11. Boschker, H. T. S. *et al.* Direct linking of microbial populations to specific biogeochemical processes by ¹³C-labelling of biomarkers. *Nature* **392**, 801–805 (1998).
12. Radajewski, S., Ineson, P., Parekh, N. R. & Murrell, J. C. Stable-isotope probing as a tool in microbial ecology. *Nature* **403**, 646–649 (2000).
13. Meselson, M. & Stahl, F. W. The replication of DNA in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **44**, 671–682 (1958).
14. Schildkraut, C. L. in *Methods in Enzymology* (eds Grossman, L. & Moldave, K.) 695–699 (Academic Press, New York, 1967).
15. Radajewski, S. & Murrell, J. C. Stable isotope probing for detection of methanotrophs after enrichment with ¹³CH₄. *Methods Mol. Biol.* **179**, 149–157 (2002).
16. Bodrossy, L. *et al.* Development and validation of a diagnostic microbial microarray for methanotrophs. *Environ. Microbiol.* **5**, 566–582 (2003).
17. Dedysh, S. N. *et al.* Isolation of acidophilic methane-oxidizing bacteria from northern peat wetlands. *Science* **282**, 281–284 (1998).
18. Morris, S. A., Radajewski, S., Willison, T. W. & Murrell, J. C. Identification of the functionally active methanotroph population in a peat soil microcosm by stable-isotope probing. *Appl. Environ. Microbiol.* **68**, 1446–1453 (2002).
19. Radajewski, S. *et al.* Identification of active methylotroph populations in an acidic forest soil by stable-isotope probing. *Microbiology* **148**, 2331–2342 (2002).
20. Hutchens, E., Radajewski, S., Dumont, M. G., McDonald, I. R. & Murrell, J. C. Analysis of methanotrophic bacteria in Movile Cave by stable isotope probing. *Environ. Microbiol.* **6**, 111–120 (2004).
21. Lin, J.-L. *et al.* Molecular diversity of methanotrophs in Transbaikal soda lake sediments and identification of potentially active populations by stable isotope probing. *Environ. Microbiol.* **6**, 1049–1060 (2004).
22. Whitby, C. B. *et al.* ¹³C incorporation into DNA as a means of identifying the active components of ammonia-oxidizer populations. *Lett. Appl. Microbiol.* **32**, 398–401 (2001).
23. Miller, L. G. *et al.* Degradation of methyl bromide and methyl chloride in soil microcosms: use of stable C isotope fractionation and stable isotope probing to identify reactions and the responsible microorganisms. *Geochim. Cosmochim. Acta* **68**, 3271–3283 (2004).
24. McDonald, I. R. *et al.* A review of bacterial methyl halide degradation: biochemistry, genetics and molecular ecology. *Environ. Microbiol.* **4**, 193–203 (2002).
25. Ginige, M. P. *et al.* Use of stable-isotope probing, full-cycle rRNA analysis, and fluorescence *in situ* hybridization–microautoradiography to study a methanol-fed denitrifying microbial community. *Appl. Environ. Microbiol.* **70**, 588–596 (2004).
26. Padmanabhan, P. *et al.* Respiration of ¹³C-labeled substrates added to soil in the field and subsequent 16S rRNA gene analysis of ¹³C-labeled soil DNA. *Appl. Environ. Microbiol.* **69**, 1614–1622 (2003).
27. Jeon, C. O. *et al.* Discovery of a bacterium, with distinctive dioxygenase, that is responsible for *in situ* biodegradation in contaminated sediment. *Proc. Natl Acad. Sci. USA* **100**, 13591–13596 (2003).
28. Wackett, L. P. Stable isotope probing in biodegradation research. *Trends Biotechnol.* **22**, 153–154 (2004).
29. Manefield, M., Whiteley, A. S. & Bailey, M. J. What can stable isotope probing do for bioremediation? *Inter. Biodeter. Biodegradation* **54**, 163–166 (2004).
30. Lueders, T., Pommerenke, B. & Friedrich, M. W. Stable-isotope probing of microorganisms thriving at thermodynamic limits: syntrophic propionate oxidation in flooded soil. *Appl. Environ. Microbiol.* **70**, 5778–5786 (2004).
31. Manefield, M., Whiteley, A. S., Griffiths, R. I. & Bailey, M. J. RNA stable isotope probing, a novel means of linking microbial community function to phylogeny. *Appl. Environ. Microbiol.* **68**, 5367–5373 (2002).
32. Manefield, M., Whiteley, A. S., Ostle, N., Ineson, P. & Bailey, M. J. Technical considerations for RNA-based stable isotope probing: an approach to associating microbial diversity with microbial community function. *Rapid Commun. Mass Spectrom.* **16**, 2179–2183 (2002).
33. Mahmood, S., Paton, G. I. & Prosser, J. I. Cultivation-independent *in situ* molecular analysis of bacteria involved in degradation of pentachlorophenol in soil. *Environ. Microbiol.* (in the press).
34. Lueders, T., Manefield, M. & Friedrich, M. W. Enhanced sensitivity of DNA- and rRNA-based stable isotope probing by fractionation and quantitative analysis of isopycnic centrifugation gradients. *Environ. Microbiol.* **6**, 73–78 (2004).
35. Lueders, T., Wagner, B., Claus, P. & Friedrich, M. W. Stable isotope probing of rRNA and DNA reveals a dynamic methylotroph community and trophic interactions with fungi and protozoa in oxic rice field soil. *Environ. Microbiol.* **6**, 60–72 (2004).
36. Singh, B. K., Millard, P., Whiteley, A. S. & Murrell, J. C. Unravelling rhizosphere–microbial interactions: opportunities and limitations. *Trends Microbiol.* **12**, 386–393 (2004).
37. Griffiths, R. I. *et al.* ¹³CO₂ pulse labelling of plants in tandem with stable isotope probing: methodological considerations for examining microbial function in the rhizosphere. *J. Microbiol. Methods* **58**, 119–129 (2004).
38. Rangel-Castro, J. I. *et al.* Stable isotope probing analysis of the influence of liming on root exudate utilization by soil microorganisms. *Environ. Microbiol.* (in the press).
39. Schloss, P. D. & Handelsman, J. Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* **14**, 303–310 (2003).
40. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685 (2004).
41. Rondon, M. R. *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).
42. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso sea. *Science* **304**, 66–74 (2004).
43. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
44. Radajewski, S., McDonald, I. R. & Murrell, J. C. Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. *Curr. Opin. Biotechnol.* **14**, 296–302 (2003).
45. Bull, I. D., Parekh, N. R., Hall, G. H., Ineson, P. & Evershed, R. P. Detection and classification of atmospheric methane oxidizing bacteria in soil. *Nature* **405**, 175–178 (2000).

