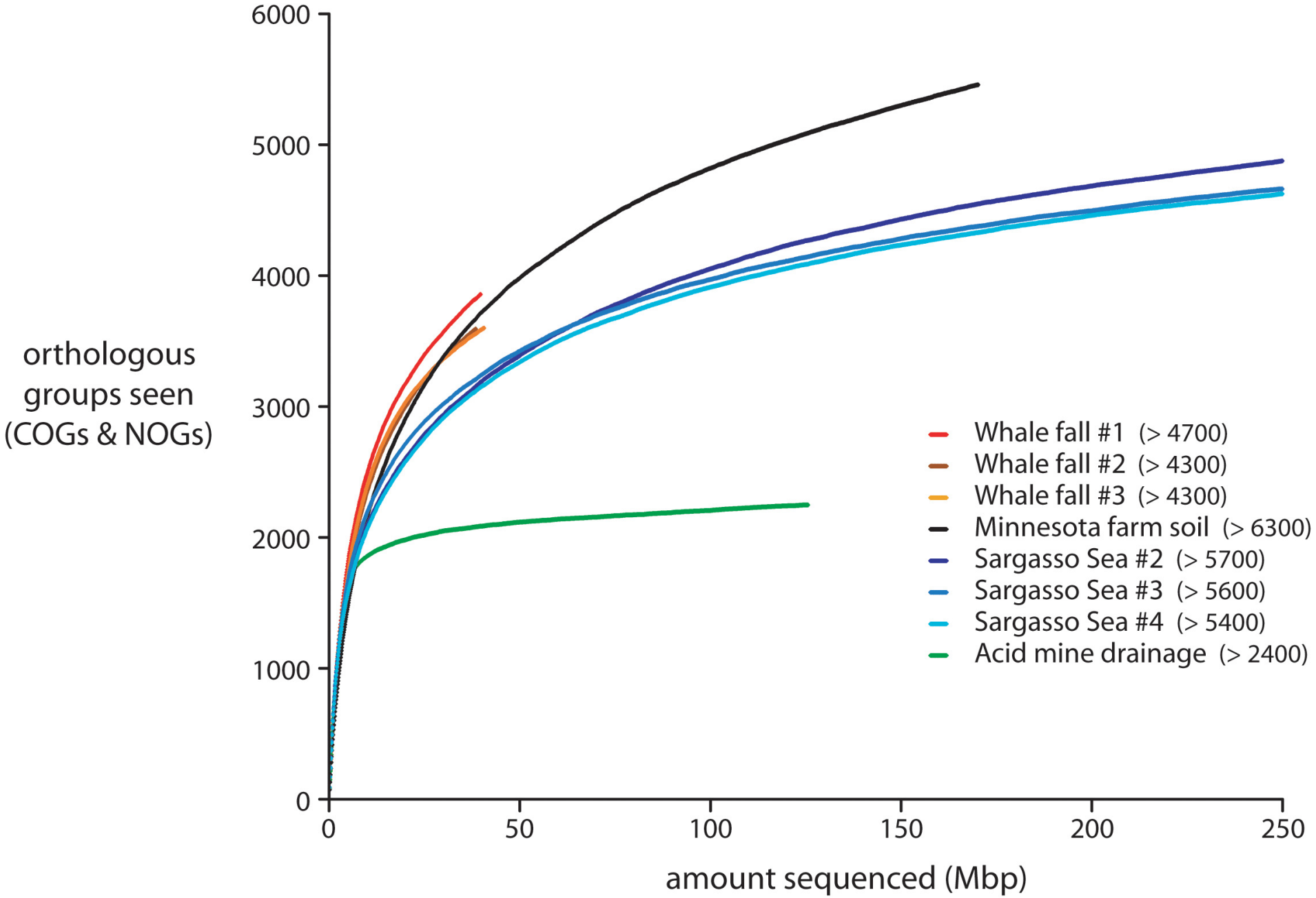
Figure 2: **Identification of orthologous groups with greater sequencing depth.** The number of orthologous groups observed at least once is shown as a function of the raw sequence generated. Numbers in brackets indicate lower limits of the total number of groups in the sample.

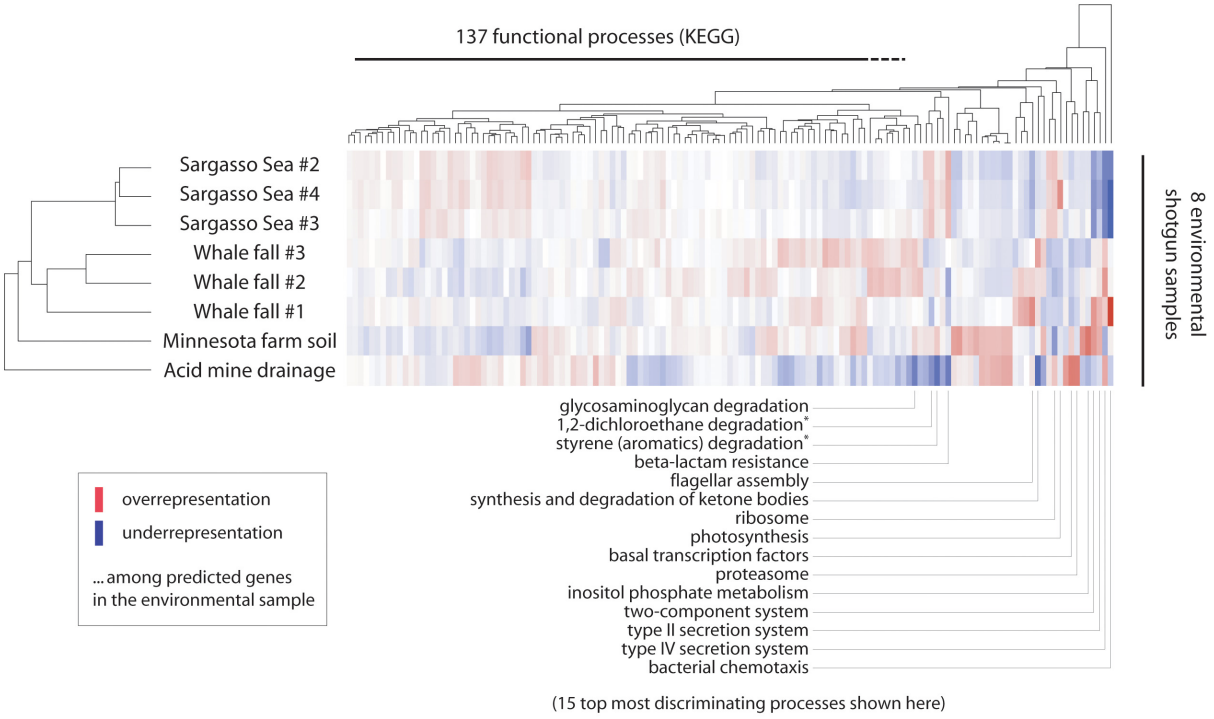
Figure 3: **Functional profiling of microbial communities.** Two-way clustering of samples and encoded functions based on relative enrichment of KEGG functional processes. The 15 most discriminating processes are highlighted.

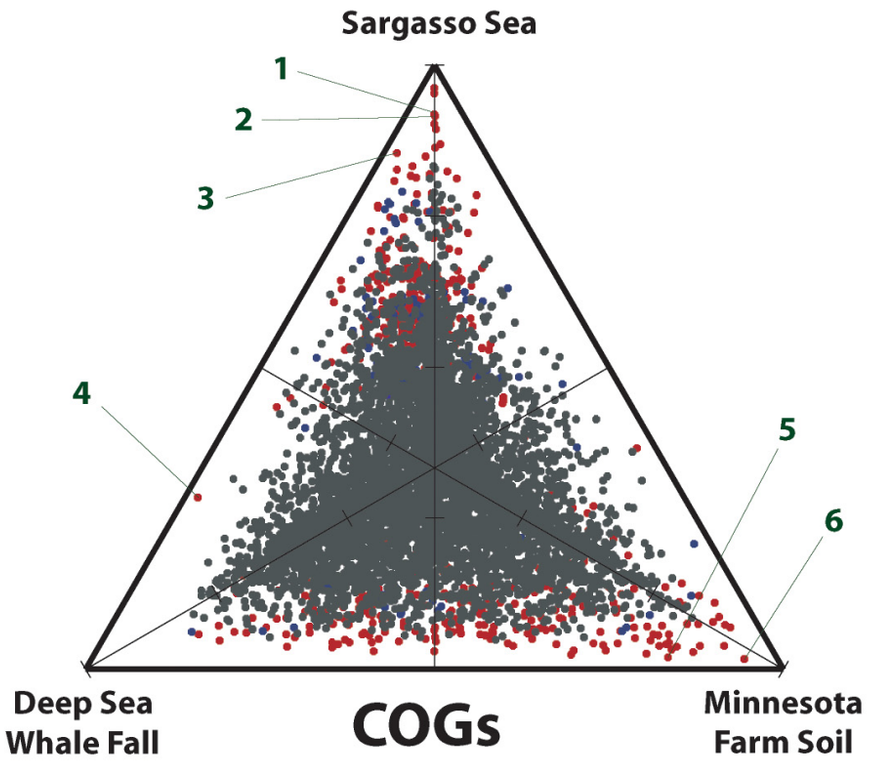
Figure 4: **Specific Enrichments**. Three-way comparisons of soil, whale fall and Sargasso Sea environments, in terms of COGs, operons, KEGG processes or COG functional categories. Each dot shows the relative abundance of an item in the three environmental samples, such that proximity to a vertex is proportional to the level of enrichment in the

