LETTERS

Functional metagenomic profiling of nine biomes

Elizabeth A. Dinsdale[1,5]*, Robert A. Edwards[1,2,3,6]*, Dana Hall[1], Florent Angly[1,4], Mya Breitbart[7], Jennifer M. Brulc[8], Mike Furlan[1], Christelle Desnues[1]†, Matthew Haynes[1], Linlin Li[1], Lauren McDaniel[7], Mary Ann Moran[10], Karen E. Nelson[11], Christina Nilsson[12], Robert Olson[6], John Paul[7], Beltran Rodriguez Brito[1,4], Yijun Ruan[12], Brandon K. Swan[13], Rick Stevens[6], David L. Valentine[13], Rebecca Vega Thurber[1], Linda Wegley[1], Bryan A. White[8,9] & Forest Rohwer[1,2]

Microbial activities shape the biogeochemistry of the planet[1,2] and macroorganism health[3]. Determining the metabolic processes performed by microbes is important both for understanding and for manipulating ecosystems (for example, disruption of key processes that lead to disease, conservation of environmental services, and so on). Describing microbial function is hampered by the inability to culture most microbes and by high levels of genomic plasticity. Metagenomic approaches analyse microbial communities to determine the metabolic processes that are important for growth and survival in any given environment. Here we conduct a metagenomic comparison of almost 15 million sequences from 45 distinct microbiomes and, for the first time, 42 distinct viromes and show that there are strongly discriminatory metabolic profiles across environments. Most of the functional diversity was maintained in all of the communities, but the relative occurrence of metabolisms varied, and the differences between metagenomes predicted the biogeochemical conditions of each environment. The magnitude of the microbial metabolic capabilities encoded by the viromes was extensive, suggesting that they serve as a repository for storing and sharing genes among their microbial hosts and influence global evolutionary and metabolic processes.

Genomic plasticity of microbes causes variations in the gene content of closely related strains[4], making predictions of community metabolism on the basis of representative genomes and signature genes such as 16S ribosomal RNA unreliable. Although it seems that core genomes are relatively stable and shared among most individuals of the same species, parts of the genome (for example, prophages, CRISPRs, pathogenicity/ecological islands, ORFans) are hyper-variable[5]. Together, these two components make up the pangenome[4]. Unlike the signature genes approach, metagenomic approaches analyse the complete genetic information of microbial and viral communities[6,7]. In this way, the relative abundances of all genes can be determined and used to generate a description of the functional potential of each community[8–14].

Here we use a comparative metagenomic approach to statistically analyse the frequency distribution of 14,585,213 microbial and viral metagenomic sequences to elucidate the functional potential of nine biomes including: subterranean (that is, mine samples); hypersaline ponds from solar salterns; marine; freshwater; coral-associated; microbialites (including stromatolites and thrombolites); aquaculture-fish-associated; terrestrial-animal-associated; and mosquito-associated (details in Supplementary Table 1 and Supplementary Fig. 1). Microbial and viral metagenomes (Supplementary Fig. 2 and Supplementary Table 2) were isolated and pyrosequenced. The sequences were compared to the 2007 SEED platform (http://www.theseed.org) using the BLASTX algorithm, and hits with an *E*-value of <0.001 were considered to be significant (Methods). A total of 1,040,665 sequences from the 45 microbial metagenomes and 541,979 sequences from the 42 viral metagenomes were significantly similar to functional genes within the SEED (Supplementary Table 1). The SEED arranges metabolic pathways into a hierarchical structure in which all of the genes required for a specific task are arranged into subsystems[15]. At the highest level of organization, the subsystems include both catabolic and anabolic functions (for example, DNA metabolism) and at the lowest levels the subsystems are specific pathways (for example, the synthesis pathway for thymidine).

Table 1 shows the relative abundances of sequences assigned to each major subsystem in the combined analysis of the microbiomes