🌐 Online links

OPINION

Viral metagenomics

Robert A. Edwards and Forest Rohwer

Abstract | Viruses, most of which infect microorganisms, are the most abundant biological entities on the planet. Identifying and measuring the community dynamics of viruses in the environment is complicated because less than one percent of microbial hosts have been cultivated. Also, there is no single gene that is common to all viral genomes, so total uncultured viral diversity cannot be monitored using approaches analogous to ribosomal DNA profiling. Metagenomic analyses of uncultured viral communities circumvent these limitations and can provide insights into the composition and structure of environmental viral communities.

The genomic age began in 1977 when ΦX174, a virus that infects *Escherichia coli*, was sequenced[1]. The metagenomics of viruses began in 2002 with the publication of two uncultured marine viral communities[2]. In both cases, the small size of viral genomes — approximately 50 kb on average[3,4] — was an advantage because less sequencing was required. However, several unique challenges are encountered when sequencing viruses that are not associated with the sequencing of cellular organisms. These challenges include the abundance of free DNA in the environment[5,6], viral genes that kill the cloning host cells[7] and unclonable, modified viral DNA[8]. These problems have now been overcome (BOX 1) and viral metagenomic libraries are starting to provide information about the types of viruses that are present in environmental samples.

Diversity of environmental viruses
Viral metagenomes mostly comprise novel sequences. There are currently five published viral metagenomic libraries, all of which contain sequences from double-stranded DNA viruses only (see BOX 1): two from near-shore marine water samples[2], a marine sediment sample[9], a human faecal sample[10] and an equine faecal sample[11]. When the marine sequences were first published, approximately 65% of them had no significant similarity (E-VALUE (see Glossary) >0.001) to any sequence in the GenBank non-redundant database (FIG. 1). Analyses 2 years later revealed that most of the viral sequences are still unique, despite the fact that the GenBank database has since more than doubled in size. Likewise, 68% of
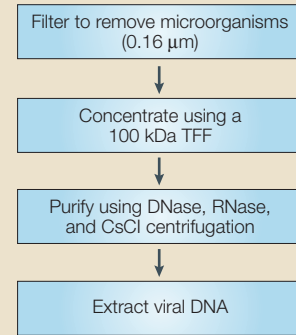
Box 1 | Cloning considerations and viral metagenomics

Isolating representative viral community DNA for metagenomic analyses is complicated by the presence of free[5,6] and cellular DNA. The viral DNA signal will be lost if the free DNA is not removed[41,42]. Similarly, at ~50 kb long[3,4], the average viral genome is about 50 times smaller than the average microbial genome[43] (2.5 Mb), so any cellular contamination will overwhelm the viral signal. A typical starting sample consists of 200 litres of seawater or 1 kg of solid material. Faecal, soil and sediment samples are resuspended in osmotically neutral solutions before filtration. A combination of differential filtration with tangential flow filters (TFF), DNase treatment and density centrifugation in caesium chloride (CsCl) is used to separate the intact viral particles from the microorganisms and free DNA. Very large or very small viruses will be lost in the filtration step, and those sensitive to CsCl will also disintegrate in this step. This protocol seems to capture most of the viral community however, as assessed by pulse-field gel electrophoresis[4] and epifluorescent microscopy[2].
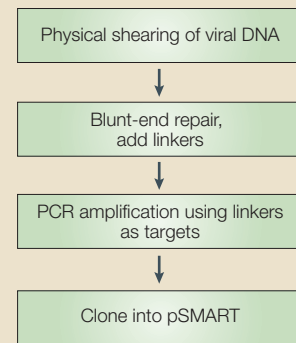
Once intact virions have been isolated, the viral DNA is extracted and cloned. Cloning representative viral metagenomes is challenging, owing to low DNA concentrations (~10^{-17} g DNA per virion), modified DNA (such as alternative bases, for example, 5-(4-aminobutyl-aminomethyl) uracil and 5-methyl cytosine[8]) and the presence of lethal viral genes such as holins and lysozymes. In most water samples, it is necessary to concentrate virions from several hundred litres to obtain enough DNA for cloning. The linker-amplified shotgun library (LASL) technique includes a PCR amplification step, which makes it possible to clone small amounts of DNA (1–100 ng). The PCR step also converts modified DNA into unmodified DNA. A shearing step disrupts lethal virus genes by shearing DNA into small fragments (~2 kb) and provides the random fragments necessary for community modelling. Using this protocol, it is possible to make representative metagenomic libraries that contain viral fragments that are proportional to their concentrations in the original sample[44]. LASLs typically contain millions of random clones.

