5. J. H. Lee *et al.*, *Mol. Cancer Res.* **1**, 674 (2003).
6. R. Kitagawa, C. J. Bakkenist, P. J. McKinnon, M. B. Kastan, *Genes Dev.* **18**, 1423 (2004).
7. P. T. Yazdi *et al.*, *Genes Dev.* **16**, 571 (2002).
8. J. H. Lee, T. T. Paull, *Science* **304**, 93 (2004).
9. C. J. Bakkenist, M. B. Kastan, *Nature* **421**, 499 (2003).
10. Materials and methods are available as supporting material on *Science* Online.
11. J. H. Lee, T. T. Paull, data not shown.
12. T. T. Paull, M. Gellert, *Mol. Cell* **1**, 969 (1998).
13. J.-H. Lee *et al.*, *J. Biol. Chem.* **278**, 45171 (2003).
14. T. T. Paull, M. Gellert, *Genes Dev.* **13**, 1276 (1999).
15. G. Moncalian *et al.*, *J. Mol. Biol.* **335**, 937 (2004).
16. R. Shroff *et al.*, *Curr. Biol.* **14**, 1703 (2004).
17. M. Lisby, J. H. Barlow, R. C. Burgess, R. Rothstein, *Cell* **118**, 699 (2004).
18. D. Nakada, K. Matsumoto, K. Sugimoto, *Genes Dev.* **17**, 1957 (2003).
19. A. Ali *et al.*, *Genes Dev.* **18**, 249 (2004).
20. A. A. Goodarzi *et al.*, *EMBO J.* **23**, 4451 (2004).
21. Molecular interaction data have been deposited in the Biomolecular Interaction Network Database (BIND) with accession codes 216020 to 216045. We thank M. Kastan and R. Abraham for expression constructs; D. Ramsden, M. Gellert, and M. O'Dea for Rag1/Rag2 protein; S. Stevens for technical advice; members of the Paull lab for their help; and R. Rothstein for a helpful word. This work was supported by NIH (grant CA094008) and by the American Cancer Society (grant RSG-04-173-01-CCG).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1108297/DC1
Materials and Methods
Figs. S1 and S2
References

6 December 2004; accepted 24 February 2005
Published online 24 March 2005;
10.1126/science.1108297
Include this information when citing this paper.

Comparative Metagenomics of Microbial Communities

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]†

The species complexity of microbial communities and challenges in culturing representative isolates make it difficult to obtain assembled genomes. Here we characterize and compare the metabolic capabilities of terrestrial and marine microbial communities using largely unassembled sequence data obtained by shotgun sequencing DNA isolated from the various environments. Quantitative gene content analysis reveals 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 16*S* 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. The latter efforts include the sequencing of individual large-insert bacterial artificial chromosome (BAC) clones as well as small-insert libraries made directly from environmental DNA (*3*–*7*). The application of high-throughput shotgun sequencing environmental samples has recently provided global views of those communities not obtainable from 16*S* rRNA or BAC clone–sequencing surveys (*6*, *7*). The sequence data have also posed challenges to genome assembly, which suggests 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 per 1000 base pairs (bp)] and currently attainable read lengths (700 to 750 bp), which result in most individual sequences containing a significant portion of at

[1]Department of Energy (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.

*S.G.T. and C.v.M. contributed equally to this work.
†To whom correspondence should be addressed. E-mail: emrubin@lbl.gov

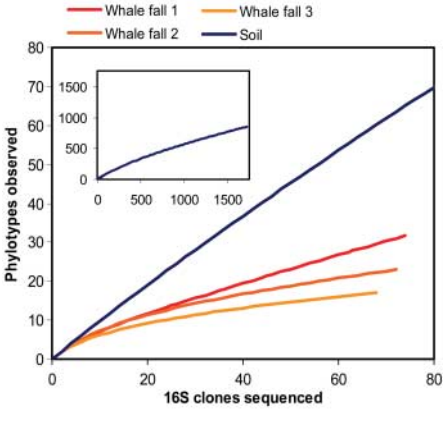

Fig. 1. Species complexity. Rarefaction curves of bacterial 16*S* rRNA clone sequences for soil and whale fall samples. (**Inset**) Rarefaction curve for all 1700 soil clones. The three whale falls are: 1, Santa Cruz Basin bone; 2, Santa Cruz Basin microbial mat; and 3, Antarctic bone.

