Comparative Metagenomics of Microbial Communities

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

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

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

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

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

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

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

Assembled genomes of environmental microbes are informative but difficult to obtain due to the resistance of most organisms to cultivation and the diversity of microbial communities. Here we show that information encoded within fragmented sequence data alone can be used to characterize distinctive metabolic capabilities of microbial communities occupying different environmental niches. We sequenced microbial DNA isolated from a phylogenetically complex soil sample and three separate deep ocean whale skeletons and compared them to each other and to previously generated environmental genomic data. Confirming the complexity that microbial communities can attain, our analysis of 100 million bp from soil revealed that at least ten-fold more sequence would be required to assemble even the single most predominant community member. Despite the resulting low sequence coverage of individual microbial genomes in the samples, protein functions predicted from the fragmented sequence data for each environment revealed habitat-specific fingerprints. Binning of these proteins into either gene families, operons or cellular processes produced distinct patterns that correlate with the known metabolic demands presented by the different environments. The identification of uncharacterized, environment-specific and -enriched genes illustrates the unique insights to be gained from gene- versus genome-centric comparative analyses of environmental samples.

Despite their importance, relatively little is known about the microbes inhabiting most ecosystems on our planet due to their resistance to culture under standard laboratory conditions. An early glimpse into the phylogenetic diversity of these

as-yet uncultured organisms has been offered by a variety of 16S ribosomal RNA (rRNA)-based sequencing projects (*1, 2*). Recently, advances in high-throughput sequencing technologies have facilitated efforts to go beyond 16S rRNA sequencing to explore the genomic complement of environmental microbes. These pursuits, dubbed "metagenomics" or "ecogenomics," have provided valuable insight into the lifestyles and metabolic capabilities of uncultured organisms. They include the sequencing of individual large-insert BAC clones and more recently high-throughput sequencing of small insert libraries made directly from environmental DNA (*3-7*). While the latter - shotgun sequencing of environmental samples - can present challenges for genome assembly, this approach has the advantage of providing a minimally biased view of the community and its entire genetic repertoire absent from 16S rRNA or BAC clone sequencing surveys.

Two studies to date have reconstructed partial genomes from environmental DNA using shotgun sequencing. The first study to report genome assembly from an environmental sample focused on an unusually simple community, a biofilm from an acidic mine environment (*6*). With a relatively modest sequencing effort (~76 Mb), sufficient genome coverage was obtained to reconstruct near complete or partial genomes for five organisms. The second metagenomic study explored a considerably more complex set of microbial communities, found in the open waters of the Sargasso Sea (*7*). Previous studies had suggested that samples from this environment contained on the order of 200 species (*8*), but extensive sequencing (~1.4 Gb total) implied the presence of more than 1000 species in the

samples examined. Near-complete assemblies were possible for the genomes of just three of these organisms, and nearly half of the sequences obtained could not be assembled beyond mate-pairing of bidirectional reads from individual clones (*7*).

An important unanswered question arising from these initial forays into shotgun sequencing of environmental samples is whether whole genome shotgun sequence data can be meaningful in the absence of significant assembly. Many microbial communities, such as those residing in soils and sediments, are substantially more complex than those found in the acid mine drainage biofilm or even the Sargasso Sea and thus less amenable to genome assembly (*6, 7, 9*). This obstacle may in part be offset by the high gene density of prokaryotes and improved sequence read lengths which result in most reads containing a significant portion of at least one gene (*10, 11*). While genomic sequencing goals targeting species ranging from humans to microbes have almost exclusively focused on determining the complete genome of a particular organism, the genetic information contained in individual sequencing reads suggests the feasibility of a less whole genome-oriented approach to the analysis of metagenomic data. Examination of the proteins encoded in a community, rather than the types of organisms producing them, could potentially distinguish samples based on the functions selected for by the local environment and reveal insights into features of that environment.

To explore the utility of a gene-centric approach to environmental samples we examined a number of communities of varying complexity from environments presenting several different metabolic challenges. One sample was from soil, a

nutrient-rich environment harboring microbial communities of high genetic and biochemical diversity whose complexity poses a daunting challenge for genomic analysis (*12, 13*). 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 (*14*). This unique ecological niche, referred to as a "whale fall," has been suggested to select for "specialist" species in geographically remote locations (*15*). Bone samples from two widely separated communities in the Pacific and Antarctic Oceans were examined, as well as a microbial mat sample from the Pacific site. We explored the gene contents of the partially assembled and unassembled reads from soil and whale fall samples and used them to compare these communities to each other and to those previously studied from the Sargasso sea and an acid mine drainage biofilm (*6, 7*).

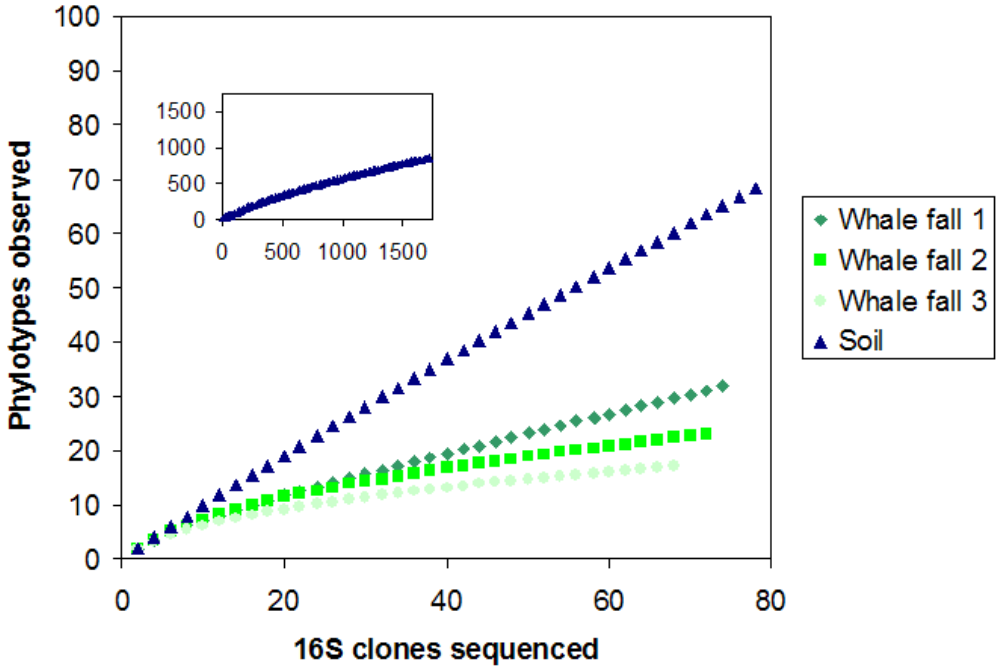
Preliminary analysis of the microbial diversity in the soil and whale fall communities was performed with PCR amplified small ribosomal RNA libraries generated for each sample using primers specific for the three domains of life (Bacteria, Archaea, and Eukaryotes). In the soil sample, rRNA gene sequences revealed the presence of a wide diversity of bacteria but very few archaeal species; some fungi and unicellular eukaryotes were also found (Supplemental Figure S1). To obtain a more accurate assessment of the bacterial species in soil and their phylogenetic distribution, we sequenced 1700 clones from two independent libraries of PCR-amplified bacterial 16S rRNA sequences prepared from the soil DNA. We observed 847 distinct ribotypes among these sequences,