RNA and single-stranded DNA (ssDNA) viruses cannot be cloned using this approach. However, preliminary studies with random-primed reverse transcriptase and random-primed strand-displacement DNA polymerases indicate that these viral groups could be analysed using metagenomic approaches (F. R., D. Mead, and Y. Ruan, unpublished data).

a Isolating uncultured viral communities from seawater

Filter to remove microorganisms (0.16 μm)
↓
Concentrate using a 100 kDa TFF
↓
Purify using DNase, RNase, and CsCl centrifugation
↓
Extract viral DNA

b Construction of LASLs

Physical shearing of viral DNA
↓
Blunt-end repair, add linkers
↓
PCR amplification using linkers as targets
↓
Clone into pSMART

---

the sequences in the newly published equine faecal metagenome have no similarity to any sequence in GenBank[11]. Genomic analyses of cultured PHAGES also show that most of the OPEN READING FRAMES (ORFs) are novel[12–14]. By contrast, only about 10% of the sequences from environmental microbial metagenomes[15,16] and cultured microbial genomes[17] are novel when analysed in similar ways. Together, these observations indicate that much of the global microbial metagenome has been sampled, whereas the global viral metagenome is still relatively uncharacterized. Daubin and Ochman have hypothesized that the unique genes in microbial genomes were acquired from the phage genomic pool[17].

Phage phylogeny and taxonomy. For cellular organisms, phylogenetic and taxonomic relationships can be derived from the universal ribosomal DNA (rDNA) sequences. However, this technique is not applicable to viruses because there is no single sequence that is present in all viral genomes[18]. Official viral classification is based on characteristics of the virions and host range, not sequence data[19]. Recently, several different approaches have been proposed for sequence-based

systems of viral classification[18,20]. The most common sequence-based approach to viral identity and taxonomy is to use a single gene locus, such as a capsid or DNA polymerase gene, to characterize a specific viral group. Primers for PCR can then be designed for these genes, and the diversity of specific genes in the environment can be assessed by cloning and sequencing of DNA products amplified directly from environmental samples. This single-locus approach has been used to show that there are many groups of uncultured viruses in the environment and that viruses move between BIOMES[21–29]. Although the single-locus approach might work for specific groups of viruses, lateral gene transfer between genomes can make interpretation of these data complicated, if not impossible[30].

The Phage Proteomic Tree[18] is a taxonomic system based on an algorithm that uses every gene in every phage genome to determine an average distance between pairs of phages. FIGURE 2 shows a new version of the Phage Proteomic Tree. In general, the additional genomes that were incorporated did not change the groupings that were proposed in the first version of the tree[18]. In many cases, the branch lengths connecting a particular

clade to other clades have increased, indicating stronger clustering within each clade. Together, these observations indicate that clades will remain the same in the future. About 20% of the phage genomes still fall outside of any clade (FIG. 2), indicating that many new phage families remain to be discovered and characterized.

Metagenomes only contain partial sequence fragments from the viral COMMUNITY. The Phage Proteomic Tree is potentially suited to analyses of metagenomic data because all of the genomic sequence is considered. To evaluate if these partial sequence fragments are useful for determining the taxonomic relationships among uncultured phages, the following *in silico* experiment was carried out. Phage genomes from each clade were independently fragmented into sequential 500-bp fragments. One thousand of these DNA fragments were picked at random and then compared against the phage protein database using blastx. The most significant hit for each fragment was the genome from which the fragment originated and this hit was therefore ignored. The second significant hit was recorded. If this second hit had an E-value <0.001 it was determined whether
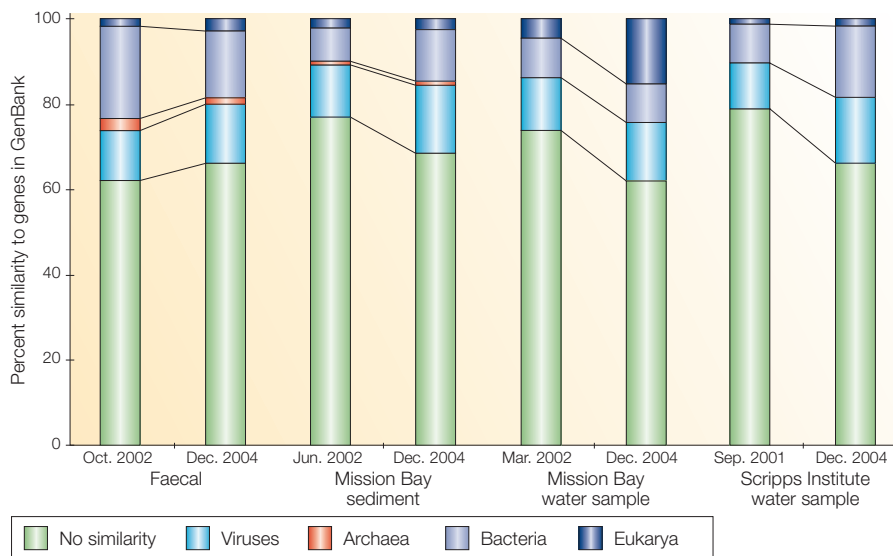
Figure 1 | **Comparison of viral metagenomic libraries to the GenBank non-redundant database.**
Viral metagenomic sequences from human faeces[10], a marine sediment sample[9] and two seawater samples[2] were compared to the GenBank non-redundant database at the date of publication and in December 2004. The percentage of each library that could be classified as Eukarya, Bacteria, Archaea, viruses or showed no similarities (E-value >0.001) is shown.

the hit was to a phage within the same clade, as outlined on the Phage Proteomic Tree. In two separate *in silico* experiments, 91% and 98% of the fragments fell into the same clade as the original phage. This shows that partial genomic-sequence fragments could be used to predict the identity of phages described in the Phage Proteomic Tree. This type of approach could easily be extended to analysis of microbial genomes and microbial metagenomes.