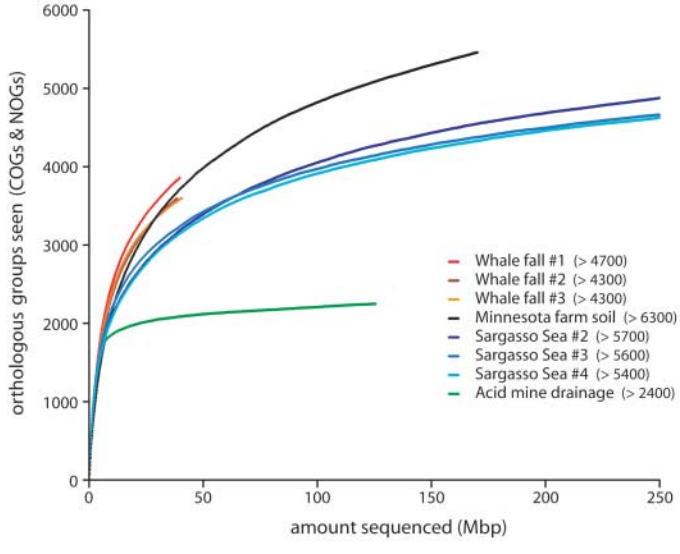

Fig. 2. Identification of orthologous groups with greater sequencing depth. The number of orthologous groups observed at least once is shown as a function of the raw sequence generated. Numbers in parentheses indicate lower limits of the total number of groups in the sample.

least one gene (*9*). Accordingly, although 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 on the basis of the functions selected for by the local environment and reveals insights into features of that environment. In the present study, 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 and 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 rRNA libraries generated for each sample by using primers specific for Bacteria, Archaea, and Eukaryota. In the soil sample, a wide diversity of bacteria, few archaeal species, and some fungi and unicellular eukaryotes were found (fig. S2). We sequenced a total of 1700 clones from two independent libraries of PCR-amplified bacterial 16*S* rRNA sequences prepared from the soil DNA, and we identified at least 847 distinct ribotypes from more than a dozen phyla (fig. 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 (Fig. 1; fig. S1), which reflected the enormous diversity found in soil (*8*). The most common ribotype accounts for 112 (6.6%) of the clones (fig. S2A) 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 to 25% of the library (Fig. 1; fig. S4). The reduced species and phyla diversity of the whale fall microbial communities as compared with 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 bp (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. On the basis of our 16*S* 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 Mbp 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 sequence with unassembled, 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 EGTs from the sequencing reads primarily used for the assembly of genomes. The gene contents of the partially