Table 1 | Mean percentage of sequences (± s.e.m.) similar to major metabolisms

Metabolic category	Microbial metagenomes	Viral metagenomes
Carbohydrates	17.218 (± 0.648)	14.353 (± 0.718)
Amino acids	12.036 (± 0.491)	10.132 (± 0.642)
Virulence	9.788 (± 0.339)	11.175 (± 0.508)
Protein metabolism	9.123 (± 0.497)	8.838 (± 0.522)
Respiration	7.139 (± 1.285)	3.718 (± 0.276)
Photosynthesis	6.965 (± 2.148)	1.984 (± 0.554)
Cofactors, vitamins, and so on	5.411 (± 0.226)	6.661 (± 0.393)
RNA metabolism	3.971 (± 0.195)	4.324 (± 0.387)
DNA metabolism	3.970 (± 0.180)	7.555 (± 0.943)
Nucleosides and nucleotides	3.316 (± 0.149)	7.666 (± 0.817)
Cell wall and capsule	3.235 (± 0.223)	5.098 (± 0.649)
Fatty acids and lipids	3.095 (± 0.160)	3.002 (± 0.242)
Membrane transport	2.736 (± 0.158)	2.371 (± 0.182)
Stress response	2.599 (± 0.115)	3.354 (± 0.326)
Aromatic compounds	2.351 (± 0.175)	2.550 (± 0.340)
Cell division and cell cycle	1.791 (± 0.091)	1.983 (± 0.212)
Nitrogen metabolism	1.547 (± 0.070)	1.135 (± 0.093)
Sulphur metabolism	1.230 (± 0.102)	1.302 (± 0.134)
Motility and chemotaxis	1.022 (± 0.096)	1.011 (± 0.083)
Phosphorus metabolism	0.909 (± 0.080)	1.319 (± 0.167)
Cell signalling	0.885 (± 0.076)	0.885 (± 0.072)
Potassium metabolism	0.796 (± 0.048)	0.846 (± 0.079)
Secondary metabolism	0.159 (± 0.014)	0.235 (± 0.047)

[1]Department of Biology, [2]Center for Microbial Sciences, [3]Department of Computer Sciences, and [4]Computational Science Research Centre, San Diego State University, San Diego, California 92182, USA. [5]School of Biological Sciences, Flinders University, Adelaide, South Australia 5042, Australia. [6]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [7]University of South Florida, College of Marine Science, 140 7th Avenue South, St Petersburg, Florida 33701, USA. [8]Department of Animal Sciences, and [9]The Institute for Genomic Biology, University of Illinois, Urbana, Illinois 61801, USA. [10]Department of Marine Sciences, University of Georgia, Athens, 30602 Georgia, USA. [11]The J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. [12]Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [13]Department of Earth Science, University of California Santa Barbara, Santa Barbara, California 93106, USA. †Present address: Unité des Rickettsies, CNRS-UMR 6020, Faculté de médecine, 13385 Marseille, France.
*These authors contributed equally to this work.

Table 2 | Mean functional diversity and evenness (± s.e.m.) of metagenomes, sampled from nine environments

	Functional diversity (H')		Functional evenness	
Biome	Microbial	Viral	Microbial	Viral
Subterranean	2.393 (± 0.030)		0.005 ($\pm 1.2\times10^{-4}$)	
Hypersaline	2.361 (± 0.006)	2.041 (± 0.021)	0.005 ($\pm 1.4\times10^{-4}$)	0.012 ($\pm 5.6\times10^{-4}$)
Marine	2.313 (± 0.021)	2.162 (± 0.026)	0.005 ($\pm 0.9\times10^{-4}$)	0.007 ($\pm 4.0\times10^{-4}$)
Freshwater	2.430 (± 0.003)	2.080 (± 0.034)	0.005 ($\pm 0.9\times10^{-4}$)	0.010 ($\pm 6.7\times10^{-4}$)
Coral	1.733 (± 0.059)	2.289 (± 0.023)	0.009 ($\pm 5.2\times10^{-4}$)	0.007 ($\pm 1.1\times10^{-4}$)
Microbialites	2.408 (± 0.015)	1.743 (± 0.115)	0.005 ($\pm 3.8\times10^{-4}$)	0.019 ($\pm 6.9\times10^{-3}$)
Fish	2.447 (± 0.001)	2.439 ($\pm 3.1\times10^{-4}$)	0.005 ($\pm 0.4\times10^{-4}$)	0.005 ($\pm 0.7\times10^{-4}$)
Terrestrial animals	2.428 (± 0.006)	2.016 (± 0.173)	0.004 ($\pm 0.1\times10^{-4}$)	0.017 ($\pm 4.5\times10^{-3}$)
Mosquito		2.395 (± 0.015)		0.004 ($\pm 0.5\times10^{-4}$)

There are no subterranean viral metagenomes and no mosquito microbial metagenomes.

compared with the viromes. Over 30% of the identifiable genes in the microbiomes were associated with carbohydrate or protein metabolism. Respiration and photosynthesis subsystems accounted for an additional ~15% of the similarities. Subsystems responsible for nucleic acid metabolism and virulence were overrepresented in the viral fractions (Table 1), whereas respiration and photosynthesis genes were less frequent.