Comparison of the available viral metagenomes with the Phage Proteomic Tree showed that, overall, Siphophage fragments are the most common fragments observed in the published metagenomic libraries[2,9–11]. In particular, Siphophage comprise 44% of phage sequences in the sediment library. Oceanic sediments, including the biologically active subsurface[31], contain the largest microbial biomass on the planet[32]. Viruses are present in these environments[33,34],

which indicates that Siphophages might be the most abundant genome arrangement on Earth. This hypothesis is supported by unpublished data (F.R. and M. Breitbart) showing that viral communities from the second largest biome on the planet, the terrestrial subsurface[32], are also dominated by Siphophages.

Whole-genome taxonomy systems for metagenome analyses also enable investigators to carry out statistical comparisons between different communities to determine their phylogenetic similarity. One example is the permutation tail probability (PTP) test, which uses trees to determine whether any particular taxonomic group is preferentially associated with one environment or another[35]. Breitbart *et al.*[9] used PTP tests to show that marine phage communities are phylogenetically similar, regardless of whether these communities are in the water column or in the sediment.

The proviral metagenome. Many viruses integrate into the genome of the bacterial host and persist as proviruses. Sixty percent of sequenced bacterial genomes contain at least one prophage[36]. The number of prophages varies by genome. For example, prophages contribute approximately 13% of the genome of *Streptococcus pyogenes* strain MGAS315, and 10% of the genome of *Xylella fastidiosa* strain Temecula1. On average, however, about 3% of genomic DNA content is composed of prophages. Approximately 75% of the genes in prophage genomes have no known function (R.A.E., unpublished results).

Some of the fragments in microbial metagenomic libraries are actually prophage genes. Comparison of 964,094 ORFs from the Sargasso Sea metagenome revealed that 3,215 ORFs had significant similarity (E-value $\leq 1 \times 10^{-5}$) to known phage genes (REF. 16 and R.A.E., unpublished results). Sargasso Sea samples with similarity to either *Shewanella* spp. or *Burkholderia* spp. were excluded from these analyses (see also the article by E.F. DeLong in this issue). All of the phage genes identified in the microbial metagenome were well-characterized, including genes encoding integrases, capsid proteins, terminases and tail fibres. This analysis does not reflect those phages or phage-encoded proteins that have not been characterized. As approximately 65% of phage genes have no homologues at all, even within other phage genomes or with sequenced phage genes, we estimate that about 1% of the microbial metagenomes encode phage proteins.

Glossary

BIOME
An important ecosystem type, usually used to describe a distinctive primary producer assemblage such as a temperate forest.

BLAST
Comparisons of sequences with databases are commonly done with BLAST and/or FASTA. Both programs allow comparison of either a nucleotide or protein query sequence with either a nucleotide or protein database.

COMMUNITY
A group of different populations within a specific area.

COMMUNITY STRUCTURE
The relative abundance of different populations in relation to each other, often graphed as a rank–abundance curve.

EXPECT VALUE (E-VALUE)
A parameter that describes the number of hits that would be 'expected' to occur by chance when searching a sequence database of a particular size. An E-value of 1 means that it would be expected to find a match with a similar score simply by chance. The lower the E-value, the more significant the match.

OPEN READING FRAME
Open reading frames (ORFs) are essentially the same as genes. They are also referred to as protein-coding regions. ORFs are identified in genomes by several algorithms, most of which search for stretches of DNA sequence without stop codons.

PHAGE
A virus that infects bacteria. Because bacteria are the most common targets on a global scale, most environmental viruses are phages.

POPULATION
The total count of individuals belonging to one species in a specific area.

RANK–ABUNDANCE CURVE
Graphs of community structure. In these graphs, the most abundant species has a rank of 1, the next most abundant is 2, and so on, on the x-axis. The y-axis represents the abundance of each species.

SUBSURFACE
The geological zone below the surface of the Earth. It is not exposed to the Earth's surface.

Viral community structure and ecology

In addition to cataloguing 'what is there?', viral metagenomics makes it possible to reconstruct the structure of uncultured viral communities. This analysis relies on the hypothesis that the occurrence of the same DNA sequence in different clones means that the same genotype has been resampled. To take advantage of this information, a modified version of the Lander–Waterman algorithm[37] was developed[2]. These analyses showed that near-shore, marine water-column viral communities contained ~5,000 genotypes per 200 litres of water. The modified Lander–Waterman approach was complemented with Monte-Carlo simulations by Breitbart et al. for analyses of marine sediments[9]. Both approaches described similar COMMUNITY STRUCTURES, indicating that the assumptions underlying the models are robust. There is now an online tool that will predict viral community structure from metagenomic data[38] (see PHACCS in Online links box).

FIGURE 3a shows a Monte-Carlo analysis of the viral communities found in human faeces, seawater and marine sediments. Each of the samples contained approximately the same number of viral particles (~10^{12}), but the community structures are dramatically different. The faecal sample only contained ~1,000 viral genotypes, the seawater samples most probably contained ~5,000 viral genotypes and the marine sediment sample contained between 10,000 and 1 million viral genotypes. The faecal and marine water-column viral communities each contained a different dominant viral genotype that constituted at least 1% of the total community, whereas the most dominant virus in the marine-sediment community made up less than 0.01% of the total community. Based on these analyses, marine sediment viral communities are the most diverse biological systems characterized to date[9].