distributed among Proteobacteria, Chloroflexi and Acidobacteria as well as 15 other phyla (Supplemental Figure S1B). Extrapolations of the accumulation curve (Figure 1)(*16*) predicted the total number of bacterial ribotypes in this sample to be more than 3000, consistent with previous estimates of soil biodiversity (*9*). The most common ribotype accounts for 112 (6.6%) of the clones, supporting the conjecture that soil communities typically lack a "dominant" member as a consequence of spatial isolation and high nutrient availability (Supplemental Figure S1D) (*17*). These numbers, based on the relatively loose cutoff of 97% or greater sequence identity (*18*), likely underestimate the number of species by an order of magnitude, as even microbes with very high similarity in 16S rRNA sequence are known to have quite divergent genome content (*19, 20*).

The whale fall samples each contained a more limited spectrum of bacteria than soil, and those from the Pacific site also contained a handful of archaea and eukaryotes (Supplement). Roughly 85 bacterial 16S clones were sequenced from each site, revealing between 17 and 37 unique sequences in each, primarily drawn from the Bacteroidetes and Proteobacteria phyla for all three libraries (Supplemental Figure S2A). Only one clone was common to all three libraries, an alpha proteobacterium most closely related to an Antarctic lake clone, but the closest relatives for virtually all of the sequences were from marine environments including hydrothermal vents, a milieu whale falls are thought to mimic in some ways (*14*). The accumulation curves suggest these communities are both less diverse and less evenly distributed than the soil cohort; each is estimated to

contain between 25 and 150 distinct ribotypes of which the most abundant accounts for 15-25% of the library (*21*) (Figure 1 and Supplemental Figure S2). 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 ecological niche.

Figure 1: **Species complexity.** Accumulation curves of bacterial 16S rRNA clone sequences for soil and whale fall samples. Inset: Accumulation 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.



We complemented the ribosomal library sequences with the sequencing of genomic small-insert libraries made from the soil and all three whale skeleton samples. In light of the organismal complexity seen in the soil sample, we generated 100 Mb of sequence from this sample and 25 Mb for each whale fall library. Partial small ribosomal RNA genes found in these data largely reflected the species and phyla found in the PCR clone libraries, demonstrating that the PCR clone data present a reasonable picture of the communities being examined (Supplement). 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 either the 16S rRNA data or the overlaps in the genomic sequence, we project that somewhere between two and five Gbp of sequence would be necessary to obtain the 8X coverage traditionally targeted for draft genome assemblies, even for the single most predominant genome in this complex community (Supplement). While the whale fall libraries also lacked contigs greater than a few kilobases, 34-47% of the reads in each library overlapped with reads from independent clones. Based on the 16S rRNA and genomic assembly data, we estimate that between 100 and 700 Mb of shotgun sequence data would be needed from each sample in order to generate a draft assembly for the most prevalent genome. Assembling genomes for low-abundance community members in either soil or whale falls would clearly require significantly more sequence data.

Given the significant hurdles to the assembly of complete genomes from these samples, we chose to study the environment from the perspective of the genes present without attempting to place them in the context of an individual genome. While such an analysis does not resolve relationships among different organisms, it may, in part, provide some insight into the properties of the community as a whole. We chose to use the spectrum and abundance of genes present in the metagenomic data to compare and contrast the communities found in soil and each of the three whale falls. To explore the utility of this "functional fingerprinting," we also extended our comparison to data from both the acid mine drainage (AMD) study (*6*) and three samples taken from different locations in the Sargasso Sea with matched prefilter and collection filter sizes (*7*).

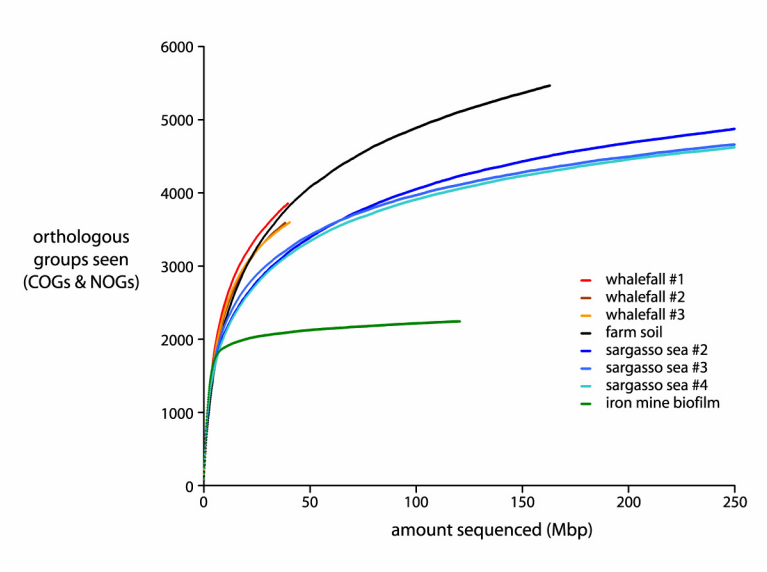
Since much of the environmental sequence data is in the form of single reads, we examined automated annotations for the five genomes assembled from an AMD biofilm microbial community (*6*) and compared them to annotations of the complete set of unassembled reads from this sample. This analysis revealed that 95% of the functionally categorized predicted genes in these genomes were apparent in the unassembled data (Supplement). With this result supporting the validity of gene predictions on unassembled reads, we then applied the annotation process to the data from soil, the three whale fall communities, and the three Sargasso Sea samples, including all unassembled reads in addition to any available contigs. As our analysis relied primarily on the predicted genes on small DNA fragments, we termed each environmental sequence an Environmental Gene Tag (EGT), to distinguish them from the sequencing reads primarily used

for the assembly of genomes. At least 90% of the sequence fragments from each sample were shown to contain putative genes, an efficiency resulting from the gene density of prokaryotes and the >700 bp length of each sequence. More than a third contained two or more predicted ORFs, raising the possibility of nearest-neighbor analysis.

All predicted genes were compared to Clusters of Orthologous Groups (COGs) (*22*). To increase the coverage and include more low-abundance orthologous groups that might discriminate among environments, we expanded the COG set from 4873 to 20334 by applying the STRING orthology assignment protocol to 179 completely sequenced genomes (*23, 24*). Roughly half the predicted proteins in each sample exhibited homology to orthologous groups; the 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. To test whether the orthologous groups from 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 AMD 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 (Figure 2). 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 whale fall and Sargasso Sea communities, consistent with their species complexity, fell between acid mine drainage and soil. We observed qualitatively similar curves when limiting the analysis to the original 4873 COGs (Supplemental Figure S3A), showing that despite incomplete gene classification one can assess functional complexity with fragmented sequence data. Since the use of orthologous groups as a reference has specific biases that may influence our results, we also examined the protein repertoire in these sequences utilizing the domain-oriented Pfam database and its native search tool, HMMer (*25*). The number of Pfam domains observed also demonstrated a leveling off with increasing sequence (Supplemental Figure S3B).