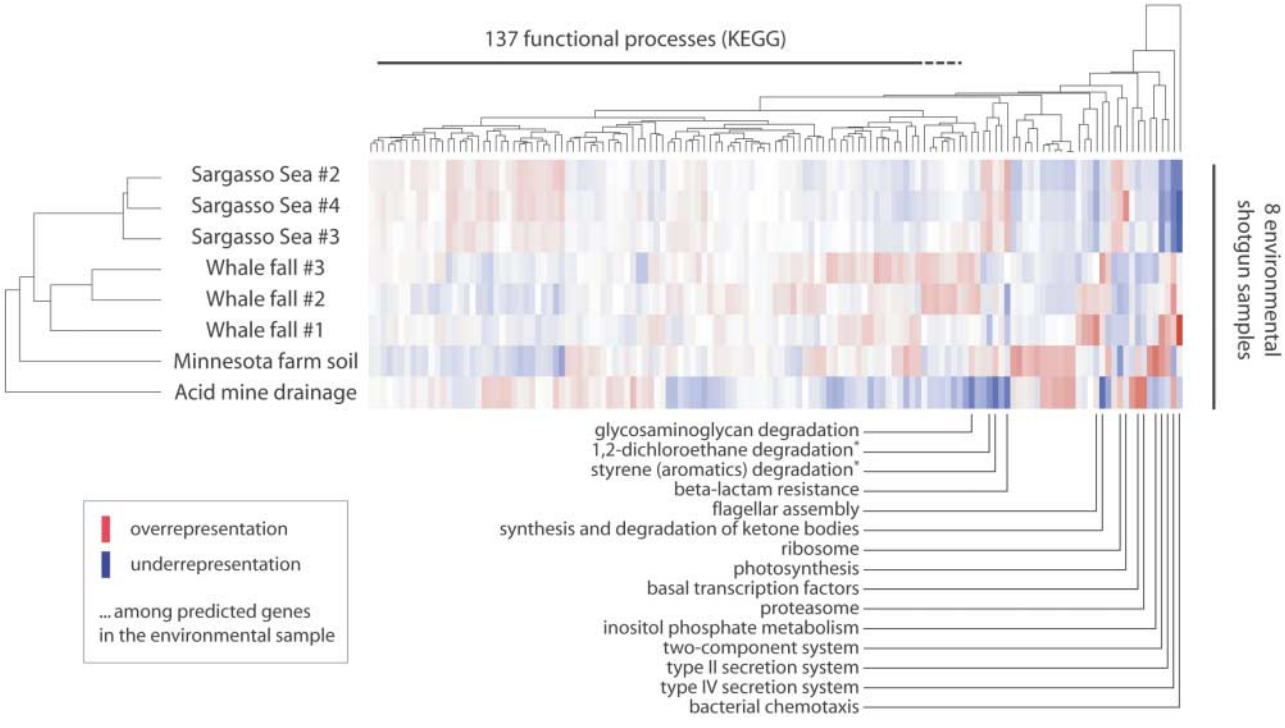

Fig. 3. Functional profiling of microbial communities. Two-way clustering of samples and encoded functions based on relative enrichment of KEGG functional processes. The 15 most discriminating processes are highlighted. Asterisks indicate that environmental genes mapping to these processes probably have a broader range of substrates than the KEGG process title indicates.

assembled and unassembled reads from soil and whale fall samples were compared with each other and with those of an acid mine drainage biofilm community (*6*) and with 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, which raised the possibility of nearest-neighbor analysis (*14*).

Roughly half of the predicted proteins in each sample showed homology to orthologous groups in an expanded in-house COG (clusters of orthologous groups of proteins) 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, whereas the general slope of the curve reveals information about community diversity (Fig. 2). In the relatively simple acid mine drainage biofilm community, 90% of the orthologous groups were detected with just 25 Mbp of 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, which suggests 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*) (fig. S5), which suggests that this phenomenon is not an artifact of comparison to a particular database.