When the community distribution shown in FIG. 3a is plotted on a standard RANK–ABUNDANCE CURVE, the shape of the resulting curve is important for developing ecological models of viral dynamics. To determine the shape of this curve, different mathematical functions are compared to the observed data. The error between different idealized functions and the observations are then determined. In this way, it is possible to determine which function best describes the community structure. Such analyses on the viral communities shown in FIG. 3a determined that a power-law function best describes the shape of the curve[38].
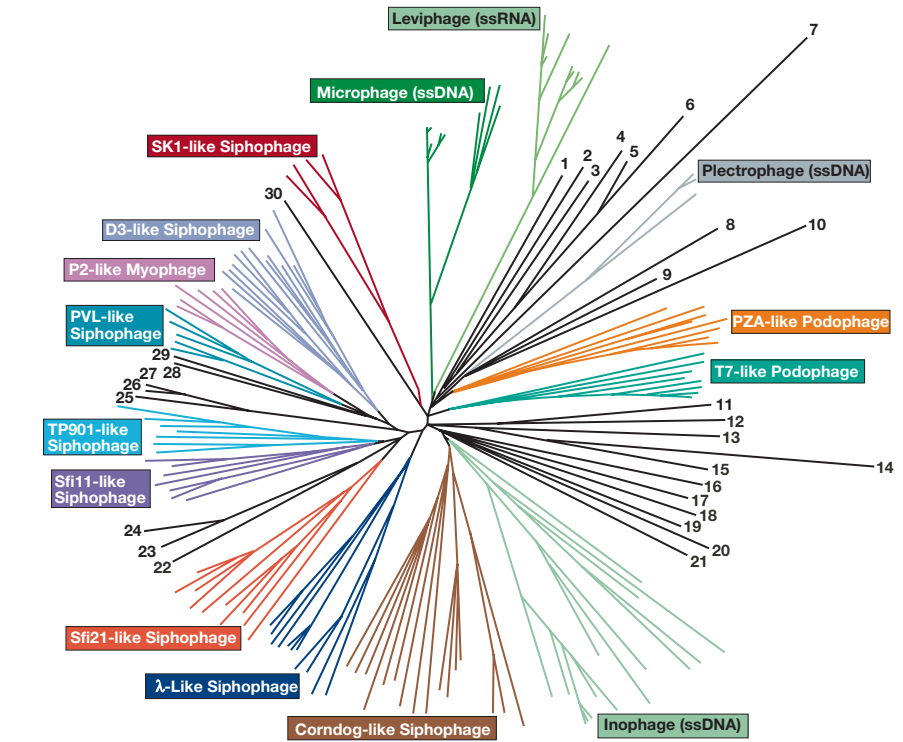
chp1-like. *Bdellovibrio bacteriovoru*s φ MH2K | *Spiroplasma citri* φ SpV4 | *Chlamydia psittaci* φ chp1 | *C. psittaci* φ 2 | *C. psittaci* φ PhiCPG1 | *Chlamydia pneumoniae* φ CPAR39
X174-like. *Escherichia coli* φ K | *E. coli* φ α3 | *E. coli* φ G4 | *E. coli* φ 174 | *Salmonella sps.* φ S13

Acinetobacter spp. φ AP205 | *E. coli* φ MX1 | *E. coli* φ M11 | *E. coli* φ SP | *E. coli* φ NL95 | *E. coli* φ MS2 | *E. coli* φ fr | *E. coli* φ KU1 | *E. coli* φ GAb | *Pseudomonas aeruginosa* φ PP7

S. citri φ SVTS2 | *Spiroplasma* φ 1-C74 | *S. citri* φ SpV1

Enteric bacteria φ PRD1 | *Sulfolobus islandicus* φ fv1 | *Mycoplasma* sp. φ P1 | *Staphylococcus* φ 44AHJD | *Staphylococcus aureus* φ P68 | *Streptococcus pneumoniae* φ Cp-1 | *Bacillus subtilis* φ GA-1 | *B. subtilis* φ PZA | *B. subtilis* φ B103

Vibrio φ VpV262 | *P. aeruginosa* φ PaP3 | *Roseobacter* SIO67 φ SIO1 | *Synechococcus* φ p.P60 | *Pseudomonas* φ gh-1 | *E. coli* φ T7 | *E. coli* φ T3 | *Yersinia enterocolitica* φ YeO3-12

Xanthomonas campestris φ Cf1c | *P. aeruginosa* φ Pf3 | *P. aeruginosa* φ Pf1 | *Vibrio parahaemolyticus* φ VfO4K68 | *V. parahaemolyticus* φ VfO3K6 | *Vibrio cholerae* φ fs1 | *V. cholerae* φ VSKK | *Vibrio* φ VSK | *V. cholerae* φ fs-2 | *E. coli* φ If1 | *E. coli* φ Ike | *E. coli* φ I2-2 | *E. coli* φ fd | *E. coli* φ f1 | *E. coli* φ M13

Mycobacterium avium φ TM4 | *Streptomyces* sp. φ C31 | *Mycobacterium smegmatis* φ L5 | *M. smegmatis* φ D29 | *M. smegmatis* φ Bxb1 | *M. smegmatis* φ Bxz2 | *M. smegmatis* φ Rosebush | *Myxococcus xanthus* φ Mx8 | *Thermus aquaticus* φ IN93 | *Mycobacterium* φ Che9c | *Mycobacterium* φ ω | *Mycobacterium* φ Corndog | *Mycobacterium* φ CJW1 | *Mycobacterium* φ Che9d | *Mycobacterium* φ Che8 | *Mycobacterium* φ Barnyard