The functional diversity represented by the metagenomes approached its theoretical limit of 2.81 in all environments (Table 2), showing that most subsystems were represented in all of the samples. Only the coral-associated microbes showed a lower functional diversity; this is because they have fewer secondary metabolisms, virulence pathways, cell signalling pathways and membrane transport pathways. Because microbes associated with corals are taxonomically diverse[11], functional reduction may have occurred in these communities, similar to microbes in other symbiotic relationships[16].

Diversity is a function of both richness (that is, the number of metabolic processes) and evenness (that is, the relative abundance of a particular metabolic process in a sample). The evenness for the metagenomes was very low (<0.1; Table 2 and Supplementary Fig. 3), showing that there are a few dominant metabolisms in each environment. Differential dominant metabolisms suggest that there are characteristic functional profiles of the metagenomes.

To test the hypothesis that each environment has a distinguishing metabolic profile, a canonical discriminant analysis (CDA) was conducted (Fig. 1). Most of the variance between the different environments (79.8% of the combined microbiome and 69.9% of the virome) was explained in this analysis, showing that metagenomes are highly predictive of metabolic potential within an ecosystem. In contrast, a recent analysis of 16S rRNA genes from multiple environments only explained about 10% of the variance[17], suggesting that different ecosystems cannot be distinguished by their taxa.

The position of each metagenome in Fig. 1 reflects the frequency combination of sequences associated with each subsystem; the vectors indicate which metabolisms most strongly determined the distribution. Using these as clues, it is possible to determine which metabolisms are important for the organisms in that environment relative to other environments. For example, subsystems involved in respiration and protein metabolism placed the coral-associated microbes apart from the microbes found within terrestrial animals. This trend is visualized in Fig. 2, which shows that ~20% of the coral-associated microbial genes were involved in respiration, compared with only 3% in the microbiomes associated with terrestrial animals. The relatively high occurrence of respiration-associated genes in the coral-associated microbiomes reflects the diurnally fluctuating oxygen environment, which is supersaturated with oxygen in the day and essentially anaerobic at night[18]. In contrast, microbes living within the stable anaerobic alimentary tracts of terrestrial animals are less likely to experience selection for multiple respiration pathways.

Similarly, virulence genes were proportionally more abundant in the organism-associated microbes than in free-living microbes. These are the factors necessary to facilitate symbiotic relationships (mutualism, parasitism or commensalisms; Fig. 2f–h). Another example of the predictive power of the metagenomes is the sulphur metabolisms associated with aquaculture fish. In particular, two subsystems—alkanesulphonate and taurine metabolism—were overrepresented in fish-associated metagenomes (Supplementary Fig. 4). Alkanesulphonates are involved in the use of both inorganic and organic sulphur, such as taurine and aliphatic sulphonates[19] (taurine is a sulphur organic acid used to supplement aquaculture fish food[20]).