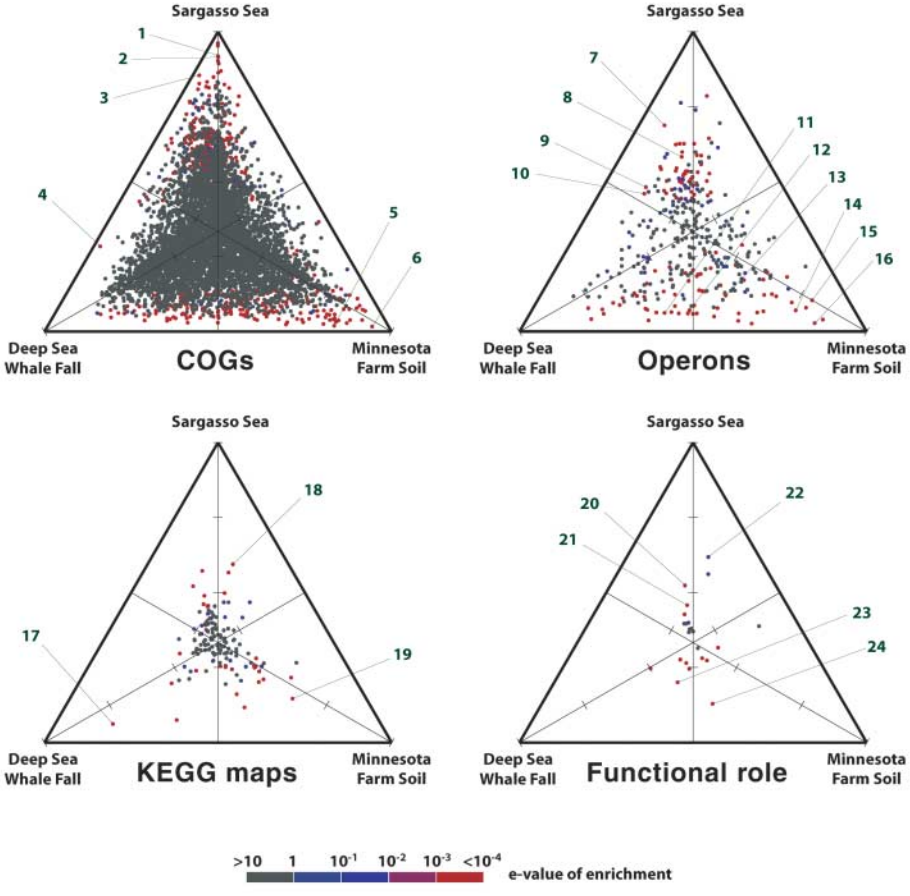

Fig. 4. Specific enrichments. Three-way comparisons of soil, whale fall, and Sargasso Sea environments in terms of COGs, operons, KEGG processes, and COG functional categories. Each dot shows the relative abundance of an item in the three environmental samples, such that proximity to a vertex is proportional 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; and 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 (*23*).

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 also library representation (*19*), of genes coding for specialized enzymes can vary with sample source. We specifically explored whether independent samples from similar, although 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 and other pathways (*20*)]; third, higher order cellular processes from the manually curated KEGG (Kyoto Encyclopedia of Genes and Genomes) 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 underrepresented categories are indicated by red and blue blocks, respectively, is displayed in Fig. 3 (fig. S6 shows 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 (Fig. 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 great dissimilarity from the other samples (Fig. 3; fig. 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 (Fig. 4, upper left). For example, 73 putative orthologs of cellobiose phosphorylase, involved in degradation of plant material, are found in the ~100 Mbp of soil sequence, but none are found in the ~700 Mbp of sequence examined from the Sargasso Sea. On the other

hand, 466 distinct homologs of the light-driven proton pump bacteriorhodopsin are found in the surface waters of the Sargasso Sea, whereas none are found in the deep-sea whale falls or in soil.

The analysis of operons likewise reveals similarities and differences in functional systems (Fig. 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 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: Whereas 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 and starch and sucrose metabolism in soil) (*7*) or population density and interspecies communication [overrepresentation of conjugation systems, plasmids, and antibiotic biosynthesis in soil (Fig. 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 "life-style" 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. At the same time, 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 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 and Notes

1. E. F. DeLong, N. R. Pace, *Syst. Biol.* **50**, 470 (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 (2003).
4. O. Beja *et al.*, *Science* **289**, 1902 (2000).
5. A. H. Treusch *et al.*, *Appl. Environ. Microbiol.* **6**, 970 (2004).
6. G. W. Tyson *et al.*, *Nature* **428**, 37 (2004).
7. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
8. V. Torsvik, L. Ovreas, T. F. Thingstad, *Science* **296**, 1064 (2002).
9. Y. A. Goo *et al.*, *BMC Genomics* **5**, 3 (2004).
10. C. R. Smith, A. R. Baco, in *Oceanography and Marine Biology: An Annual Review*, R. N. Gibson, R. J. A. Atkinson, Eds. (Taylor & Francis, London, 2003), vol. 41, pp. 311–354.
11. J. B. Hughes, J. J. Hellmann, T. H. Ricketts, B. J. Bohannan, *Appl. Environ. Microbiol.* **67**, 4399 (2001).
12. R. K. Colwell, personal communications (1994–2004).
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 (1999).
15. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
16. C. von Mering *et al.*, *Nucleic Acids Res.* **33**, D433 (2005).
17. A. Bateman *et al.*, *Nucleic Acids Res.* **32** (Database special issue), D138 (2004).
18. S. K. Rhee *et al.*, *Appl. Environ. Microbiol.* **70**, 4303 (2004).
19. D. E. Robertson *et al.*, *Appl. Environ. Microbiol.* **70**, 2429 (2004).
20. C. von Mering *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15428 (2003).
21. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, *Nucleic Acids Res.* **32** (Database special issue), D277 (2004).
22. R. Daniel, *Curr. Opin. Biotechnol.* **15**, 199 (2004).
23. http://string.embl.de/metagenome_comp_suppl/