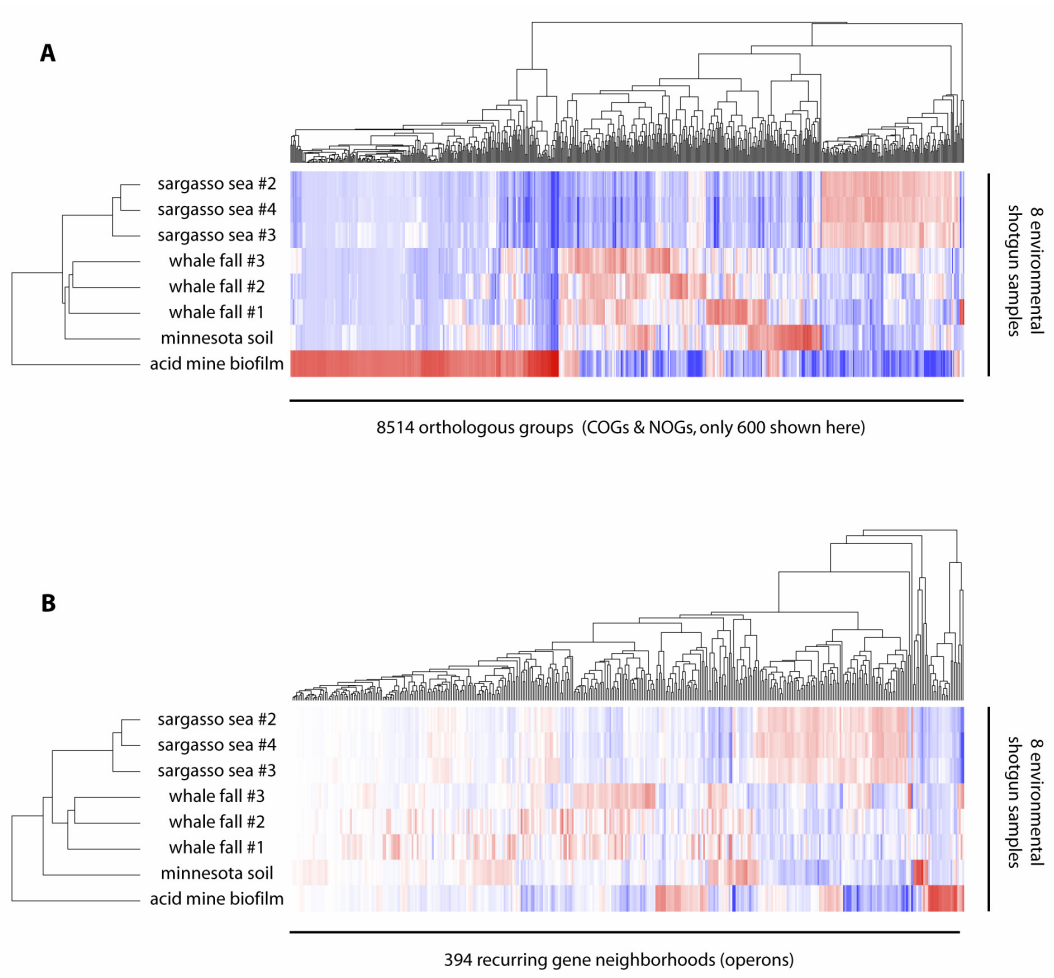
Figure 2: **Identification of orthologous groups with greater sequencing depth.** The number of new orthologous groups predicted in the sequence data from each library is shown as a function of the raw sequence generated. (Raw sequence numbers were used because quality information was not available for all samples; for JGI data raw read lengths are typically 64% greater than quality base counts determined with a Phred score 15 threshold.)

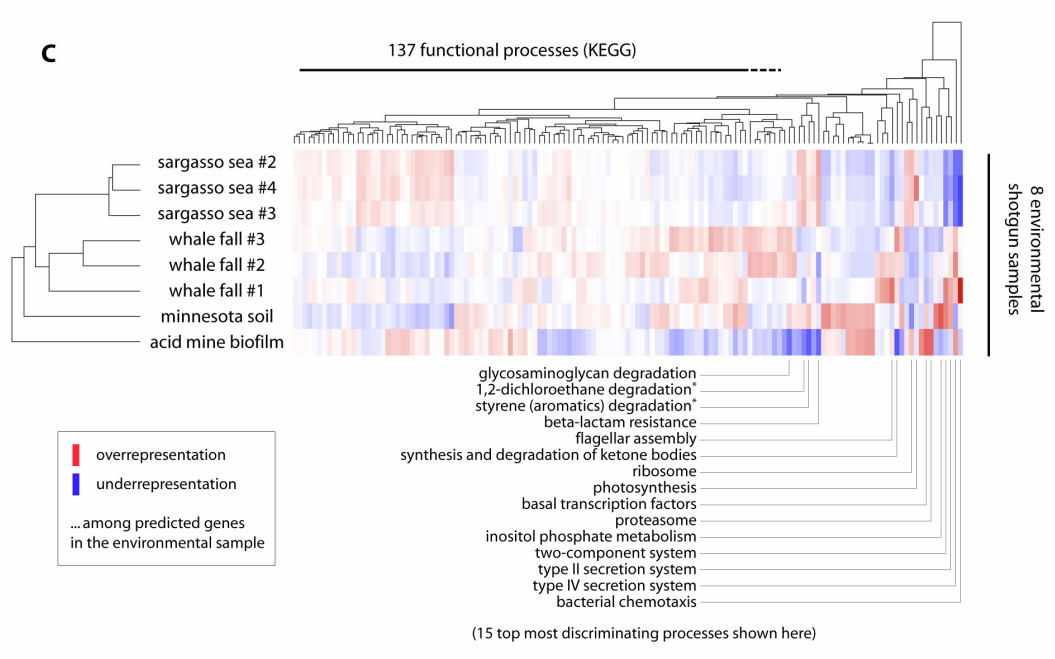


What may be more relevant than the total number of different orthologous groups in a sample is the relative proportion of the total protein sets devoted to particular functions, which could aid in revealing the predominant metabolic interactions of these microbes with their environments. We therefore explored whether quantitative variations in the relative distribution of protein sequences among communities reflected the specific demands placed on the organisms by their local environment. Our prediction here was that samples from different environments would have different profiles, while independent samples from similar environments would exhibit similar functional profiles. The sample characterization exploited functional binning at three levels: i) individual genes (orthologous groups), ii) conserved "operons" that usually encode individual pathways, and iii) higher order cellular processes that combine a number of such individual pathways.