Salmonella enterica serovar Typhimurium φ P22 | *S. typhimurium* φ ST64T | *E. coli* φ HK620 | Pea Aphid φ APSE-1 | *E. coli* φ 933W | *E. coli* φ Stx2 | *E. coli* φ VT2-Sa | *E. coli* φ HK022 | *E. coli* φ HK97 | *E. coli* φ λ | *E. coli* φ N15

Streptococcus thermophilus φ Sfi21 | *S. thermophilus* φ Sfi19 | *S. thermophilus* φ DT1 | *S. thermophilus* φ 7201 | *Lactobacillus* φ adh | *Lactobacillus lactis* φ BK5-T | *Lactobacillus* φ 4268 | *L. lactis* φ bIL286 | *L. lactis* φ bIL309 | *Lactobacillus* sp. φ G1e

S. thermophilus φ Sfi11 | *S. thermophilus* φ O1205 | *Streptococcus pyogenes* φ 315.5 | *S. pneumoniae* φ MM1 | *S. pyogenes* φ 315.4 | *S. pyogenes* φ NIH1.1 | *S. pyogenes* φ 315.6

L. lactis φ TP901-1 | *L. lactis* φ Tuc2009 | *L. lactis* φ ul36 | *L. lactis* φ r1t | *S. pyogenes* φ 315.3 | *S. aureus* φ ETA | *S. aureus* φ 11

S. aureus φ 12 | *S. aureus* φ SLT | *S. aureus* φ PVL | *S. aureus* φ 13 | *S. aureus* pro-φ PV83

V. parahaemolyticus φ Vp16T | *V. parahaemolyticus* φ Vp16C | *Vibrio harveyi* φ VHML | *P. aeruginosa* φ CTX | *E. coli* f P2 | *E. coli* φ 186

Listeria φ 2389 | *Leuconostoc oenos* φ L5 | *L. lactis* φ bIL285 | *L. lactis* φ bIL310 | *L. lactis* φ bIL311 | *L. lactis* φ bIL312 | *Clostridium perfringens* φ 3626 | *Shigella flexneri* φ V (Podo) | *S.typhimurium* φ ST64B] | *E. coli* φ P27 | *P. aeruginosa* φ D3 | *S. pyogenes* φ 315.2 | *S. pyogenes* φ 315.1

L. lactis φ bIL67 | *L. lactis* φ c2 | *L. lactis* φ sk1 | *L. lactis* φ bIL170

1. *Staphylococcus* spp. φ K | **2**. *E. coli* φ T4 (Myo) | **3**. *P. aeruginosa* φ KZ (Myo) | **4**. *Pseudomonas syringae* φ 8 (Cysto) | **5**. *P. syringae* φ 12 | **6**. *P. syringae* φ 6 | 7. *Burkholderia thailandensis* φ E125 | **8**. Propionibacteria φ B5 | **9**. *Pseudoalteromonas espejiana* φ PM2 (Corticol) | **10**. *P. syringae* φ 13 | **11**. *Sinorhizobium meliloti* φ PBC5 | **12**. *E. coli* φ P4 (Myo) | **13**. *E. coli* φ Mu (Myo) | **14**. *Acholeplasma* sp. φ MV-L1 (Ino) | **15**. *Sulfolobus shibitae* φ 1 (Fusello) | **16**. *Mycobacterium* φ Bxz1 | **17**. *Halorubrum coriense* φ HF2 | **18**. *Burkholderia cepacia* Bcep781 | **19**. *Natrialba magadii* φ Ch1 (Myo) | **20**. *Mycoplasma arthritidis* φ MAV1 | **21**. *B. subtilis* φ SPP1 (Sipho) | **22**. *Acholeplasma laidlawii* φ L2 (Plasma) | **23**. *Methanobacterium thermoautotrophicum* φ ψM2 (Sipho) | **24**. *Methanothermobacter wolfeii* pro-φ M100 | **25**. *V. cholerae* φ K139 | **26**. *Haemophilus influenzae* φ HP2 | **27**. *H. influenzae* φ HP1 (Myo) | **28**. *L. casei* φ A2 (Sipho) | **29**. *B. subtilis* φ 105 (Sipho) | **30**. *Listeria monocytogenes* φ A118 (Sipho)