24. This work was performed under the auspices of the DOE's Office of Science, Biological and Environmental Research Program; 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. S.G.T. was supported by grant no. THL007279F, an NIH National Research Service Award (NRSA) Training and Fellowship grant to E.R. K.C. was supported by NSF grant no. EF 03-31494. Sequencing of the environmental libraries was performed under a license agreement with Diversa (J. R. Short, U.S. patent no. 6455254). We gratefully acknowledge the efforts of C. Baptista, L. Christoffersen, J. Garcia, K. Li, J. Ritter, P. Sammon, S. Wells, D. Whitney, J. Eads, T. Richardson, M. Noordewier, and L. Bibbs. We thank C. Smith for providing the whale fall samples; K. Remington for providing Sargasso Sea sample information; N. Ivanova, N. Kyrpides, and members of the Rubin laboratory for helpful comments on the manuscript; and J. Chapman, I. Grigoriev, E. Szeto, J. Korbel, T. Doerks, K. Foerstner, E. Harrington, and M. Krupp for assistance with data processing and analysis. These Whole Genome Shotgun projects have been deposited with the DNA Data Bank of Japan, the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, and the GenBank in collaboration (DDBJ/EMBL/GenBank) under the project accessions AAFX00000000 (soil), AAFY00000000 (whale fall 1), AAFZ00000000 (whale fall 2), and AAGA00000000 (whale fall 3). For each project, the version described in this paper is the first version, AAFX01000000, AAFY01000000, AAFZ01000000, and AAGA01000000. The 16*S* rRNA sequences from the soil and three whale fall samples have been deposited under GenBank accession nos. AY921654 to AY922252. The metagenomic data will also be incorporated into the U.S. Department of Energy Joint Genome Institute Integrated Microbial Genomes system (www.jgi.doe.gov/) to facilitate detailed comparative analysis of the data in the context of all publicly available complete microbial genomes.

Supporting Online Material
www.sciencemag.org/cgi/content/full/308/5721/554/DC1
Materials and Methods
Figs. S1 to S7
References and Notes

23 November 2004; accepted 4 February 2005
10.1126/science.1107851

A Cellular MicroRNA Mediates Antiviral Defense in Human Cells

Charles-Henri Lecellier,[1]* Patrice Dunoyer,[1] Khalil Arar,[2] Jacqueline Lehmann-Che,[3] Stephanie Eyquem,[4] Christophe Himber,[1] Ali Saïb,[3] Olivier Voinnet[1]*

In eukaryotes, 21- to 24-nucleotide-long RNAs engage in sequence-specific interactions that inhibit gene expression by RNA silencing. This process has regulatory roles involving microRNAs and, in plants and insects, it also forms the basis of a defense mechanism directed by small interfering RNAs that derive from replicative or integrated viral genomes. We show that a cellular microRNA effectively restricts the accumulation of the retrovirus primate foamy virus type 1 (PFV-1) in human cells. PFV-1 also encodes a protein, Tas, that suppresses microRNA-directed functions in mammalian cells and displays cross-kingdom antisilencing activities. Therefore, through fortuitous recognition of foreign nucleic acids, cellular microRNAs have direct antiviral effects in addition to their regulatory functions.

In plants and insects, viral double-stranded RNA is processed into small interfering RNAs (siRNAs) by the ribonuclease (RNase) III enzyme Dicer. These siRNAs are incorporated into the RNA-induced silencing complex to target the pathogen's genome for destruction (*1*, *2*). Plant and insect viruses can counter this defense with silencing suppres-