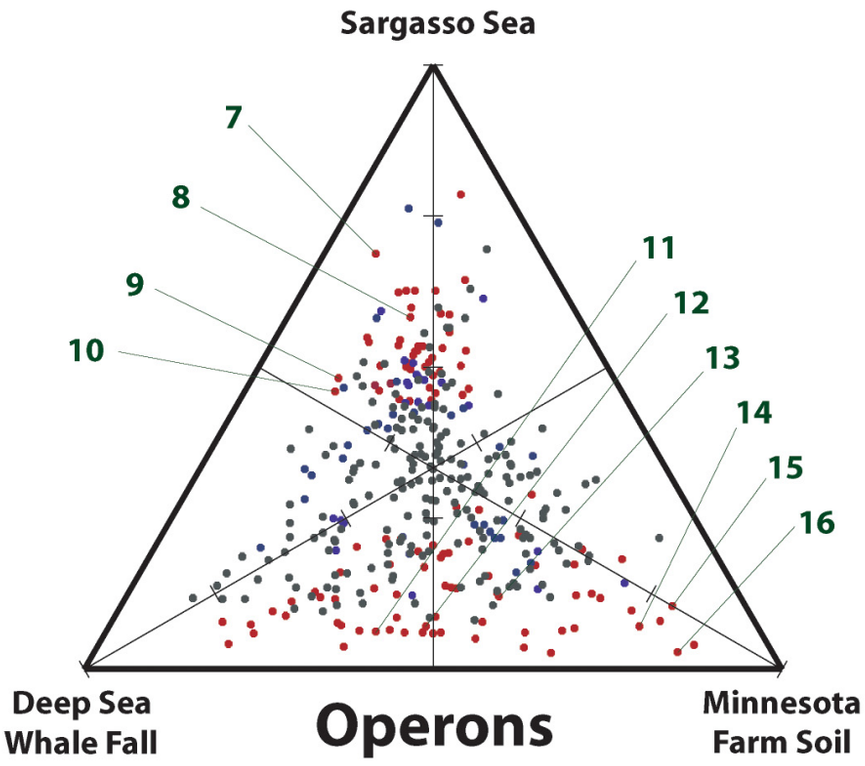
respective sample. Color indicates statistical significance of the enrichment. Marked items discussed in main text: 1) COG5524 – Bacteriorhodopsin. 5) COG3459 – Cellobiose phosphorylase. 7) ABC-type proline/glycine betaine transport system. 10) Na^+-transporting NADH:ubiquinone reductase. 14) Osmosensitive, active K^+-transport system. 18) Photosynthesis. 19) Type I polyketide biosynthesis (antibiotics). A complete listing of numbered items is available in the SOM, and an enhanced version of the figure is at http://string.embl.de/metagenome_comp_suppl/.

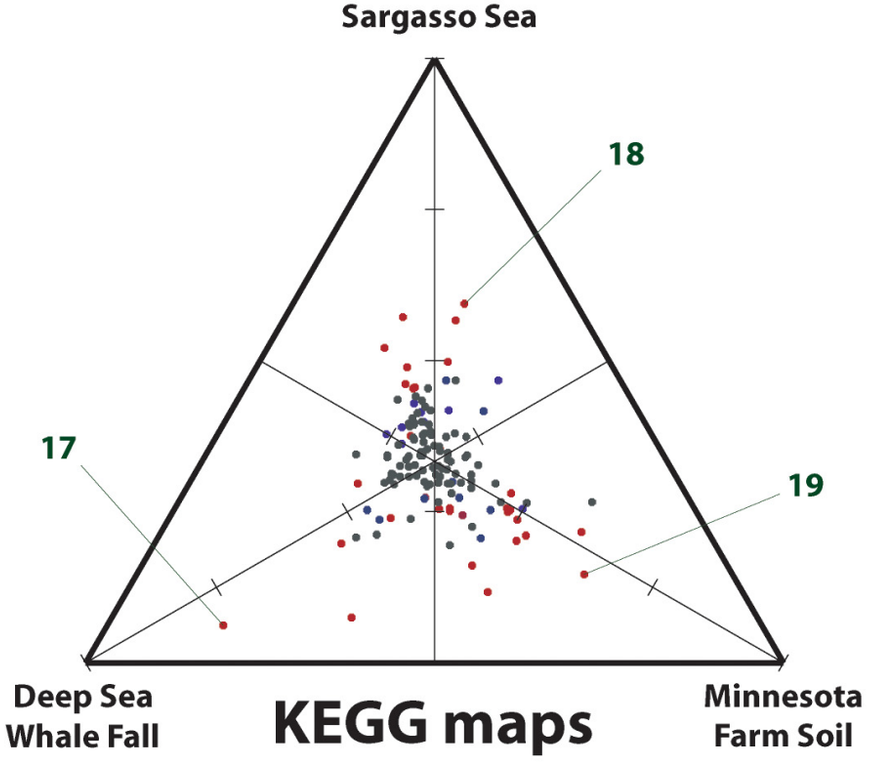

80
70
60
50
40
30
20
10
0
Phylotypes observed
0
20
40
60
80
16S clones sequenced
1500
1000
500
0
0
500
1000
1500
Whale fall 1
Whale fall 2
Whale fall 3
Soil

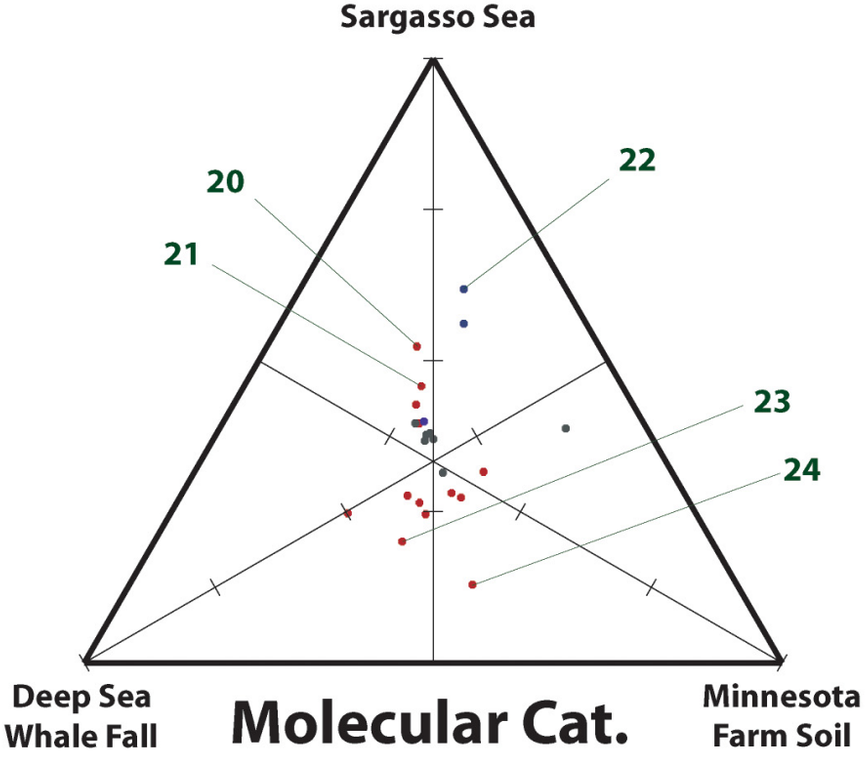

orthologous groups seen (COGs & NOGs)
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
amount sequenced (Mbp)
Whale fall #1 (> 4700)
Whale fall #2 (> 4300)
Whale fall #3 (> 4300)
Minnesota farm soil (> 6300)
Sargasso Sea #2 (> 5700)
Sargasso Sea #3 (> 5600)
Sargasso Sea #4 (> 5400)
Acid mine drainage (> 2400)


137 functional processes (KEGG)
Sargasso Sea #2
Sargasso Sea #4
Sargasso Sea #3
Whale fall #3
Whale fall #2
Whale fall #1
Minnesota farm soil
Acid mine drainage
8 environmental shotgun samples
overrepresentation
underrepresentation
... among predicted genes in the environmental sample
glycosaminoglycan degradation
1,2-dichloroethane degradation*
styrene (aromatics) degradation*
beta-lactam resistance
flagellar assembly
synthesis and degradation of ketone bodies
ribosome
photosynthesis
basal transcription factors
proteasome
inositol phosphate metabolism
two-component system
type II secretion system
type IV secretion system
bacterial chemotaxis
(15 top most discriminating processes shown here)


Sargasso Sea
1
2
3
4
5
6
Deep Sea
Whale Fall
COGs
Minnesota
Farm Soil


Sargasso Sea
7
8
9
10
11
12
13
14
15
16
Deep Sea
Whale Fall
Operons
Minnesota
Farm Soil


Sargasso Sea
17
18
19
Deep Sea
Whale Fall
KEGG maps
Minnesota
Farm Soil


Sargasso Sea
20
21
22
23
24
Deep Sea
Whale Fall
Molecular Cat.
Minnesota
Farm Soil


>10
1
10-1
10-2
10-3
<10-4
e-value of enrichment

Materials and Methods

Sample collection:

Surface soil (0-10 cm) was collected in September 2001 from a farm in Waseca County, Minnesota. The surrounding area had been used for livestock, including sheep, cattle, and pigs, and was in the drainage path of a silage storage bunker that had been used for sweet corn and pea silage waste operations from 1990 to 1997. The sample was collected and sealed in polyethylene bags and stored at 4°C for processing prior to archiving at -80°C. Biochemical analysis (Wallace Laboratories) on 20 g of soil from the same site revealed it to be clay loam, with fair to low organic matter content and high levels of most essential elements. Potassium was present at 926.15 mg/kg dry weight and sodium at 75.38 mg/kg; levels of most nonessential elements were low. Microscopic analysis, including Sybr green staining, found the organisms in the sample to be primarily prokaryotic.