Functional relationships among all individual genes (orthologous groups) in the expanded COG database were inferred from their genomic neighborhoods in sequenced genomes. Groups were then clustered into "operons" containing on average 4-5 genes. These operons have been shown to correlate well with known metabolic pathways (*23, 26*). Proteins were also grouped into higher order cellular processes, containing on average 15 genes, according to the manually curated KEGG database (*27*). Predicted genes in the environmental samples were assigned to orthologous groups by sequence comparison and weighted according to their clone depth. Profiles generated by binning the genes into orthologous groups, operons and cellular processes were then used to compare and cluster the samples. The results of these analyses are displayed in Figures 3A-3C, each of which is a two-way clustering of samples and functional bins in which over- and under-represented categories are indicated by red and blue blocks respectively. Regardless of the functional binning employed, trees with roughly similar topology resulted. A similar tree was also obtained when only 10 Mb of unassembled sequence from each community was included (Supplement). Each sample had a unique profile, but, importantly, the independent Sargasso Sea samples clustered together in all of the trees, as did the whale fall samples. Thus the predicted protein complement of a community is similar to that of other communities whose environments of origin pose similar metabolic challenges. These results support the hypothesis that the "functional" profile of a community is influenced by its environment and that EGT data can serve as a fingerprint for particular environments.

Figure 3: **Functional profiling of microbial communities.** Two-way clustering of samples and encoded genes / functions based on relative enrichment of A) COGs, B) operons and C) KEGG functional processes. In A) and B), only those categories with significant inter-sample variation are shown; all KEGG maps are shown in C with the 15 most discriminating processes highlighted. Maps marked with an asterisk are xenobiotic degradation pathways whose apparent presence in the samples is likely indicative of more general degradative processes.





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 at four different levels of resolution; apart from the previously introduced COG, operon and KEGG map levels, we also included 23 broad functional categories representing on average 170 COGs each into this analysis (Figure 4)(*22*). 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 revealed environment-specific variations in both biochemistry and phylogeny.

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 soil but not a single one in the Sargasso Sea. On the other hand, 466 distinct copies of the light-driven proton pump bacteriorhodopsin are found in the surface water, but not a single one in the deep sea or in soil. Strikingly, there are several uncharacterized genes displaying equally extreme distributions across samples, e.g. 444 members of COG4338 in Sargasso Sea but none in the soil or whale fall data.

The analysis of operons 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 Na+ exporters coupled to oxidative phosphorylation. Soil, on the other hand, has a strong enrichment in operons responsible for active K+-channeling including KDP-type channels, which are thought to use K+ to regulate intracellular osmolarity and turgor. 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 (Supplement). In terms of available electron acceptors,

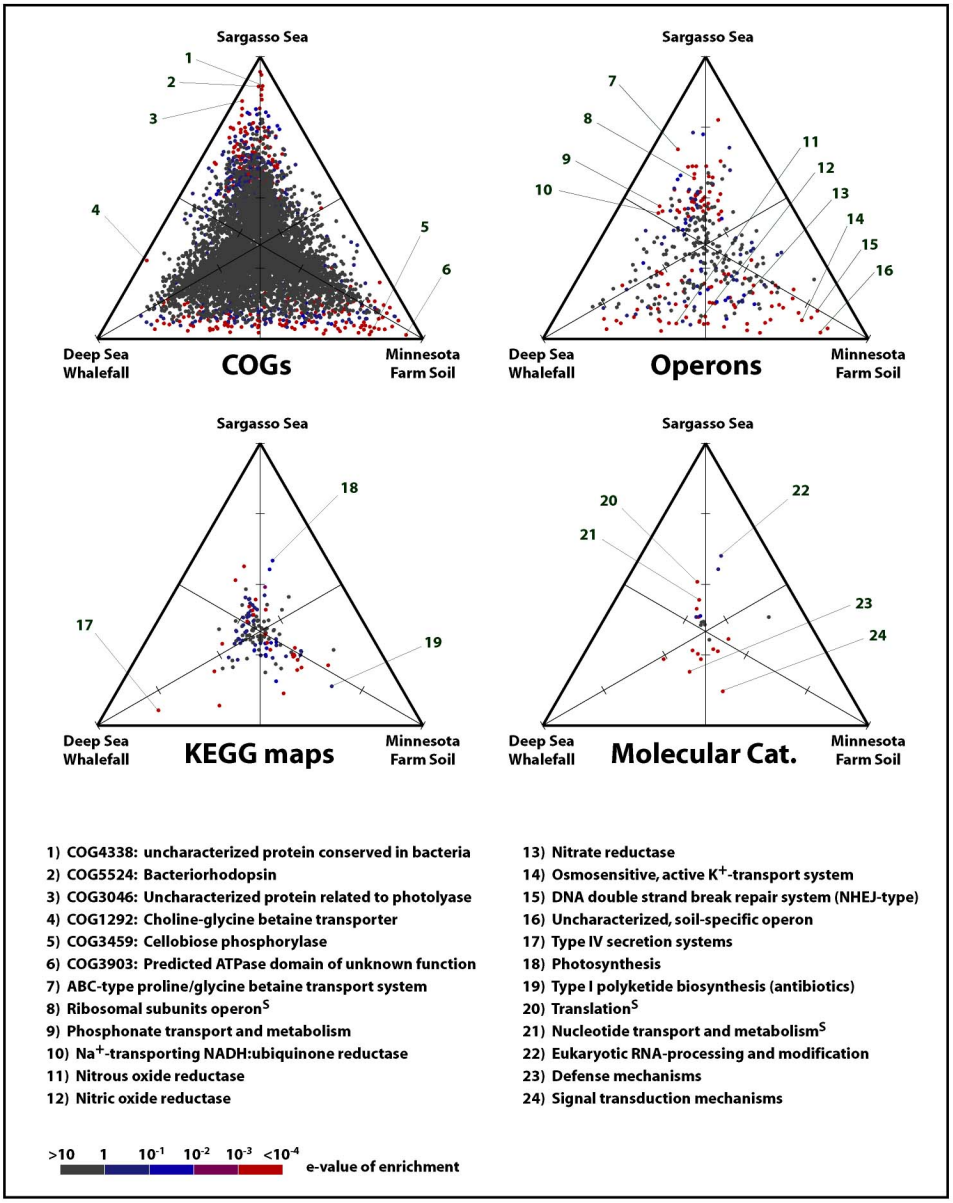
however, the deep sea whale falls share much in common with soil, including an enrichment of all three types of nitrate respiration processes (i.e. subunits of nitrous oxide reductase, nitrite oxide reductase, and nitrate oxide reductase).

Apart from relatively obvious adaptations to the environment, we also observe community differences that are unexpected at first glance. For example, we 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)(*28*). This suggests that soil microbes may have 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.

Examination of the higher order processes reveals known differences in energy production (e.g. photosynthesis in the oligotrophic waters of the Sargasso Sea) (*29*) or population density and interspecies communication (overrepresentation of conjugation systems, plasmids, and antibiotic biosynthesis in soil; Figure 4, lower left) (*12*). The broad functional COG categories, on the other hand, primarily suggest differences in genome size and phylogenetic composition. 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)(*30, 31*). The greater prevalence of RNA processing genes in the Sargasso Sea is indicative of a significant eukaryotic component in these samples (*7*).