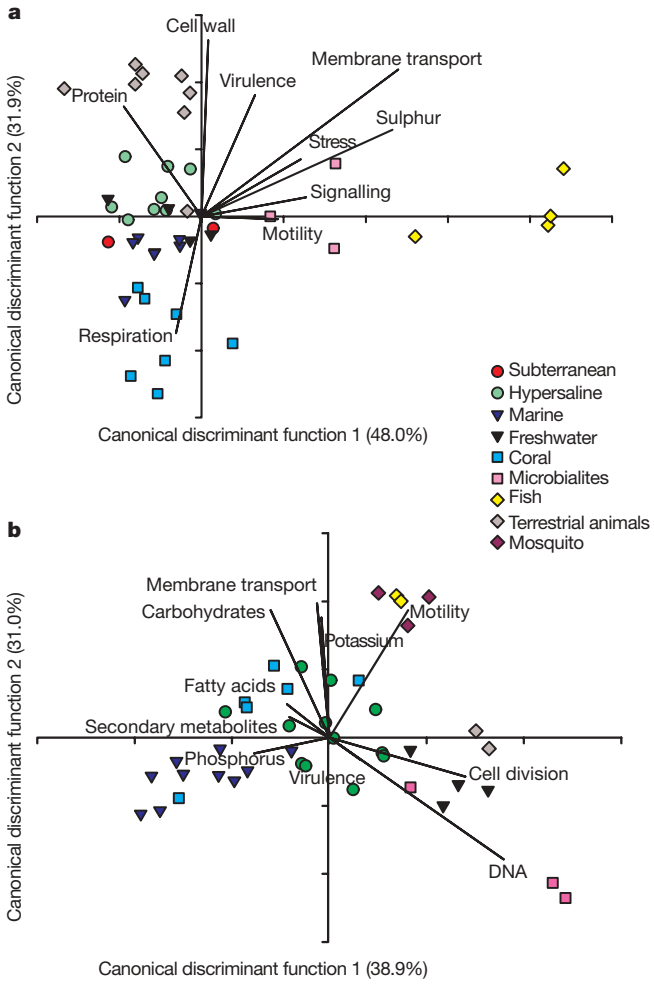

Figure 1 | Functional analysis of microbial and viral metagenomes. The CDA of the microbial (**a**) and viral (**b**) metagenomes identified that the metabolic processes grouped these communities in the two-dimensional spaced described by canonical discriminant functions 1 and 2. The symbols represent the position of each metagenome and the vectors represent the structural matrix for subsystems that were identified as influencing the separation of the metagenomes using the stepwise procedure. The length of the vectors represents the strength of influence of the particular metabolic process. The cross-validation scores for the microbial and viral metagenomes were 66.7 and 59.9%, respectively.

Together, these examples show that metagenomes predict important, emergent biological characters of the environments. By substituting environmental groups in multiple CDAs, the predictive nature of metagenomes was confirmed (Supplementary Figs 5 and 6).

Shifting of a metagenome from its sister group in the CDA was also predictive of ecological differences. For example, one of the marine metagenomes (number 27 Supplementary Table 1) was positioned more negatively than the rest of the marine metagenomes (Fig. 1a). This sample was taken from waters that were unusually rich in nitrogen, phosphate and dissolved organic carbon[21]. The ability to determine subtle differences in metabolic potential will allow the detection of environmental changes at early stages of perturbation and identify previously unknown pathways for therapeutics.

The viromes are dominated by phage, which are expected to have similar lifestyles in every environment (infection, replication, host lysis and release of free virions). Phage have also been shown to move between environments[22], which suggests that their metabolic profiles are similar in different ecosystems. In contrast, other studies have shown that phage carry 'specialization' genes[23], including phosphate metabolism[24] and cyanobacterial photosystems[25], to manipulate host metabolisms associated with a particular ecosystem. Phage 'sample' their host's genetic material and incorporate extra pieces of DNA called MORONS[26], suggesting that phage metagenomes may instead show distinctive profiles based on their environment. As shown in Figs 1b and 2, the viromes have highly predictive metabolic profiles that suggest enrichment for specific genes in different environments, and thus support the latter hypothesis (69.9% of the variance).

Because phages and viruses are non-motile, the abundance of motility and chemotaxis proteins within the combined viral metagenomes was the most unexpected example of specialized metabolisms being carried within the viromes (Fig. 3). A total of 130 SEED-annotated motility and chemotaxis proteins (out of a possible 157) were present in the viromes. There was a non-random acquisition of these proteins by the viral community, shown by the variation in relative abundances of these proteins between the microbial and viral metagenomes (Supplementary Table 3). In the viromes, flagellar biosynthesis protein FlhA, the chemotaxis response regulator proteins CheA and CheB and deacylases were overrepresented (Supplementary Table 3), whereas the twitching motility protein PilT, type II secretary pathways and GldJ were overrepresented in the microbiomes. *cheA* and *cheB* genes within microbes work together to control flagella motor switching rates[27], but their role within the phage remains an outstanding question.

Essentially all of the functional diversity was represented in the viromes. Unlike their cellular hosts, most viruses must carry a specific amount of DNA to correctly pack their capsids (that is, viruses are not evolutionarily penalized for carrying 'extra' DNA). If there is a selective advantage of the extra DNA (resulting in increased phage progeny), these genes are fixed in the phage genome; otherwise they will be lost. Because there are an estimated 10^{31} phages on the planet and they can move between environments, the potential reservoir of genes that can be transferred both locally and globally[12] by phage is enormous[28]. As our research shows, there is little restriction to the types of genes carried by the viral community, suggesting that they influence a wide range of processes, including biogeochemical cycling, short-term adaptation and long-term evolution of microbes.