Three other samples were from microbial communities growing on sunken whale skeletons, a lipid-rich nutrient source that can foster the growth of a flourishing ecosystem in an otherwise nutrient-poor environment (*S1*). This unique ecological niche, referred to as a "whale fall," has been suggested to select for "specialist" species in geographically remote locations (*S2*). Three independent whale fall sample libraries were examined. "Whale fall 1" is a section of a rib bone from a gray whale carcass experimentally sunk in 1998 in the Pacific Ocean, Santa Cruz Basin (N33.30 W119.22), at a depth of 1674 meters (*S3*). "Whale fall 2" is an orange microbial mat from the same whale carcass; both samples were collected using a remote operated vehicle (ROV). "Whale fall 3" is a whale bone of uncertain age and species collected by otter trawl on a muddy seafloor at a depth of 560 meters off the West Antarctic Peninsula Shelf (S65.10 W64.47).

Library construction:

DNA for all libraries was isolated as described (*S4*). For analysis of small ribosomal RNA sequences, three sets of primers were used to individually target bacterial (27F and 1392R), archaeal (21F and 958R) and eukaryotic (forward

primer: 5'-ACCTGGTTGATCCTGCCAG-3', reverse primer: 5'-TGATCCTTCYGCAGGTTCAC-3') genomes. Products were then cloned into the pCR4-Topo vector (Invitrogen).

For soil genomic sequencing, community DNA from 0.5 g material was cut with 6-base recognition site restriction enzymes and cloned into the lambda ZAP vector (*S5*). The library was amplified once then *in vivo* excised to form a phagemid library according to manufacturer's protocol. Average insert size was determined to be 2.4 kb by gel electrophoresis. All three whale fall libraries were made from mechanically sheared community DNA cloned into the lambda ZAP vector, then *in vivo* excised, without amplification, to form a phagemid library. Average insert sizes were 3.3 – 3.5 kb.

Clones for all libraries were picked and bidirectionally sequenced by standard protocols (http://www.jgi.doe.gov/).

16S/18S rRNA sequence analysis:

Paired reads from 16S and 18S rRNA clones were assembled using phrap (www.phrap.org); 18S clones with two successful sequencing reads that failed to assemble were manually assembled with Ns filling the central gap. Chimeric sequences were identified by the Bellerophon program (*S6*) and removed from further analysis. However, any sequences that appeared in both independent bacterial PCR libraries from soil were flagged as non-chimeric and retained. Species abundances were determined by a Perl script that utilized Megablast alignment and single-linkage clustering to group together any sequences with >97% identity over the full length of the insert; clustering was also performed with 98% and 99% identity cutoffs for comparison. Rarefaction curves and total species estimates were generated using EstimateS (Version 7, R. K. Colwell, http://purl.oclc.org/estimates). For phylogenetic assignment, all bacterial and archaeal sequences were blasted against an internal ARB database of curated 16S rRNA sequences; any sequences without hits of >95% identity, as well as all eukaryotic 18S rRNA sequences, were blasted against the NR database. Sequences with >95% identity to a database sequence were assigned to the same

phylum. For clusters that remained unassigned, a representative member was phylogenetically classified by incorporation into the ARB database tree via clustalw and manual refinement based on secondary structure. Singlets that could not be automatically assigned to a phylum remained unclassified.

Genomic sequence analysis:

Prior to annotation, low-quality and duplicate sequencing reads were removed from the soil sequence. Among the original set of 198529 reads, 8164 had fewer than 200 bases with phred score >20 and were therefore removed as being unlikely to contribute useful information. The remaining reads were then scanned for duplicate clones resulting from the amplification step in the library preparation. 41,280 reads were defined as duplicates and removed, using the criteria that any reads that matched each other with >95% identity over at least 400 bp or 90% of the insert length, and had the same insert orientation, were considered duplicates. When pairs of duplicates were found, the read with fewer high-quality bases was deleted from the data set. The remaining 149085 reads were subjected to phrap assembly, to identify overlapping reads from independent clones, and functional annotation for EGT analysis.

Metagenome size calculations:

To calculate the amount of metagenomic sequence needed to assemble the genome of the most common species in soil, we estimate based on the 16S rRNA data that this species represents roughly 5% of the library. Assuming an average genome size of 6 Mb (*S7-9*), and a desired coverage level of 8X, we would need to sequence 48 Mb of DNA from this organism. Accordingly, nearly a gigabase of sequence from this community would be necessary. However, significantly more could be needed if the 5% representation of this clone is inflated by biases such as preferential PCR amplification: if the ~3000 taxa were present in equal abundance, >150 Gb could be required.

To estimate the sequence coverage based on the assembly statistics from soil, we considered two extremes, in which either one species dominates or all species are present in equal abundance. In total, 744 contigs were identified in the phrap

assembly that contained reads from at least two independent clones and were longer than 850 kb. Within these contigs, roughly 0.3 Mb of sequence were covered more than once. We first assumed that these overlapping sequences all derived from the same 6 Mb genome. The Lander-Waterman equation indicates that the number of bases covered more than once will be equal to $G * (1 - e^{-c} - ce^{-c})$, for coverage c of a genome (or metagenome) of size G. Solving this equation, we estimate that the most abundant genome is covered at a depth of about 0.35 in our data, so achieving the 8X coverage desired for assembly would therefore require more than 2 Gb additional sequence. On the other extreme, if we assume that all species are present in equal abundance, the same equation predicts a total "metagenome" size of 16.7 Gb (~2800 individual genomes of 6 Mb) and implies that more than 130 Gb of sequence would be required for genome assembly. Thus both the 16S rRNA data and the assembly statistics independently project the need for an amount of sequence on the order of one to a hundred gigabases in order to assemble one or more prokaryotic genomes from the soil community.

Whale fall "metagenome size" estimates were calculated by determining the coverage of each base sequenced and fitting the resulting histogram. Assuming an average genome size of 6 Mb and a desired coverage of 8X, the amount of sequence necessary to assemble the three most abundant genomes (roughly 50% of the community) in whale falls 1, 2 and 3 respectively are: 257 – 520 Mb, 270 – 698 Mb, and 240 – 486 Mb. Achieving sufficient coverage of all genomes present at an abundance of at least 2% in any sample would require 2.4 Gb of sequence.

Functional annotation:

All genomic sequences from soil, whale falls and acid mine drainage were analyzed by the program FGENESB from Softberry, which predicts genes and operons as well as functional RNAs (described at http://www.softberry.com). Analysis of Sargasso Sea sequences utilized the previously reported protein sequence predictions (*S10*). Functional annotation of the predicted proteins

utilized an extended version of the COG database, covering 26201 protein families (orthologous groups) in 179 completely sequenced organisms as compared to 10740 orthologous groups in 73 organisms found in the current COG database (*S11*). The extension has been performed using an unsupervised procedure in the context of the STRING project (*S12, 13*). As a result of the extension, additional members have been added to existing COGs, and novel orthologous groups have been created which are termed "non-supervised orthologous groups" (NOGs). The procedures used for extending the database were very similar to the original COG procedures (including a 'COGNITOR'-type protocol for extension of existing COGs, and full all-against-all similarity searches to define novel groups as triangles of reciprocal best hits; see the last chapter of the STRING documentation for details: http://string.embl.de/). The extended COGs used here are those of STRING version 6; they are transitional in that they will be replaced when updated versions of the original COG database are released.

Predicted proteins from all environments, including those from unassembled reads and those annotated as 'miscellaneous feature' in the Sargasso Sea data, were compared to this extended COG database using BLASTP. Predicted proteins were assigned to one of the orthologous groups if they showed a similarity score of 60 bits or better to any of the proteins in that group. BLASTP was run using the BLOSUM62 matrix and low-complexity filtering disabled (under these settings, 60 bits corresponds to an e-value of roughly 10^{-8} in searches against nrdb). As is the case in the original COG database, a protein was allowed to map to several orthologous groups, provided all of these were detected above the 60 bits cutoff and overlapped by no more than 50% of the shortest assignment. Based on the COG assignments, proteins were also assigned to operons and higher functional categories for further analysis (described below). A separate BLASTP against the KEGG database was used to assign proteins to KEGG maps. This was again done using BLASTP at a cutoff of 60 bits, but each environmental protein was mapped to at most one protein in the KEGG database.