Notably, many uncharacterized genes and processes are among the most significant overrepresentations in each sample. This hints at an abundance of hitherto unknown functional systems, specific to each environment; they represent a rich source for further, more directed experimental and computational investigations (*32*). Taken together, the 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 interactions.

Figure 4: **Specific Enrichments**. Three-way comparisons of soil, whale fall and Sargasso Sea environments, in terms of single genes (i.e. orthologous groups, COGs) or functional groupings of genes (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 while items with unbiased occurrences are positioned in the middle. Color indicates statistical significance of the enrichment, determined by comparison to randomized data. Items marked with [S] are indicative of differences in average genome size between the environments (see text).



These analyses demonstrate that whole genome assemblies of individual organisms from environmental communities are not required for an informed characterization of the biochemical potential of microbes present in a spectrum of ecological niches. Rather, we were able to show that the predicted metaproteome,

obtained from fragmented sequence data, identified functional fingerprints which could be used to discriminate environments. The reflection of known biochemical properties of these environments in the fingerprints suggests that this information may be predictive of environmental features of interest such as nutrient supply or 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. A. H. Treusch *et al.*, *Environ Microbiol* **6**, 970 (Sep, 2004).
4. M. R. Liles, B. F. Manske, S. B. Bintrim, J. Handelsman, R. M. Goodman, *Appl Environ Microbiol* **69**, 2684 (May, 2003).
5. O. Beja *et al.*, *Science* **289**, 1902 (Sep 15, 2000).
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. T. P. Curtis, W. T. Sloan, J. W. Scannell, *Proc Natl Acad Sci U S A* **99**, 10494 (Aug 6, 2002).

9. V. Torsvik, L. Ovreas, T. F. Thingstad, *Science* **296**, 1064 (May 10, 2002).

10. Y. A. Goo *et al.*, *BMC Genomics* **5**, 3 (Jan 12, 2004).

11. C. Jenkins, V. Kedar, J. A. Fuerst, *Genome Biol* **3**, RESEARCH0031 (2002).

12. R. Daniel, *Curr Opin Biotechnol* **15**, 199 (Jun, 2004).

13. V. Torsvik, J. Goksoyr, F. L. Daae, *Appl Environ Microbiol* **56**, 782 (Mar, 1990).

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

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

16. J. B. Hughes, J. J. Hellmann, T. H. Ricketts, B. J. Bohannan, *Appl Environ Microbiol* **67**, 4399 (Oct, 2001).

17. J. Zhou *et al.*, *Appl Environ Microbiol* **68**, 326 (Jan, 2002).

18. S. G. Acinas *et al.*, *Nature* **430**, 551 (Jul 29, 2004).

19. O. Beja *et al.*, *Appl Environ Microbiol* **68**, 335 (Jan, 2002).

20. G. Sabehi, O. Beja, M. T. Suzuki, C. M. Preston, E. F. DeLong, *Environ Microbiol* **6**, 903 (Sep, 2004).

21. T. D. Olszewski, *OIKOS* **104**, 377 (February, 2004).

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

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

24. C. von Mering, P. Bork, *Nucleic Acids Res* **Database issue, in press** (Jan 1, 2005).

25. A. Bateman *et al.*, *Nucleic Acids Res* **32 Database issue**, D138 (Jan 1, 2004).

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

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

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

29. J. Wu, W. Sunda, E. A. Boyle, D. M. Karl, *Science* **289**, 759 (Aug 4, 2000).

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

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

32. K. A. Gray, T. H. Richardson, D. E. Robertson, P. E. Swanson, M. V. Subramanian, *Adv Appl Microbiol* **52**, 1 (2003).

33. 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. 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; Natalia Ivanova, Nikos Kyrpides, Tanja Woyke and members of the Rubin lab for helpful comments on the manuscript; and Paul Richardson, Jarrod Chapman, Alex Copeland, Igor Grigoriev, Ernest Szeto, Jan Korbel, Tobias Doerks, Konrad Foerstner, Eoghan Harrington and Marcus Krupp for assistance with data processing and analysis.

Supporting Online Material

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 no evidence of eukaryotic cells in the sample and no normalization or fractionation methods were applied to enrich for a particular component of the community.

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 (*1*). "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 (*2*). 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 Express vector (Stratagene). The library was amplified once then *in vivo* excised to form a pBK-CMV 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 Express vector, then in vivo excised, without amplification, to form a pBK-CMV 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; 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 (*3*) 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 single-linkage clustering to group together any sequences with >97% identity over the full length of the insert. Accumulation 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. 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 (*4-6*), 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 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 twice 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 were analyzed by the program FGENESB from Softberry, which predicts genes and operons as well as functional RNAs (described at http://www.softberry.com). Functional annotation of the predicted proteins utilized an extended version of the COGs 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 COGs database (*7*). The extension has been performed using an unsupervised procedure in the context of the STRING project (*8*). 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 COGs 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 COGs 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 COGs 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.

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 (*9*). 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 3A of the main text 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 (*10*), 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 (*11*) 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 (*11*), 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). 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. 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 between 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.

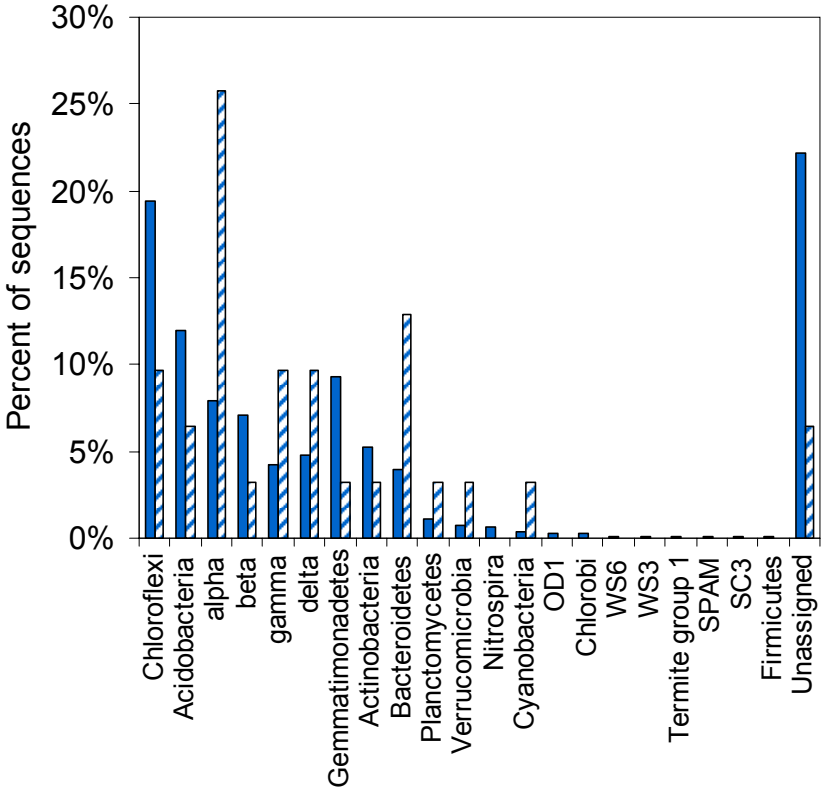
First, the actual items were binned into abundance classes. An observed relative enrichment is statistically more significant for an abundant item (e.g. a widespread orthologous group or a large operon) than for a rare item. Comparison of items to randomized data was done separately for each abundance 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.