The low functional evenness measured for both microbial and viral metagenomes is even lower than the functional diversity calculated for individual bacterial genomes (Table 2 and Supplementary Fig. 3). This finding is diametrically opposed to the high taxonomic evenness reported for both microbial and viral communities[2,12], ranging from 0.6 to 1 for human faecal and marine viruses[9,12] and about 0.9 for soil microbes[29]. To resolve this apparent dilemma, we propose that the frequency of a gene encoding a particular metabolic function reflects its relative importance in an environment, and that genetic sweeps favour particular gene frequencies regardless of their taxonomical background. That is, rather than changing taxa, variation in gene content, presumably by means of horizontal gene transfer[30] between

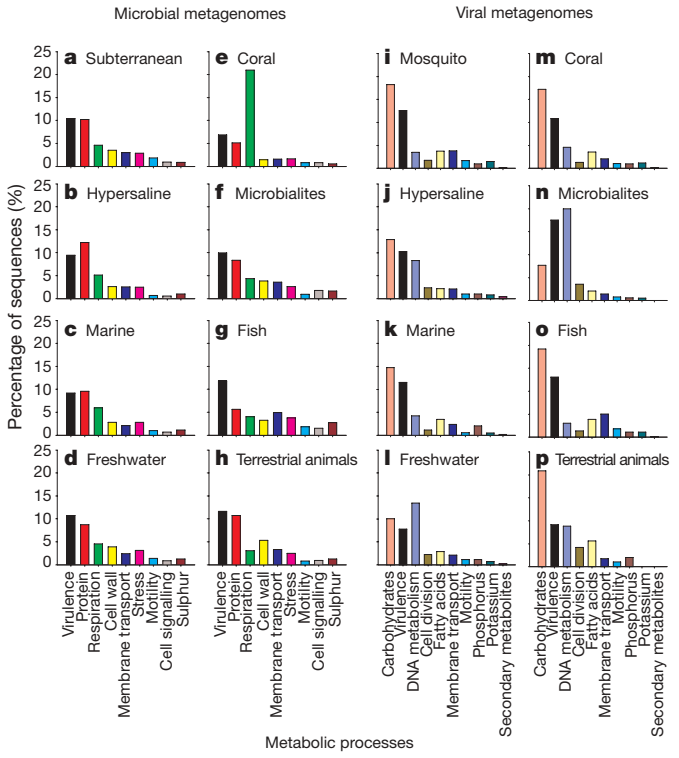

Figure 2 | A one-dimensional representation of the environmental metabolic profiles for the microbial and viral metagenomes sampled from the nine environments. Microbial metagenomes are shown in **a**–**h**, and viral metagenomes are shown in **i**–**p**. Each bar represents the mean for each metabolic category. For subterranean, $n = 2$ (**a**); for hypersaline, $n = 9$ (**b**); for marine, $n = 8$ (**c**); for freshwater, $n = 4$ (**d**); for coral, $n = 7$ (**e**); for microbialites, $n = 3$ (**f**); for fish, $n = 4$ (**g**); for terrestrial animals, $n = 8$ (**h**); for mosquito, $n = 3$ (**i**); for hypersaline, $n = 12$ (**j**); for marine, $n = 10$ (**k**); for freshwater $n = 4$ (**l**); for coral $n = 6$ (**m**); for microbialites, $n = 3$ (**n**); for fish, $n = 2$ (**o**); and for terrestrial animals, $n = 2$ (**p**).

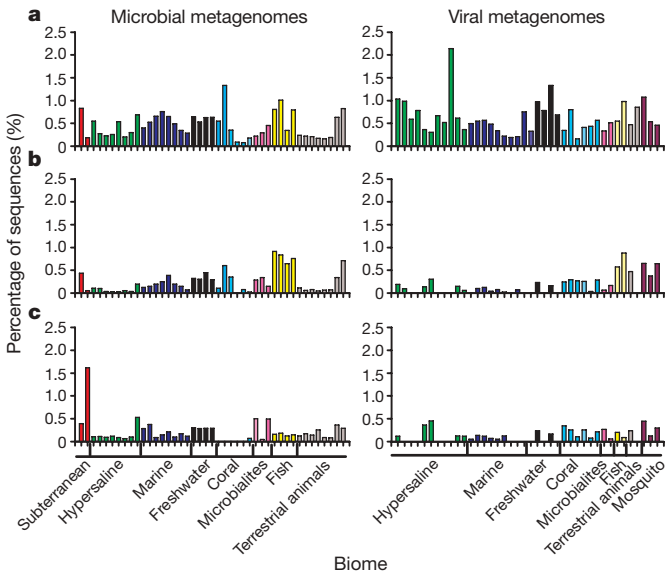

Figure 3 | A comparison of the distribution of sequences similar to motility and chemotaxis genes identified within the microbiomes ($n = 43$) and viromes ($n = 41$). Microbial metagenomes are shown on the left, and viral metagenomes are shown on the right. The abundance of sequences identified within each of three fine-scale subsystems including flagellum (**a**), bacterial chemotaxis (**b**) and gliding motility (**c**), as described by the SEED platform.

sympatric microbes, is controlling gene distribution within an environment. The large amount of variation (~70%) explained by the functional analysis presented here supports this hypothesis.