When this procedure was applied to the environmental sequences, the predicted soil proteins mapped to 5467 distinct orthologous groups (3394 to the original COGs and 2127 to additional, automatically derived, non-supervised orthologous groups or NOGs), each whale fall library contained representatives of ~3600 groups and each Sargasso Sea library contained representatives of ~4800 groups. The predicted AMD proteins, on the other hand, mapped to just 2244 groups, consistent with the limited diversity of this community.

COG accumulation curves were generated by examining each read individually, in random order, and assessing the number of bases in the read and the number of previously unseen COGs assigned to that read. This analysis utilized raw, untrimmed reads; for Joint Genome Institute data the number of quality bases determined with a Phred score 15 threshold is typically 64% of the raw read length. Chao1 estimates of total COG content were obtained using EstimateS (Version 7, R. K. Colwell, http://purl.oclc.org/estimates).

Two-way clustering analysis:

To investigate whether independent samples taken from related environments show a similar functional profile in terms of encoded proteins, a two-dimensional cluster analysis was performed - akin to the clustering of microarray data (*S14*). Each sample was treated independently, including each of the separate sample libraries from the Sargasso Sea; we chose to focus on samples 2, 3 and 4 because they were isolated from different locations utilizing the same sampling protocol, specifically the same size prefilter and collection filter.

A two-dimensional matrix was constructed of environmental samples and orthologous groups, wherein each cell indicates how often genes of a particular orthologous group were seen within a particular environmental sample. To achieve optimal sensitivity and specificity, this was done based on assembled data wherever possible, correcting for the read-depth of the assembled contigs (a contig with a high read-depth is more frequently represented within the sample and correspondingly receives a higher count). Corrections for mated reads and contig sizes were also performed: mated reads do not constitute independent

observations, and large contigs are clearly covered by more reads than short contigs. Thus, final counts were expressed as number of independent clones per 1000 base pairs of assembly, and those final counts were equally applied to all orthologous groups found within a contig. In a last step, we added to those final counts a small amount of pseudocounts, in order to suppress meaningless statistical fluctuations caused by very rare orthologous groups (the amount of pseudocounts added to each cell was the sum of all cells of that environment, divided by 10000).

At this point, the matrix was normalized to account for the varying amounts of sequence acquired for each environmental sample, and for the varying overall frequency of orthologous groups. Normalization of the rows to unity (i.e. the environments) corrected for sequencing depth, and a subsequent normalization of columns to unity corrected for the overall frequency of orthologous groups (some groups such as unspecific methylases or dehydrogenases are generally very frequent in microbial genomes, and would dominate over less-frequent, but more specific groups without this last normalization). The matrix was then clustered independently in each dimension, using UPGMA clustering of Euclidian distances (PHYLIP package). Figure S6A shows the final matrix, rearranged according to the result of the clustering - whereby cells in the matrix are colored to indicate whether the orthologous group in that particular environment is seen more often than expected, or less often (colors represent log-ratios, i.e. observation divided by the unbiased expectation: two-fold overrepresentation is shown in full red, two-fold underrepresentation is shown in full blue, white color means observation is as expected). The matrix shown is truncated after 600 orthologous groups due to space constraints, but the clustering of samples is based on all available groups. The 600 groups shown are those with the overall largest deviation from the expectation (i.e. the product of their matrix cells is minimal).

The above analysis was repeated for functionally binned genes (as opposed to single genes), in order to assess whether the resulting tree of environmental samples was robust, and to assess which functional systems differed most between samples. Grouping of genes was performed at two levels: at the level of

operons (averaging 4.5 genes per operon), and at the level of the functional process (as defined in the KEGG database (*S15*), averaging 15 genes per process and species).

Not all bacterial operons are known, but a comprehensive list of presumed operons can be constructed by searching for repeatedly occurring gene neighborhoods in fully sequenced prokaryotic genomes. We have previously executed such a search (*S16*) and have extended it here to cover 179 fully sequenced genomes. In short, all orthologous groups in all genomes were assayed for neighboring occurrences or instances where two groups mapped to the same ORF (gene fusions). The resulting links between orthologous groups were scored according to frequency and specificity of the interaction, and then clustered to reveal entire operons. The procedure and cutoff applied here were essentially identical to those used previously (*S16*), except that neighborhood and fusion were considered but not the phylogenetic co-occurrence of genes across species. The resulting set of conserved operons consisted of 565 operons of at least three orthologous groups each. Of those, 394 operons were found within at least one of the environments. Note that this does not require the presence of multiple genes on a single contig – what is counted are still the individual orthologous groups (as in the above paragraph), but these are subsequently grouped according to their membership in known operons. Construction of the two-dimensional matrix and clustering were done as described above; pseudocounts were 1 in 10000, and full color is shown for enrichments of 1.5-fold or higher.

For the two-way clustering according to KEGG processes, the predicted environmental proteins were directly compared to proteins in the KEGG database (bypassing the COG-assignment). The two-dimensional matrix was then constructed using entire KEGG-processes, each grouping the counts for several proteins. Filling and clustering of the matrix were done as above; pseudo-counts were 1 in 2000 (reflecting the larger size of KEGG processes), and full color was shown for enrichments of 1.3-fold or higher.

Specific enrichments (three way comparisons):

Having established that similar environmental samples can be grouped together based on their gene content, the next task was to assess which genes were particularly enriched in each of the environments (hinting at functional differences among the microbial communities). For this analysis, the three whale fall samples were pooled as one environment, as were the three Sargasso Sea samples #2, #3 and #4. The acid mine drainage sample was not considered here, because it is the least diverse sample and because it is from a relatively recent, man-made environment. Samples 2-4 from the Sargasso Sea were chosen because they were independent samples utilizing identical sampling procedures. A triangular representation was chosen to display the specific enrichments of genes or functional processes in each of the environments (Figure 4, main text). Assessing the relative counts of orthologous groups, operons or KEGG processes was done exactly as described in the previous section (two-way clustering analysis). Additionally, as a fourth binning the assignment of orthologous groups to broad functional categories was used (categories were as defined in the COG database).

For each item, one dot is shown within a triangle – the position of the dot signifies the relative enrichment of the item in one or several of the samples. Items that are equally frequent in all three environments appear in the middle of the triangle. Items that appear in one of the corners of the triangle are found primarily in one of the environments, and items that appear along one of the edges of the triangle are found primarily in two of the three samples, but are largely absent from the third. For each item, the relative counts for the three environments were normalized to add up to 1 (after addition of pseudocounts to select against rare items). This permitted the display of three-dimensional data in two dimensions (using three axes at 120 degree angles). In order to estimate the statistical significance of each observation, the data were compared to randomized data, as follows.

First, the actual items were binned into abundance classes. An observed relative enrichment is statistically more significant for an abundant item (e.g. a widespread orthologous group or a large operon) than for a rare item. Comparison of items to randomized data was done separately for each abundance class. Randomization was done by repeated sampling of items from reservoirs matching the size distributions of the three environmental samples. For each random sampling, the addition of pseudocounts and normalization were done in exactly the same way as for the actual data, and the position of the random dot in the triangle was noted. After at least $2 \cdot 10^6$ randomizations in each abundance class, the density of random dots in the triangle was assessed, on a grid spanning 20 bins on each axis (i.e. 20 * 20 * 20 = 8000 gridpoints). This allowed the computation of p-values for each of the actual items, by checking the density of random dots at the position of the item: the p-value corresponded to the number of random dots in bins of equal or lower density, divided by the total number of randomizations. E-values were then computed by multiplying the p-values with the total number of items under consideration. For each of the triangles, dot positions and e-values of all items are available as flat files on request.

Supplemental Data

16S / 18S ribosomal RNA analyses:

The bacterial 16S rRNA sequences from soil (1700 total) clustered into 847 unique groups mapping to 18 different phyla when single-linkage clustering was performed with a 97% identity threshold. The number of unique groups rose to 1034 when this threshold was raised to 98%; at a 99% threshold, essentially the limit of the error inherent in the sequencing quality, 1467 unique sequences were identified. Total diversity, based on the Chao1 estimator, was estimated to be more than 3500 phylotypes (at 97% threshold) but may be considerably more as

the estimate continued to increase with sequencing (Figure S1A). Values of the alternative ACE (Abundance Coverage Estimator) estimate of total species richness were more stable and plateaued at 3000 phylotypes. Most of these sequences were singlets, and the largest cluster contained 112 clones, or 6.6% of the total (Figure S2A). As a result of the single-linkage clustering, however, some sequences within the cluster were as little as 95% identical; at a higher identity cutoff of 98% the cluster broke into several smaller clusters, the largest of which contained 81 (4.8%) of the clones. At a 99% identity cutoff, the cluster essentially disappeared. The 58 archaeal clones formed just seven clusters, all within two major euryarchaeal branches (Figure S2B), and the 106 eukaryotic 18S sequences analyzed fell into 35 distinct groups in at least 8 different phyla, primarily fungi and unicellular eukaryotes. 33 partial 16S rRNA sequences were found in the soil genomic data, representing 31 distinct bacteria, one archaeon and one chloroplast; one eukaryotic 18S sequence was also found.