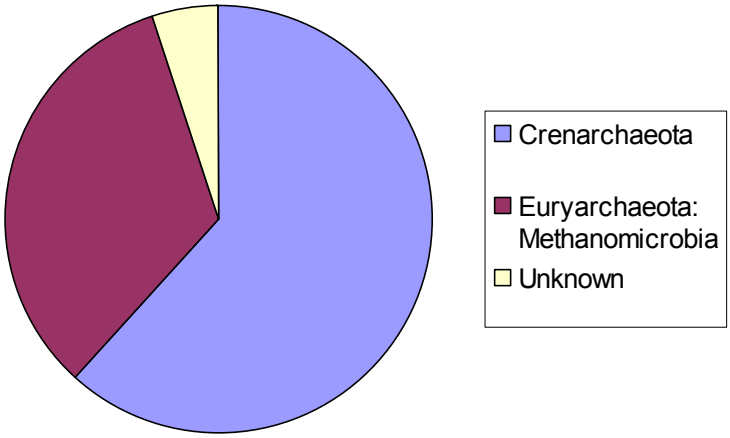
Supplemental Data

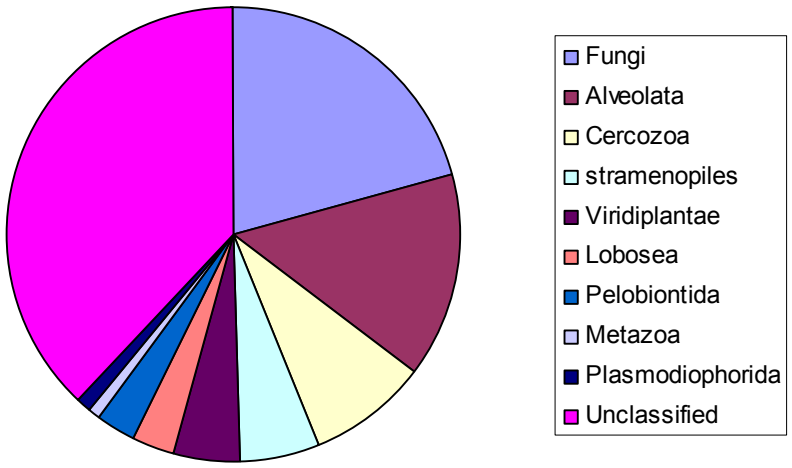
16S / 18S ribosomal RNA analyses

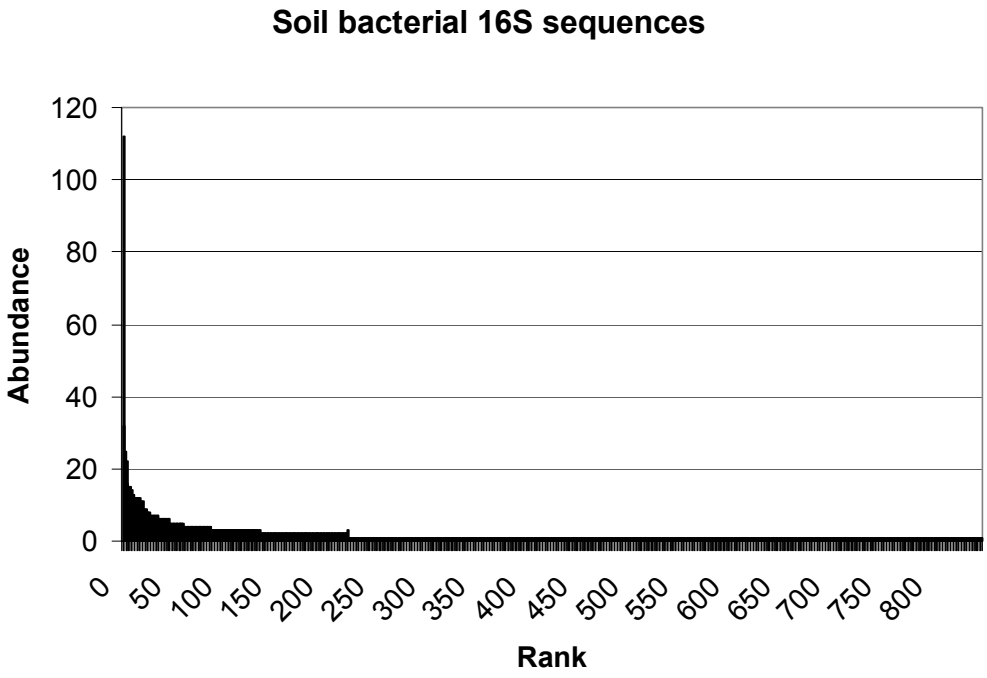
The bacterial 16S rRNA sequences from soil (1700 total) clustered into 847 unique groups mapping to 18 different phyla. Most of these sequences were singlets, and the largest cluster contained just 112 clones, or 6.6% of the total (Supplementary Figure 1A). The 58 archaeal clones formed just seven clusters, all within two major euryarchaeal branches (Supplementary Figure 1B), 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.

Supplemental Figure S1: 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.