Figure 2 | The Phage Proteomic Tree. The Phage Proteomic Tree is a whole-genome-based taxonomy system that can be used to identify similarities between complete phage genomes and metagenomic sequences. This new version of the tree contains 167 phage genomes. Phages in black cannot be classified into any clade. In the key, each phage is defined in a clockwise direction.
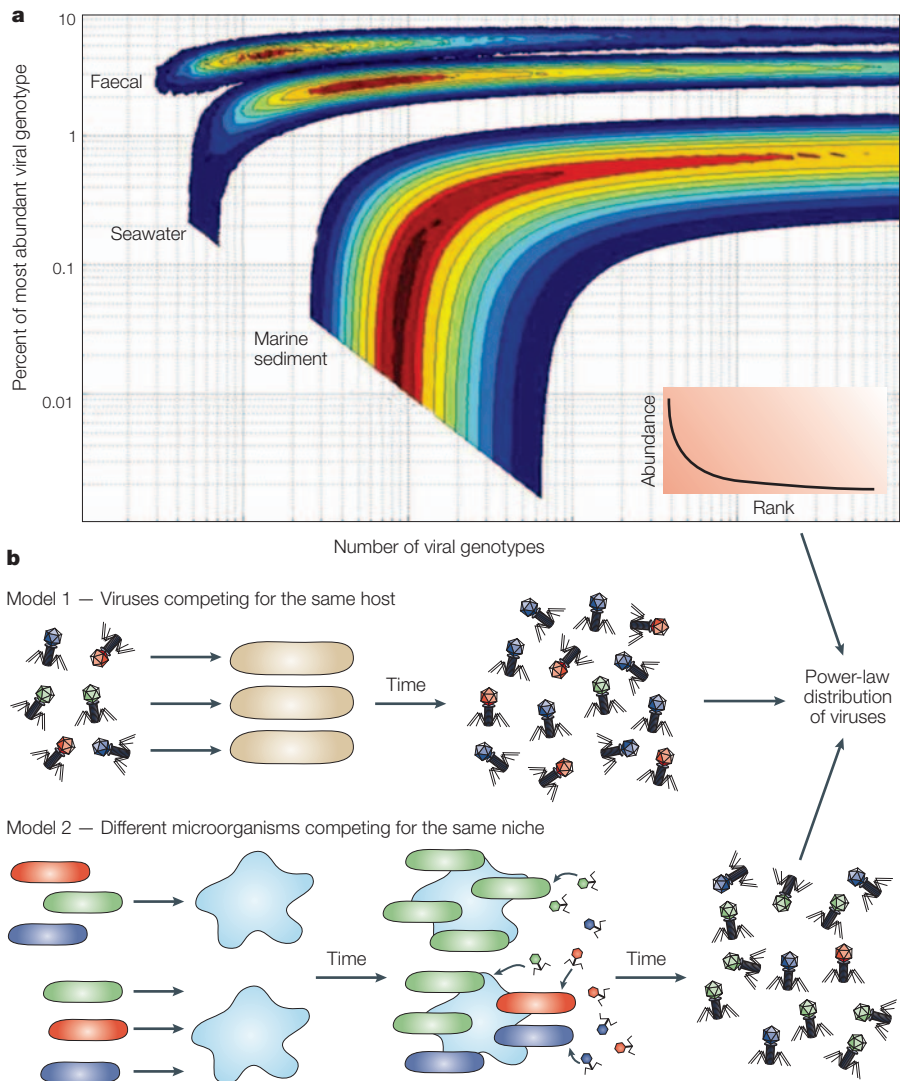
Figure 3 | **Metagenomics and viral diversity, community structure and ecology. a** | Monte-Carlo simulations were used to determine the most probable structure for the phage communities observed in REFS 2,9,10. The red regions are the most probable explanation of the observed data, blue is the least probable explanation. Only the Scripps Pier seawater sample is shown, because it completely overlaps with the Mission Bay seawater sample. This indicates that the two seawater samples have identical community structure even though they are from different places and times. The inset shows a standard rank–abundance curve, where the single most abundant viral genotype is given a rank of 1, the second most abundant genotype is ranked 2, and so on. The rank–abundance curve is used to determine the mathematical function that best describes the community structure. **b** | The community structure of all four viral communities is best described by a power-law function[38]. Models 1 and 2 are two possible explanations for this observation and they are described in the text.

In this example, the microorganisms are competing for the same food source. By chance, one microorganism obtains more food than its competitors and divides more frequently. This will lead to a power-law distribution of the microbial community. When viruses that include those that can infect only one species infect this community, the result will be a power-law distribution of the viral community (that is, 'a power law begets a power law'). Both models are examples of the 'rich-get-richer' idiom. Studies are currently underway to differentiate between these models. These types of approaches will allow metagenomics to advance the theoretical ecology of microbial communities.

Bioinformatics and viral metagenomes

Analysis of metagenomes presents substantial computational challenges. Assembly programs like Phred/Phrap and Sequencher (Gene Codes Corporation) are designed to connect fragments from the same genome. The assembly programs assume, for example, that single-base mismatches represent errors in base-calling. In metagenomic analysis, this assumption is invalid because single-base mismatches might represent different sequences from unique individuals in the metagenome. Currently, the problem is exacerbated in viral metagenomics by the large number of genomes in the samples (possibly several million) and by repeated sequences, such as insertion elements and transposons. These problems might be overcome with different assembly algorithms, longer sequence reads and deeper coverage of the environmental samples.

Sequence assembly facilitates gene identification by providing whole ORFs and operons for analysis. Current gene identification algorithms are optimized to identify ORFs in bacterial or eukaryotic genomes. Little work has been directed towards the identification of ORFs in viral genomes. ORF identification is of little benefit in viral metagenomic libraries that consist of single-sequence reads — it is probable that many of the ORFs will be missed because of sequencing errors or because the read is too short to contain a large enough fraction of the ORF. Again, these limitations will be overcome with longer reads and deeper coverage of the metagenomic libraries.

Sequence assembly is not a prerequisite for comparing metagenomic sequences with databases to determine the gene content of the environment. Most metagenomic comparisons (such as those described above) use the BLAST algorithm for

Power laws are important mathematical functions because they arise from a series of connected, exponential events[39]. Two possible models that might explain why the uncultured viral communities have power-law dynamics are shown in FIG. 3b. In the first model, different viruses are competing for the same microbial host. Stochastic behaviour causes one of the viral genotypes to find a few more hosts than the other virus genotypes. When the viruses go through a lytic cycle, each productive infection produces ~25 new phages (reviewed in REF. 40). In the next round of infections, there are more copies of the phage genotype that found more hosts, so this phage will therefore be more successful at finding a new host and replicating. After a couple of replication cycles, this results in a power-law distribution of these viruses, in which the virus that originally found more hosts dominates the community, while the other viruses are rare. In the second model, one virus can infect only one microbial species.