Each whale fall bacterial 16S library contained 17-37 unique sequences mapping primarily to the Proteobacteria and Bacteroidetes. In contrast to soil, more than half of the sequences were distributed among the top few clusters (Figure S3A). The archaeal 16S sequences from the two Pacific samples fell into a limited number of clusters, primarily within the Methanomicrobia and, for the mat sample, the C1 archaea. The eukaryotic 18S sequences from the mat sample were all from the same deeply branching eukaryote while those from the bone derived mainly from two alveolates; singlet representatives of a cercozoan and a fungus were also found in this library. We found partial 16S rRNA sequences in 74, 36

and 64 clones from the three whale fall libraries respectively, all of which were bacterial. Comparing these to the sequences found in the PCR clone libraries revealed that for each sample, the same phyla (and proteobacterial classes) were typically represented in the two types of libraries (Supplementary Figure 2D).

Comparison of assembled acid mine drainage biofilm genomes with unassembled reads:

Automated annotation was applied to the assembled genomic scaffolds from the acid mine drainage biofilm as well as to all unassembled reads from the same sample. In the five genome "bins" assembled from the acid mine drainage sequence data, a total of 7173 distinct proteins were predicted in 1629 different COG categories. In the complete set of unassembled reads, 77685 proteins were predicted in 1824 different COG categories (of 144771 total predicted ORFs), including all but 8 of the categories predicted in the assembled genomes. 203 additional COGs were predicted in the unassembled data that were not predicted in the assembled genomes, of which slightly more than half (107) were predicted in reads that were discarded because they did not form large contigs. More stringent methods for assigning proteins to COGs, such as requiring multiple hits to the same category in different organisms, did not substantially change the number of apparent false positives or false negatives.

Sample-specific enrichments:

Beyond the gene content variations described in the main text, numerous differences in distribution of functional proteins across samples were observed, both expected and unexpected. Several of these, labeled with numbers on Figure

4, are 1) COG5524 – Bacteriorhodopsin. 2) COG4338 – uncharacterized protein conserved in bacteria. 3) COG3046 – uncharacterized protein related to photolyase. 4) COG1292 – Choline-glycine betaine transporter. 5) COG3459 – Cellobiose phosphorylase. 6) COG3903 – Predicted ATPase domain of unknown function. 7) ABC-type proline/glycine betaine transport system. 8) Ribosomal subunit operon. 9) Phosphonate transport and metabolism. 10) Na^+-transporting NADH:ubiquinone reductase. 11) Nitrous oxide reductase. 12) Nitric oxide reductase. 13) Nitrate reductase. 14) Osmosensitive, active K^+-transport system. 15) DNA double strand break repair system (NHEJ-type). 16) Uncharacterized, soil-specific operon. 17) Type IV secretion systems. 18) Photosynthesis. 19) Type I polyketide biosynthesis (antibiotics). 20) Translation. 21) Nucleotide transport and metabolism. 22) Eukaryotic RNA-processing and modification. 23) Defense mechanisms. 24) Signal transduction mechanisms.

At the COG level, there are several uncharacterized genes displaying extreme bias toward particular environments; for example, COG4338, COG3046 and COG4240 are found almost exclusively in the Sargasso Sea; COG3903 is heavily skewed towards soil and COG3550 is primarily observed in whale falls. Several putative genes categorized as NOGs are also quite unevenly distributed.

Among the operons, there were a number of apparent variations in systems other than transport. In terms of available electron acceptors, the deep sea whale falls share much in common with soil, including an enrichment of all three types of nitrate respiration processes (i.e. subunits of nitrous oxide reductase, nitrite oxide reductase, and nitrate oxide reductase). We also observe a strong enrichment (e-

value < 0.05) in the soil of a small operon recently shown to encode a prokaryotic double-strand break (DSB) repair system (Figure 4, upper right)(*S17*). This suggests that the resident microbes may have had a higher chance of suffering a DSB, or greater difficulty repairing it via recombinational repair, possibly because of factors such as larger genome sizes, slower growth, desiccation or attack from DSB-inducing genotoxins.

One of the most prominent overrepresentations among the KEGG maps was an abundance of proteins involved in Type IV secretion systems in the whale fall samples; evidence of this was also apparent in the COG, operon and higher functional category analyses. Chemotaxis and flagellar assembly pathways were also prominent in this exotic environment, providing potential clues to the processes involved in its colonization.

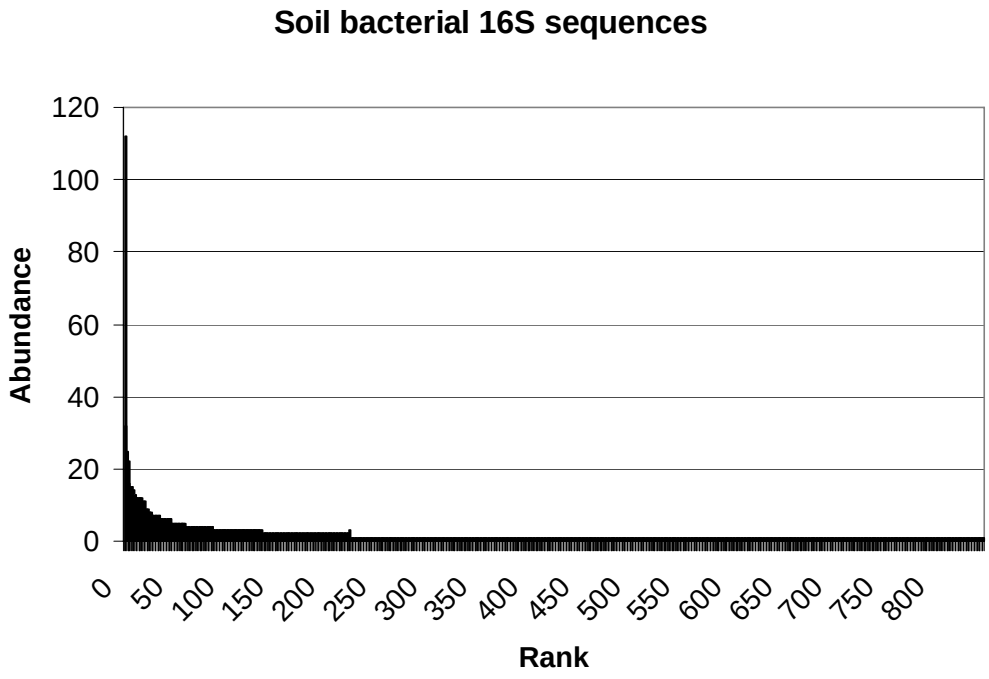
As mentioned in the main text, the differences among higher functional categories seem to suggest differences in genome size and/or phylogeny. Signal transduction genes, known to be more common in large genomes, are overrepresented in soil and whale falls while housekeeping functions like translation are overrepresented in the smaller genomes of Sargasso sea organisms (Figure 4, lower right)(*S18, 19*). The greater prevalence of RNA processing genes in the Sargasso Sea is indicative of a significant eukaryotic component in these samples (*S10*).

Figures and legends

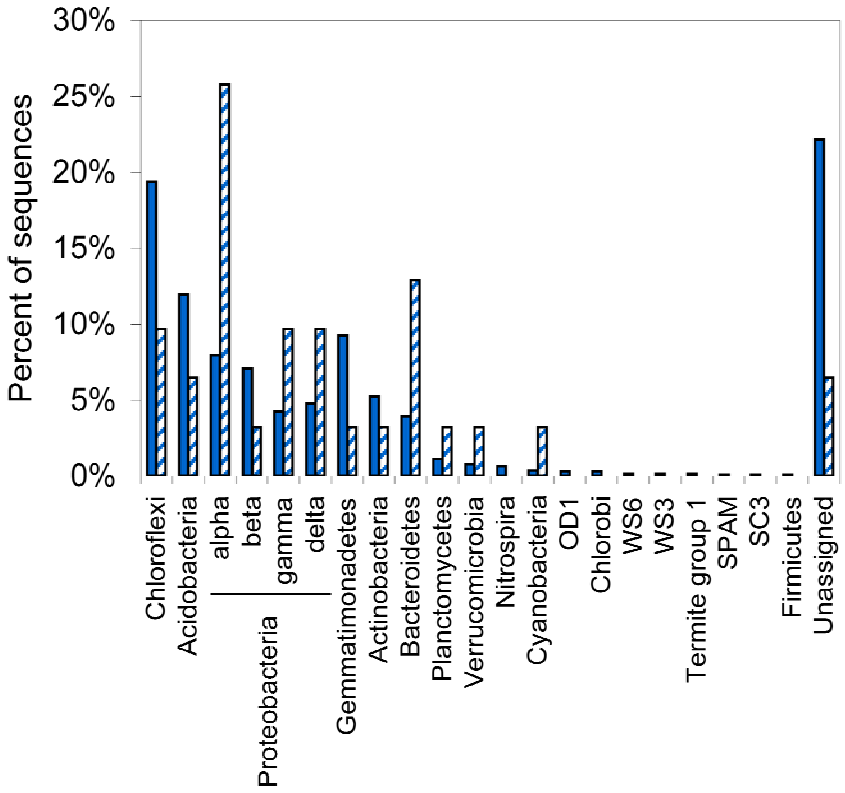
Figure S1: Rarefaction curves for 16S phylotypes observed (blue triangles), Chao1 total richness estimator (blue line), and ACE total richness estimator (dotted blue line) for soil.