METHODS SUMMARY

Samples for metagenomes were collected and fractioned using standard techniques, sequenced using pyrosequencing and compared to the functional genes in the SEED platform[11,12] (Methods). All statistics were performed on the percentage of sequences showing similarities to known functions. For the CDA, sequences were grouped according to the SEED classification scheme and the analysis was conducted on the principal metabolic functions. The CDA builds a model for group membership. A discriminative value is calculated for each metagenomic sample, which is a linear combination of the response variables (metabolic processes) represented in the new dimensional space. These values are used to visualize group membership.

An advantage of the CDA is that it identifies which variables are driving the separation between the groups; it uses these to build the model and discards those that are not influential. Identification of influential variables was conducted by a stepwise method, using Wilk's lambda with $P = 0.05$, and was confirmed with analysis of variance (ANOVA; Supplementary Table 4). The level of influence of each variable is provided by the structural matrix and can be visualized using an *h*-plot, in which the length of the line is representative of the level of influence. The CDA also performs a cross-validation analysis that identifies the likelihood of correctly classifying each sample. Cross validation removes the predetermined grouping for each sample and uses the response variables to align the individual sample to a group. Because the data were divided into nine predetermined groups (biomes), the number of samples correctly identified by chance alone is 11%. The percentage-correct classification has to be substantially larger than this number for the metabolic processes to be useful for classifying the metagenomes into environments.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 November 2007; accepted 6 February 2008.
Published online 12 March 2008.

1. Newman, D. K. & Banfield, J. F. Geomicrobiology: how molecular-scale interactions underpin biogeochemical systems. *Science* **296,** 1071–1076 (2002).
2. Prosser, J. I. *et al.* The role of ecological theory in microbial ecology. *Nature Rev. Microbiol.* **5,** 384–392 (2007).
3. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
4. Medini, D. *et al.* The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15,** 589–594 (2005).
5. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311,** 1768–1770 (2006).
6. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).
7. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Genet.* **6,** 805–814 (2005).
8. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7,** 57 (2006).
9. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185,** 6220–6223 (2003).
10. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99,** 14250–14255 (2002).
11. Wegley, L., Breitbart, M., Edwards, R. A. & Rohwer, F. Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environ. Microbiol.* **9,** 2707–2719 (2007).
12. Angly, F. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4,** e368 (2006).
13. Breitbart, M. & Rohwer, F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* **39,** 729–736 (2005).
14. Fierer, N. *et al.* Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of Bacteria, Archaea, Fungi, and viruses in soil. *Appl. Environ. Microbiol.* **73,** 7059–7066 (2007).
15. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33,** 5691–5702 (2005).
16. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Microbiol.* **17,** 589–596 (2001).
17. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl Acad. Sci. USA* **104,** 11436–11440 (2007).
18. Shashar, N., Cohen, Y. & Loya, Y. Extreme diel fluctuations of oxygen in diffusive boundary layers surrounding stony corals. *Biol. Bull.* **185,** 455–461 (1993).
19. Iwanicka-Nowicka, R. *et al.* Regulation of sulfur assimilation pathways in *Burkholderia cenocepacia*: identification of transcription factors CysB and SsuR and their role in control of target genes. *J. Bacteriol.* **189,** 1675–1688 (2007).
20. Aksnes, A., Hope, B., Hostmark, O. & Albrektsen, S. Inclusion of size fractionated fish hydrolysate in high plant protein diets for Atlantic cod, *Gadus morhua*. *Aquaculture* **261,** 1102–1110 (2006).
21. Dinsdale, E. A. *et al.* Microbial ecology of four coral atolls in the Northern Line Islands. *Plos One* **3,** e1584 (2008).
22. Sano, E., Carlson, S., Wegley, L. & Rohwer, F. Movement of virus between biomes. *Appl. Environ. Microbiol.* **70,** 5842–5846 (2004).
23. Davis, B. M. & Waldor, K. in *Mobile DNA II* (ed. Craig, N. L., Gragie, R., Gellert, M. & Lambowitz, A. M.) 1040–1055 (ASM, Washington DC, 2002).
24. Rohwer, F. *et al.* The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* **45,** 408–418 (2000).
25. Mann, N. *et al.* Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424,** 741 (2003).
26. Hendrix, R. W., Smith, M. C. M., Burns, R. N. & Ford, M. E. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl Acad. Sci. USA* **96,** 2192–2197 (1999).
27. Wadhams, G. H. & Armitage, J. P. Making sense of it all: bacterial chemotaxis. *Nature Rev. Mol. Cell Biol.* **5,** 1024–1037 (2004).
28. Hendrix, R. W. Bacteriophage: evolution of the majority. *Theor. Popul. Biol.* **61,** 471–480 (2002).
29. Dunbar, J., Ticknor, L. O. & Kuske, C. R. Assessment of microbial diversity in four southwestern United States soils by 16S rRNA gene terminal restriction fragment analysis. *Appl. Environ. Microbiol.* **66,** 2943–2950 (2000).
30. Frigaard, N.-U., Martinez, A., Mincer, T. J. & Delong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439,** 847–850 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project was supported by the Gordon and Betty Moore Foundation Marine Microbial Initiative, National Science Foundation grants (F.R. and D.L.V.), a Department of Commerce ATP grant (F.R.), a National Research Initiative Competitive Grant from the USDA Cooperative State Research, Education and Extension Service (B.W.), the National Institute of Allergy and Infectious Diseases, the National Institutes of Health and the Department of Health and Human Services (R.S.).