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


Crenarchaeota
Euryarchaeota: Methanomicrobia
Unknown



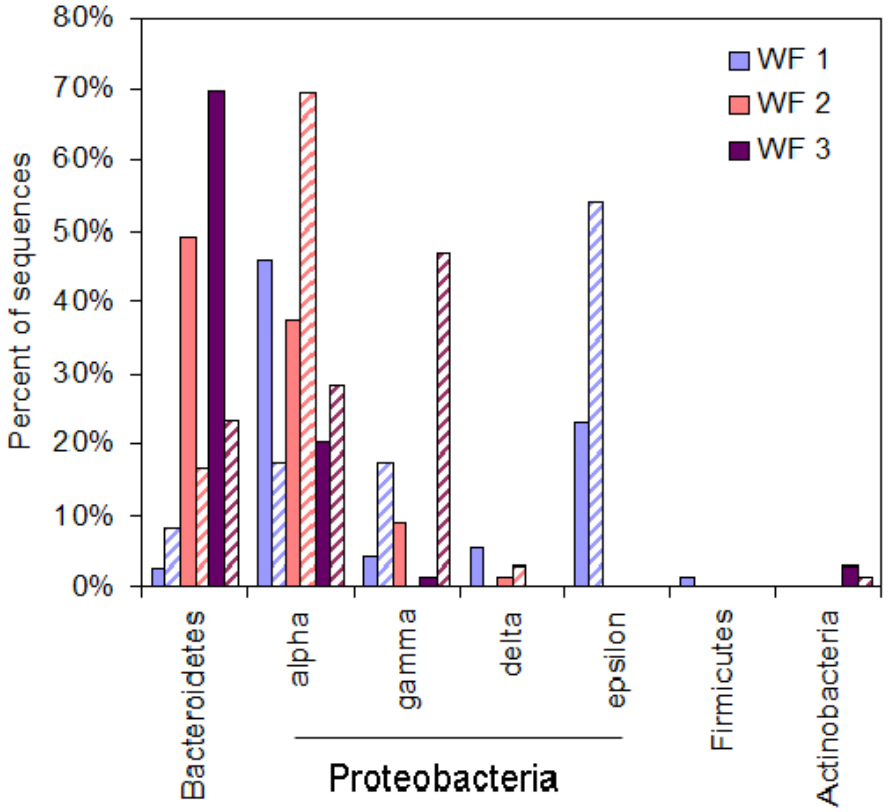


Each whale fall bacterial 16S library contained 17-37 unique sequences mapping primarily to the Proteobacteria and Bacteroidetes. In contrast to soil, more than half of the sequences were distributed among the top few clusters (Supplementary

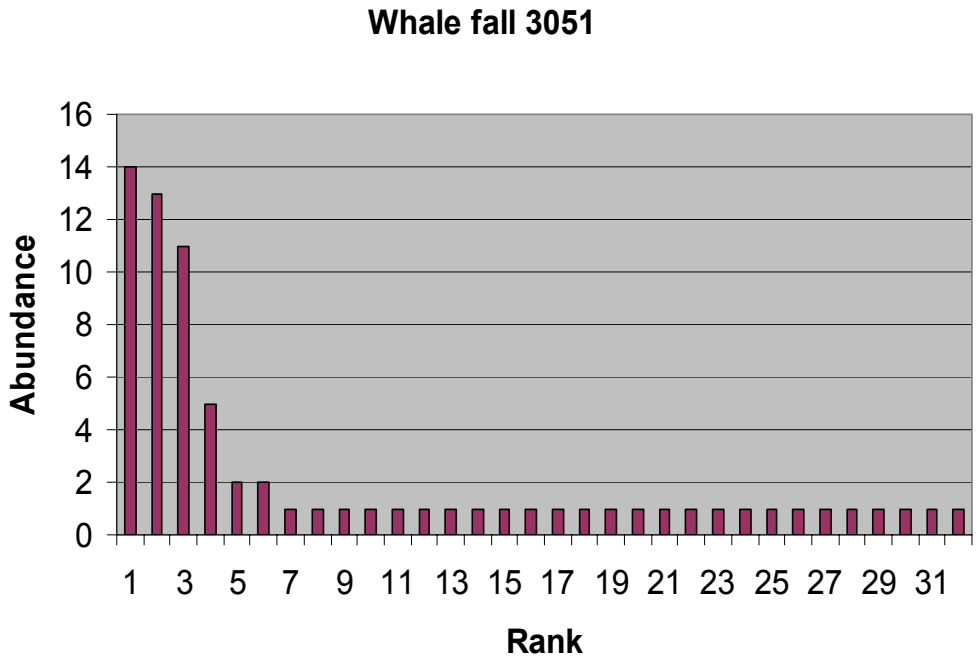
Figure 2A). The archaeal 16S sequences from the two Pacific samples fell into a limited number of clusters, primarily within the Methanomicrobia and, for the mat sample, the C1 archaea. The eukaryotic 18S sequences from the mat sample were all from the same deeply branching eukaryote while those from 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).

Supplemental Figure S2: Rank-abundance curves for whale fall bacterial 16S sequences. A) Whale fall 1, Santa Cruz bone; B) Whale fall 2, Santa Cruz microbial mat; C) Whale fall 3, Antarctic bone. D) Assignment of 16S rRNA sequences to bacterial phyla for both PCR clone libraries (solid bars) and genomic libraries (hatched bars).

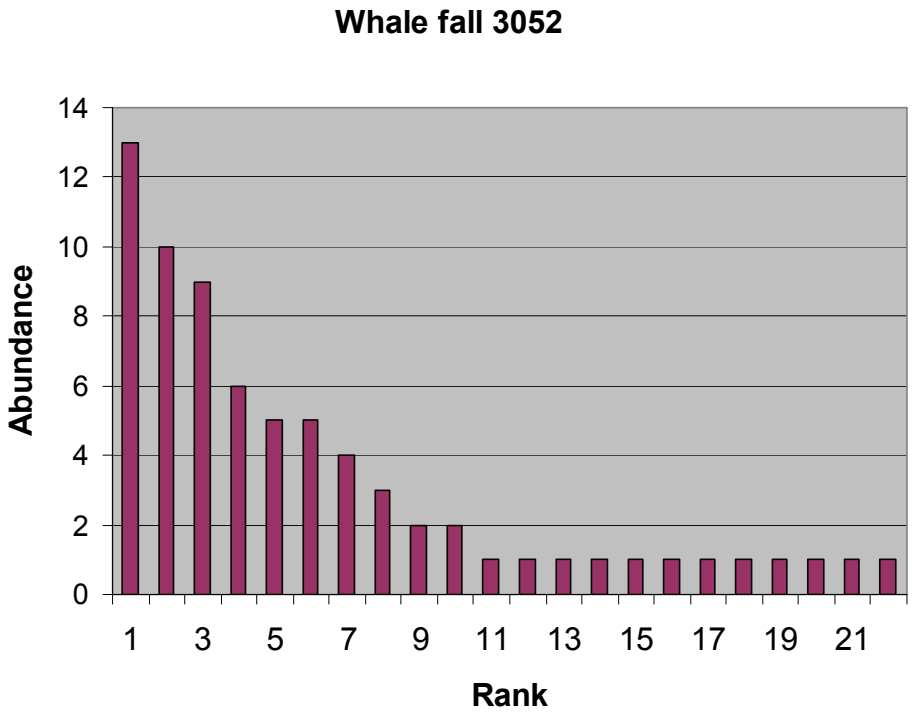
A


80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent of sequences
WF 1
WF 2
WF 3
Bacteroidetes
alpha
gamma
delta
epsilon
Firmicutes
Actinobacteria
Proteobacteria