Figure S2: rRNA analysis of soil. A) Rank-abundance curve for bacterial 16S rRNA sequences. B) Phylogenetic distribution of soil 16S rRNA sequences from PCR clone library (solid) and genomic library (hatched). C and D) Allocation of C) archaeal 16S and D) eukaryotic 18S rRNA sequences into phyla.

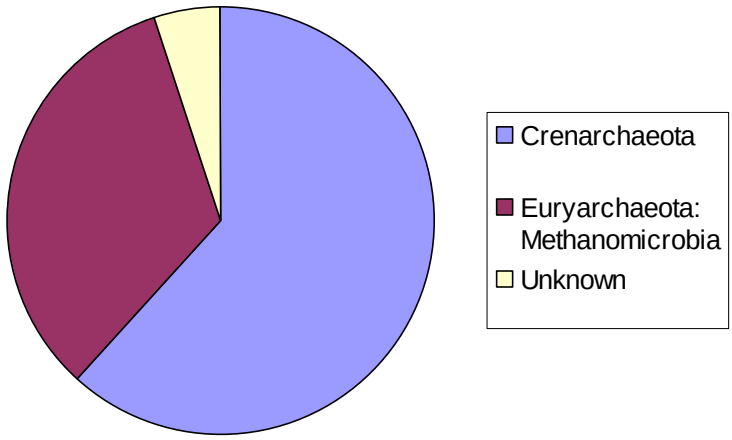
A)



B)


30%
25%
20%
15%
10%
5%
0%
Percent of sequences
Chloroflexi
Acidobacteria
alpha
beta
gamma
delta
Proteobacteria
Gemmatimonadetes
Actinobacteria
Bacteroidetes
Planctomycetes
Verrucomicrobia
Nitrospira
Cyanobacteria
OD1
Chlorobi
WS6
WS3
Termite group 1
SPAM
SC3
Firmicutes
Unassigned

C)


Crenarchaeota
Euryarchaeota: Methanomicrobia
Unknown

D)

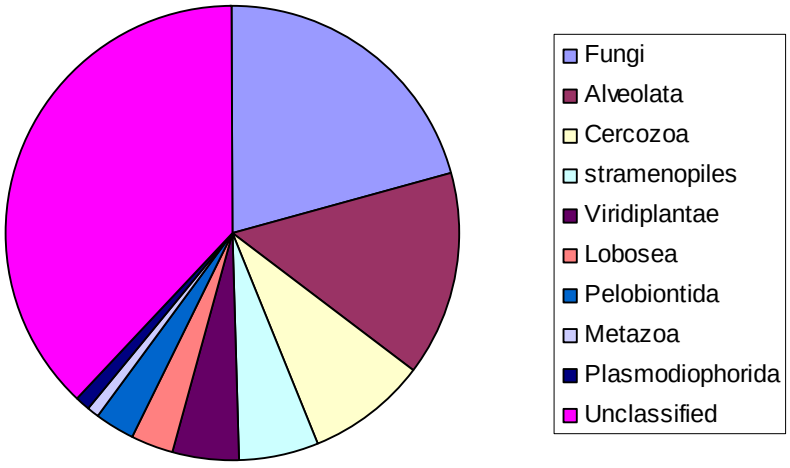

Fungi
Alveolata
Cercozoa
stramenopiles
Viridiplantae
Lobosea
Pelobiontida
Metazoa
Plasmodiophorida
Unclassified

Figure S3: Rarefaction curves for 16S phylotypes observed (triangles), Chao1 total richness estimator (lines), and ACE total richness estimator (dotted lines) for 3 whale falls. Whale fall 1, dark green; whale fall 2, bright green, whale fall 3, light green.

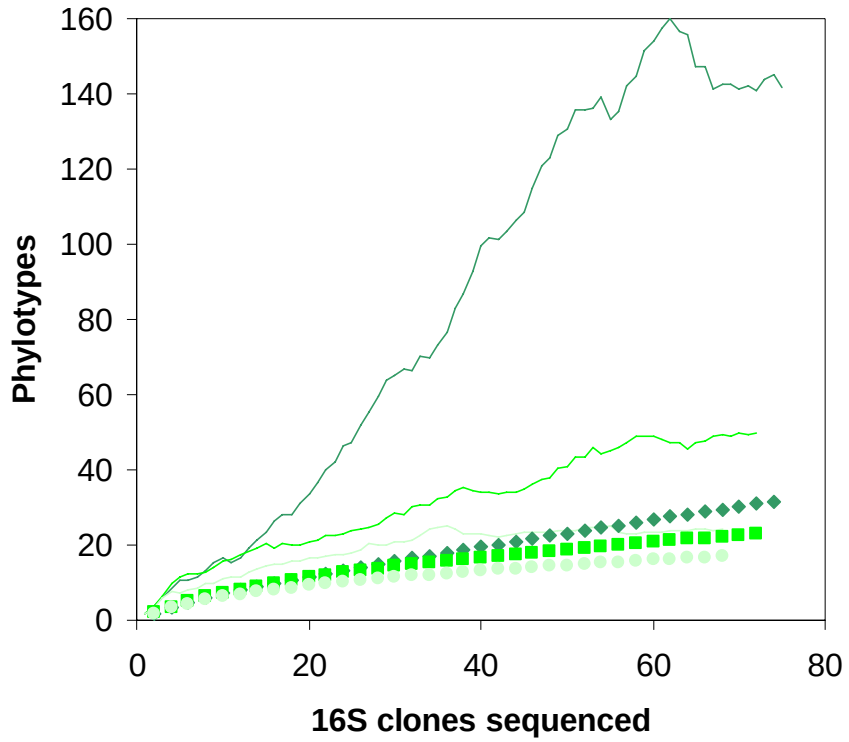
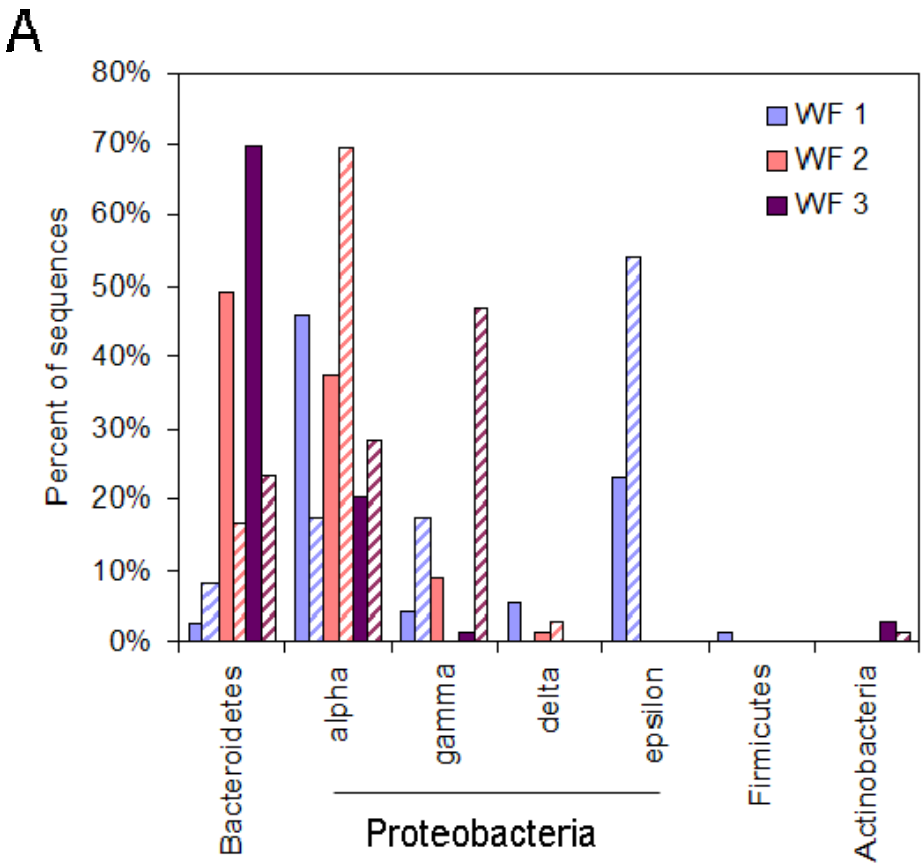
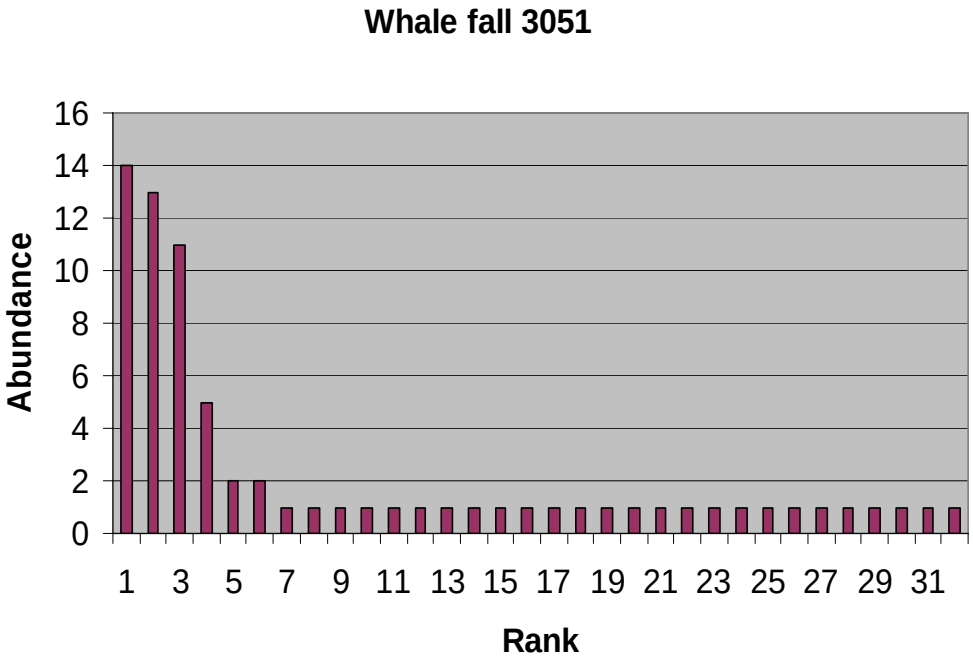