Author Contributions E.A.D. conceptualized the project, conducted the CDA and wrote the manuscript. R.A.E., R.O. and R.S. performed the bioinformatics. D.H. conducted the non-parametric statistical analysis. F.R. oversaw most of the metagenomic projects. All other authors collected the metagenomes and provided comments on the manuscript.

Author Information The metagenomes used in this paper are freely available from the SEED platform and are being made accessible from CAMERA and the NCBI Short Read Archive. The accession numbers are shown in Supplementary Table 1. The NCBI genome project IDs used in this study are: 28619, 28613, 28611, 28609, 28607, 28605, 28603, 28601, 28599, 28597, 28469, 28467, 28465, 28463, 28461, 28459, 28457, 28455, 28453, 28451, 28449, 28447, 28445, 28443, 28441, 28439, 28437, 28435, 28433, 28431, 28429, 28427, 28425, 28423, 28421, 28419, 28417, 28415, 28413, 28411, 28409, 28407, 28405, 28403, 28401, 28395, 28393, 28391, 28389, 28387, 28385, 28383, 28381, 28379, 28377, 28375, 28373, 28371, 28361, 28359, 28357, 28355, 28353 and 28351. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.A.D. (elizabeth_dinsdale@hotmail.com).

METHODS

Collection of the metagenomes. Metagenomic samples were collected and DNA was prepared by the different groups involved; each laboratory used slight modifications on the basic protocol. Sample locations were widely dispersed or separate organisms (Supplementary Fig. 1 and Supplementary Table 1). Metagenomes were collected to answer broad ecological questions such as viral community dynamics in the lungs of healthy and cystic fibrosis patients and the microbial communities on coral reefs (Supplementary Table 1). Typically, the microbiome process starts by filtering samples onto 0.22 μm Sterivex filters, removing the filter membranes and extracting DNA using a bead-beating protocol (MoBio). In some samples, the DNA was amplified with Genomiphi (GE Healthcare Life Sciences) in six to eight 18-h reactions[22,31]. The reactions were pooled and purified using silica columns (Qiagen). The DNA was precipitated with ethanol and resuspended in water at a concentration of approximately 300 ng μl^{-1}. Microbial metagenomes capture Bacteria, Archaea, some small protists as well as a few trapped viral-like particles (Supplementary Table 2).

The viruses in the small metagenomic fractions (that is, 0.22-μm filtrate treated with chloroform) were purified using caesium chloride (CsCl) step gradients to remove free DNA and any cellular material[10,12]. Viral samples were visually checked for microbial contamination using epifluorescent microscopy. Viral DNA was isolated using CTAB (cyltrimethylammonium bromide) and 25:24:1 phenol:chloroform:isoamyl alcohol mix extractions and amplified using Genomiphi reactions. These reactions were pooled and purified using silica columns (Qiagen). The DNA was precipitated with ethanol and resuspended in water at a concentration of approximately 300 ng μl^{-1}. One viral metagenome (number 40, Supplementary Table 1) was prepared by concentrating a natural microbial sample and inducing it with mitomycin C. All metagenome libraries consisted of approximately 5 μg of DNA. The viral metagenomes contained viruses, phage and prophage, and as expected the proportion of phage and prophage are higher in these metagenomes than in the microbial fraction (Supplementary Table 2).