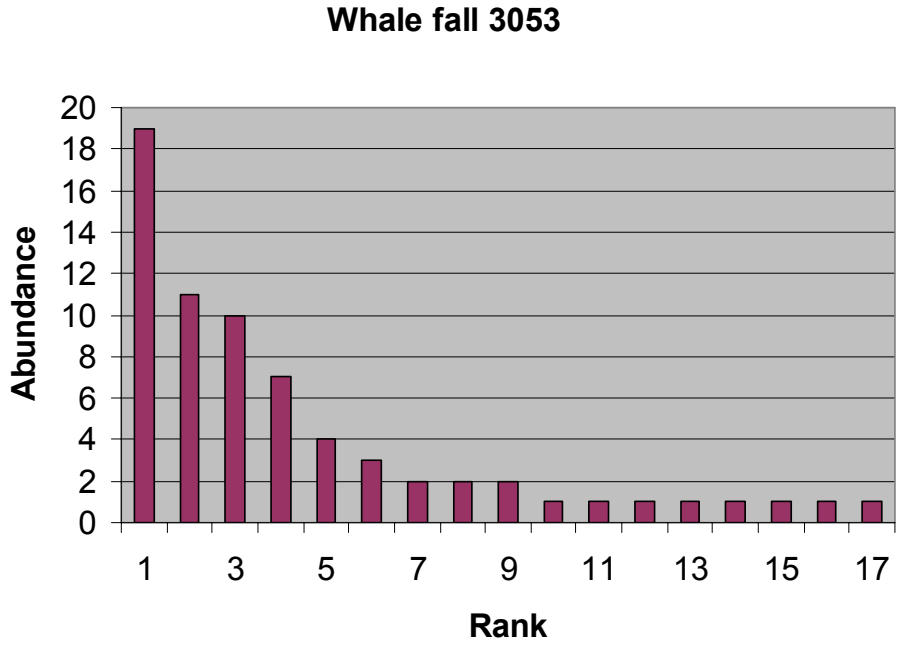
B)


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)

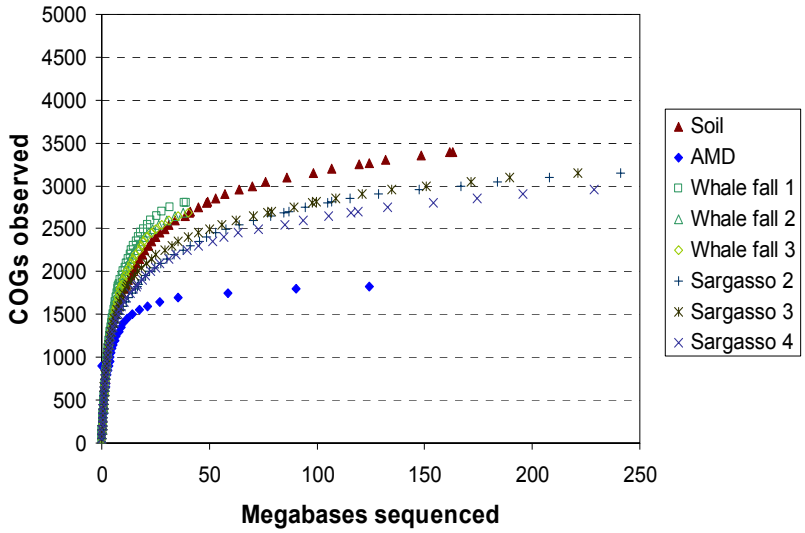

Whale fall 3053
Abundance
20
18
16
14
12
10
8
6
4
2
0
1
3
5
7
9
11
13
15
17
Rank

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

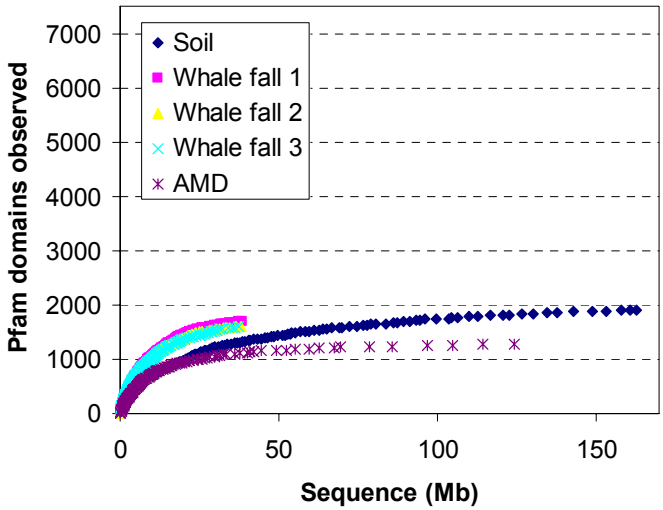
Automated annotation was applied to the assembled genomic scaffolds from the acid mine drainage biofilm as well as to all unassembled reads from the same sample. In the five genome "bins" assembled from the acid mine drainage sequence data, a total of 7173 distinct proteins were predicted in 1629 different 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.

Supplemental Figure S3: Functional accumulation curves for all samples. Number of unique hits in the A) COG and B) Pfam database as a functional of sequence depth. The y-axis maximum is set to the total number of categories in each database.

A)



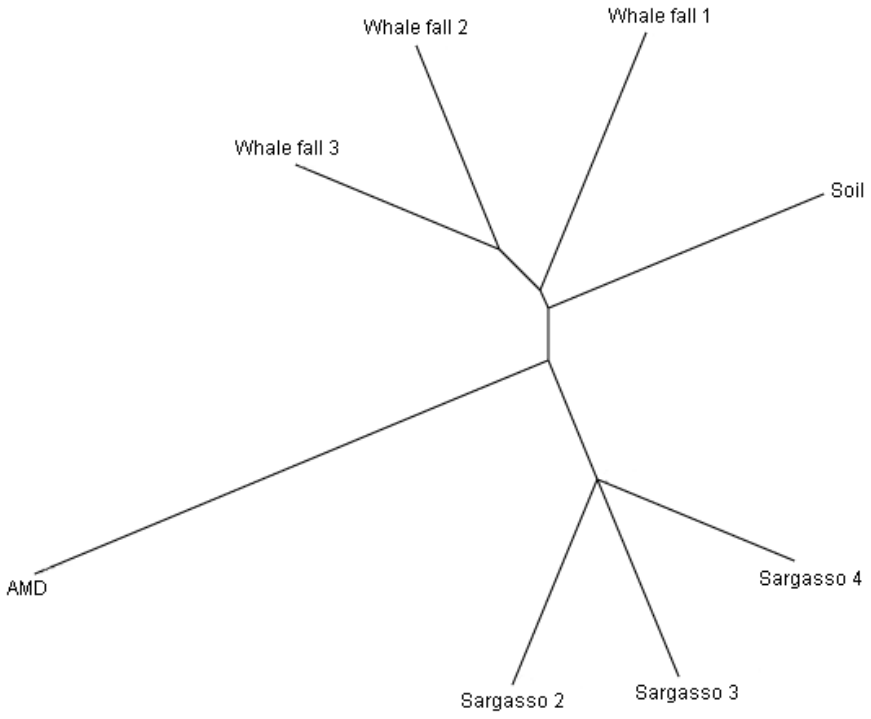
B)



Supplemental Figure S4: 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).



1. N. J. Debenham, P. J. D. Lambshead, T. J. Ferrero, C. R. Smith, *Deep Sea Research Part I: Oceanographic Research Papers* **51**, 701 (2004/5, 2004).
2. E. M. Gabor, E. J. de Vries, D. B. Janssen, *FEMS Microbiology Ecology* **44**, 153 (May 15, 2003).
3. T. Huber, G. Faulkner, P. Hugenholtz, *Bioinformatics* (Apr 8, 2004).
4. T. Kaneko *et al.*, *DNA Res* **7**, 381 (Dec 31, 2000).

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

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

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

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

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

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

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