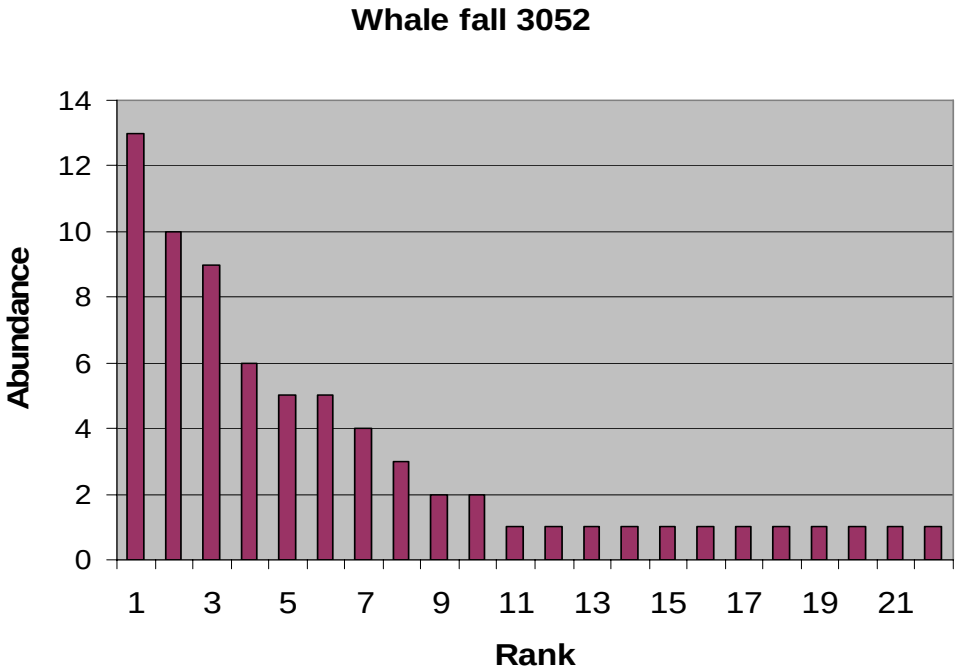
Figure S4: Rank-abundance curves for whale fall bacterial 16S sequences. A) Assignment of 16S rRNA sequences to bacterial phyla for both PCR clone libraries (solid bars) and genomic libraries (hatched bars). B) Whale fall 1, Santa Cruz bone; C) Whale fall 2, Santa Cruz microbial mat; D) Whale fall 3, Antarctic bone.



B)


Whale fall 3051
Abundance
16
14
12
10
8
6
4
2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Rank

C)


Whale fall 3052
Abundance
14
12
10
8
6
4
2
0
1 3 5 7 9 11 13 15 17 19 21
Rank

D)

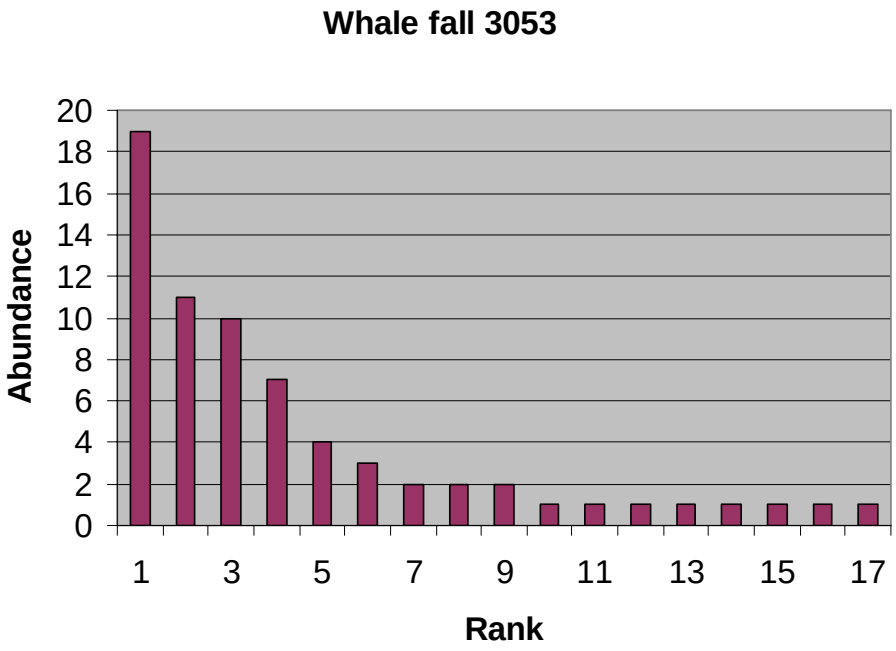

Whale fall 3053
Abundance
20
18
16
14
12
10
8
6
4
2
0
1
3
5
7
9
11
13
15
17
Rank

Figure S5: Functional accumulation curves for all samples. Number of unique hits in the A) COG and B) Pfam database as a function of sequence depth. The y-axis maximum is set to the total number of categories in each database.

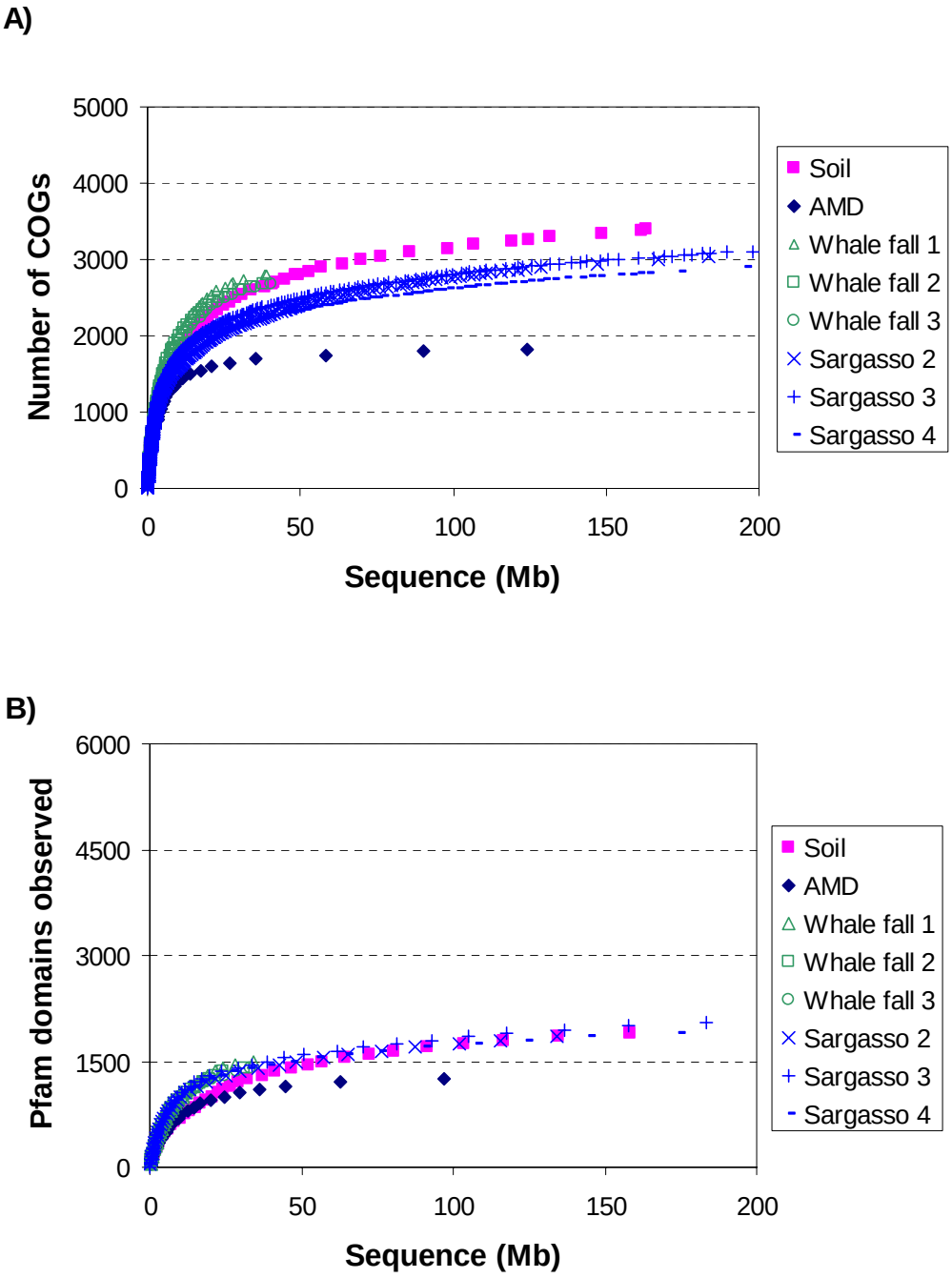


Figure S6: Two-way clustering of data based on A) COGs and B) operons.