Sequencing and bioinformatics. Sequencing was performed using pyrosequencing on Roche Applied Sciences and 454 Life Sciences GS20 platforms[32] with a practical limit of 105 bp. DNA sequences were analysed in the metagenomics RAST pipeline—an open-access metagenome curation and analysis platform (http://metagenomics.theseed.org/)[33]. First, sequences were screened to remove exactly duplicated sequences—a known artefact of the pyrosequencing approach. The sequences were compared to the SEED platform, which comprises all known protein sequences, using the NCBI BLASTX algorithm on the NMPDR compute cluster (Argonne National Laboratory; http://www.nmpdr.org/). The SEED platform includes all available genome data, DNA and protein sequences, and is supplemented with data from genome sequencing centres as available. Every metagenome was compared to exactly the same data set using the same BLAST parameters at the same time to ensure congruity of the data. Connections between the metagenomes and the SEED subsystems were calculated by identifying matches to the SEED platform where the matched protein was curated to be in a subsystem, and the expect value from the BLAST search was less than 0.001. The SEED subsystems are manually curated collections of proteins with related functions and are available at http://www.theseed.org/. Simultaneously, all sequences were compared to the 16S databases using BLASTN. The databases were extracted from GreenGenes[34], the Ribosomal Database Project[35] and the European Ribosomal Database Project[36].

Several metagenomes were constructed from environments that were likely to contain DNA from other organisms such as humans, corals and mosquitoes. To test and to remove contaminants, 20,000 sequences were chosen at random from every metagenome and compared to the March 2006 build of the human genome and the February 2003 build of the *Anopheles gambiae* genome (both downloaded from http://genome.ucsc.edu/). The comparisons were performed using BLASTN with an expect (E) value cutoff of 1×10^{-5}. Every sample (including the mosquito samples) had less than 1% of their sequences with significant similarity to the *A. gambiae* genome, and only two samples had >5% of sequence similarity to the human genome. These two samples, from the human virome studies, were compared in full and human sequences excluded. To identify and remove dinoflagellate sequences, such as *Symbiodinium* (a coral symbiont), a custom database was created from the nucleotide and RNA (expressed sequence tag) sequences in GenBank; all coral reef water and coral samples were analysed as described above and dinoflagellates sequences were excluded.

Statistical analysis. Statistics were performed on the proportions of sequences within each subsystem, thus normalizing data across metagenomes and removing differences in reaction efficiencies. Total numbers of sequences and numbers of sequences that showed similarities to the SEED are provided in Supplementary Table 1, and ~11% of sequences were similar to functional genes. The SEED platform housed 654 well-documented subsystems that were used to calculate the Shannon index (H'). Maximum diversity occurs when every functional category is present in equal numbers, thus $H_{max} = \log S$, where S is number of categories. Evenness is H' divided by the number of subsystems in each sample (evenness ranges from 0 to 1, which is even). As a comparison to the metagenomic analyses, the diversity and evenness was calculated for all 842 sequenced bacterial genomes. These calculations were conducted on the number of genes within each subsystem, rather than on the number of sequences that was used for the metagenomes (Supplementary Fig. 3).

To analyse the stability of the CDAs, an experiment was conducted in which several of the metagenomic groups were removed and the analysis re-run. In the first trial, the subterranean, fish and mosquito metagenomes were removed (Supplementary Fig. 5). In the second trial, these metagenomes were re-added and the hypersaline metagenomes removed (Supplementary Fig. 6). Multiple trials were required because CDAs are sensitive to the number of samples (that is, metagenomes) relative to the number of variables (that is, metabolic processes).

The data were further analysed using a non-parametric ANOVA, a Kruskal–Wallis test and a median test, and the results compared to ensure that stable results could be obtained (Supplementary Table 3). Environments driving the variation were identified using Duncan comparisons (degrees of freedom were set at 7).

All metagenomes were provided by authors of this manuscript. Further material, including direct access to the data, is available at http://www.theseed.org/DinsdaleSupplementalMaterial/. The NCBI genome project IDs used in this study that were associated with previous publications are: 28369, 28367, 28365, 28363, 28349, 28347, 28345, 28343, 19145 17771, 17769, 17767, 17765 17635, 17633 and 17401.

31. Gunn, M. R. *et al.* A test of the efficacy of whole-genome amplification on DNA obtained from low-yield samples. *Mol. Ecol. Notes* **7**, 393–399 (2007).
32. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
33. Meyer, F. *et al.* The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinf.* (submitted).
34. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
35. Cole, J. R. *et al.* The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* **35**, D169–D172 (2007).
36. Wuyts, J., Perriere, G. & de Peer, Y. V. The European ribosomal RNA database. *Nucleic Acids Res.* **32**, D101–D103 (2004).