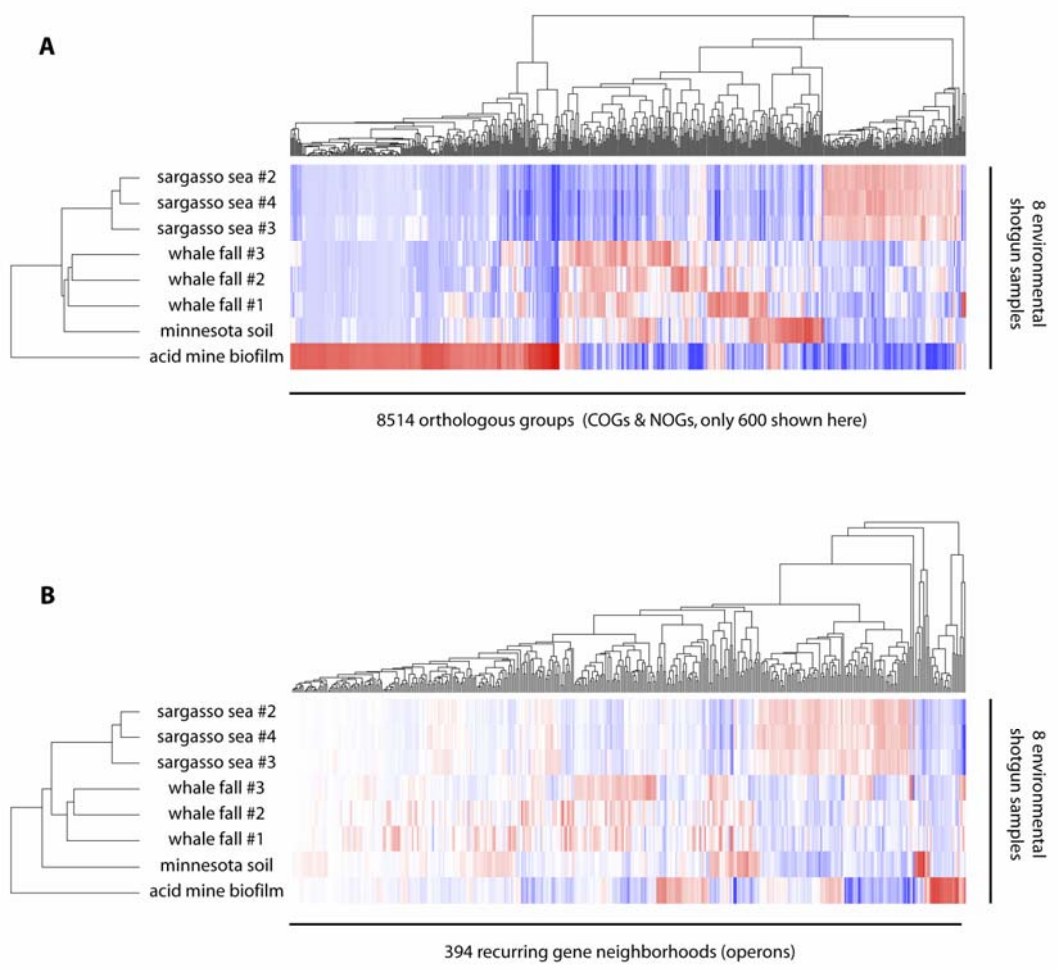
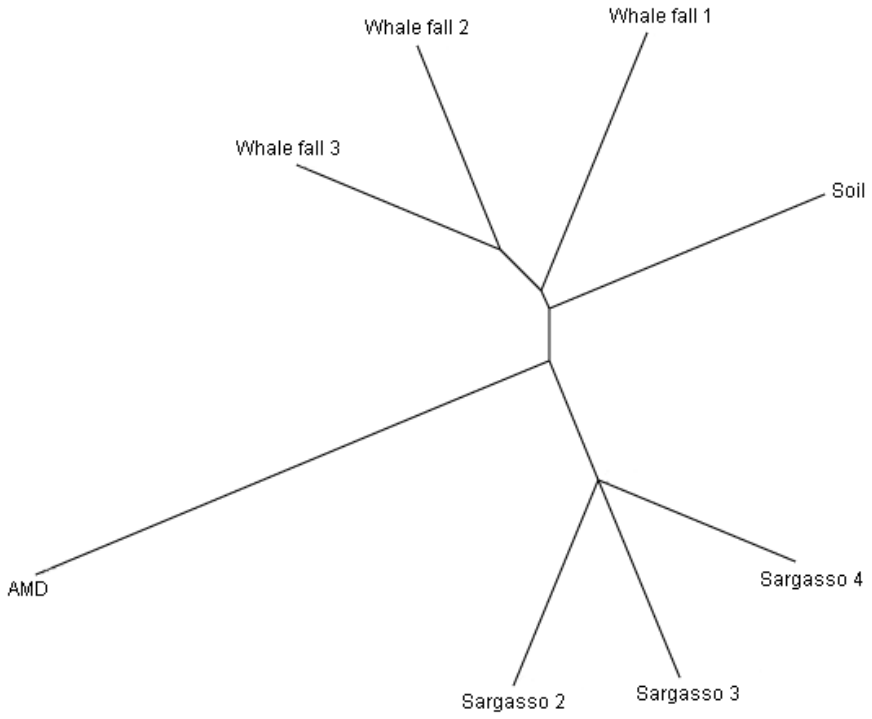


Figure S7: Sample tree based on 10 Mb of unassembled sequence from each sample. Total hits to each of 4873 COGs were taken as components of a COG vector; Euclidean distances were calculated among the vectors to create a distance matrix. Tree was generated using Phylip (University of Washington, http://evolution.genetics.washington.edu/phylip.html) and visualized with Phylodendron (University of Indiana, http://www.es.embnet.org/Doc/phylodendron/treeprint-form.html).



References

S1. C. R. Smith, H. Kukert, R. A. Wheatcroft, P. A. Jumars, J. W. Deming, *Nature* **341**, 27 (Sep 7, 1989).

S2. G. W. Rouse, S. K. Goffredi, R. C. Vrijenhoek, *Science* **305**, 668 (Jul 30, 2004).

S3. C. R. Smith, A. R. Baco, in *Oceanography and Marine Biology: an Annual Review* R. N. Gibson, R. J. A. Atkinson, Eds. (Taylor & Francis, 2003), vol. 41, pp. 311-354.

S4. D. E. Robertson *et al.*, *Appl Environ Microbiol* **70**, 2429 (Apr, 2004).

S5. J. M. Short, J. M. Fernandez, J. A. Sorge, W. D. Huse, *Nucleic Acids Res* **16**, 7583 (Aug 11, 1988).

S6. T. Huber, G. Faulkner, P. Hugenholtz, *Bioinformatics* (Apr 8, 2004).

S7. T. Kaneko *et al.*, *DNA Res* **7**, 381 (Dec 31, 2000).

S8. S. D. Bentley *et al.*, *Nature* **417**, 141 (May 9, 2002).

S9. C. K. Stover *et al.*, *Nature* **406**, 959 (Aug 31, 2000).

S10. J. C. Venter *et al.*, *Science* **304**, 66 (Apr 2, 2004).

S11. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (Sep 11, 2003).

S12. C. von Mering *et al.*, *Nucleic Acids Res* **31**, 258 (Jan 1, 2003).

S13. C. von Mering *et al.*, *Nucleic Acids Res* **33 Database Issue**, D433 (Jan 1, 2005).

S14. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (Dec 8, 1998).

S15. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, *Nucleic Acids Res* **32 Database issue**, D277 (Jan 1, 2004).

S16. C. von Mering *et al.*, *Proc Natl Acad Sci U S A* **100**, 15428 (Dec 23, 2003).

S17. G. R. Weller *et al.*, *Science* **297**, 1686 (Oct 6, 2002).

S18. E. van Nimwegen, *Trends Genet* **19**, 479 (Oct, 2003).

S19. K. T. Konstantinidis, J. M. Tiedje, *Proc Natl Acad Sci U S A* **101**, 3160 (Apr 2, 2004).

*This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.