12. Pedulla, M. L. *et al.* Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171–182 (2003).
13. Rohwer, F. *et al.* The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with non-marine phages. *Limnol. Oceanogr.* **42**, 408–418 (2000).
14. Chen, F. & Lu, J. Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol.* **68**, 2589–2594 (2002).
15. Tyson, G. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
16. Venter, J. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
17. Daubin, V. & Ochman, H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* **14**, 1036–1042 (2004).
18. Rohwer, F. & Edwards, R. The Phage Proteomic Tree: a genome based taxonomy for phage. *J. Bacteriol.* **184**, 4529–4535 (2002).
19. Büchen-Osmond, C. ICTVdB: The universal virus database. *Index of Viruses* [online], <http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/index.htm> (2002).
20. Nelson, D. Phage taxonomy: we agree to disagree. *J. Bacteriol.* **186**, 7029–7031 (2004).
21. Dorigo, U., Jacquet, S. & Humbert, J. Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. *Appl. Environ. Microbiol.* **70**, 1017–1022 (2004).
22. Zhong, Y., Chen, F., Wilhelm, S. W., Poorvin, L. & Hodson, R. E. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl. Environ. Microbiol.* **68**, 1576–1584 (2002).
23. Fuller, N. J., Wilson, W. H., Joint, I. R. & Mann, N. H. Occurence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl. Environ. Microbiol.* **64**, 2051–2060 (1998).
24. Chen, F., Suttle, C. A. & Short, S. M. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Appl. Environ. Microbiol.* **62**, 2869–2874 (1996).
25. Chen, F. & Suttle, C. A. Evolutionary relationships among large double-stranded DNA viruses that infect microalgae and other organisms as inferred from DNA polymerase genes. *Virology* **219**, 170–178 (1996).
26. Short, S. M. & Suttle, C. A. Sequence analysis of marine cirus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl. Environ. Microbiol.* **68**, 1290–1296 (2002).
27. Short, C. & Suttle, C. Nearly identical bacteriophage structural gene sequences are widely distributed in marine and freshwater environments. *Appl. Environ. Microbiol.* **71**, 480–486 (2005).
28. Hambly, E. *et al.* A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc. Natl Acad. Sci. USA* **98**, 11411–11416 (2001).
29. Breitbart, M. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 245–252 (2004).
30. Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. The imbroglios of viral taxonomy: genetic exchange and the failings of phenetic approaches. *J. Bacteriol.* **184**, 4891–4905 (2002).
31. Schippers, A. *et al.* Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria. *Nature* **433**, 861–864 (2005).
32. Whitman, W., Coleman, D. & Wiebe, W. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
33. Danovaro, R., Manini, E. & Dell'Anno, A. Higher abundance of bacteria than viruses in deep Mediterranean sediments. *Appl. Environ. Microbiol.* **66**, 1857–1861 (2002).
34. Danovaro, R. & Serresi, M. Viral density and virus-to-bacterium ratio in deep-sea sediments of the Eastern Mediterranean. *Appl. Environ. Microbiol.* **66**, 1857–1861 (2000).
35. Martin, A. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* **68**, 3673–3682 (2002).

Box 2 | Bioinformatics for metagenomics — the future

Almost all comparisons between metagenomic libraries are currently carried out using sequence-similarity algorithms like BLAST and FASTA. The observation that most sequences in viral metagenomic libraries have no recognizable similarity to the GenBank database indicates that other computational methods are needed. Future analyses of metagenomic sequences should include GC/AT content, codon usage and oligomer skews using different-sized sequence strings such as dinucleotides or trinucleotides. Sequence skews might associate viral sequences with their host or might associate groups of viral sequences[45]. At the protein level, future classifications will be done by predicting protein structure. Viral proteins often contain unexpected folds and architecture (A. Godzik, personal communication). A deeper understanding of the structure of viral proteins and a more diverse selection of crystallized viral proteins are required before structures can be readily used to group metagenomic sequences.

searching sequence space (BOX 2). Typically, translated DNA sequences that are queried against a translated DNA database (tblastx searches) are used to explore similarities between the metagenomic libraries and the sequence databases. Translated searches have the advantage of being less susceptible to errors that would be introduced by frameshifts in the sequence caused by incorrect base-calling than other searches. However, translated searches also have the disadvantage of requiring substantial computing power, and such searches take longer than other sequence comparisons.

The bioinformatics techniques available for phage metagenomic libraries still have several limitations. As discussed above, approximately 65% of metagenomic sequences have no homologues in the non-redundant databases. It is currently unclear whether this is a limitation of the search algorithms, a limitation of the diversity represented in the GenBank database or a combination of both. This problem should be alleviated as more viral sequences are sampled and characterized and distant relationships become more clear.

Future directions in viral metagenomics

The amazing diversity and novelty of viral metagenomes mean that large-scale sequencing efforts like the acid mine drainage[15] and Sargasso Sea[16] projects need to be carried out on the viral component. These surveys will provide the raw data necessary for understanding the size of the viral metagenome and community structure. Methods to clone and sequence ssDNA and RNA viruses also need to be developed and incorporated into these surveys to include all viruses in these analyses. At the bioinformatics level, tools need to be automated and made freely available so individual labs can carry out viral metagenomic analyses on communities of interest. At the evolutionary level, the relationships between horizontally transferred DNA and mobile genetic elements need to be further investigated. In particular, the relationships between unidentified genes in microbial genomes and viral metagenomes need to be explored in more detail. Finally, mathematical models to study POPULATION and community dynamics incorporating metagenomic data need to be developed.

Robert A. Edwards is at the Department of Biology, LS301, and the Center for Microbial Sciences, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA, and at the Fellowship for Interpretation of Genomes, Chicago, Illinois 60527, USA.

Forest Rohwer is at the Department of Biology, LS301, and the Center for Microbial Sciences, San Diego State University, San Diego, California 92182, USA.

*Correspondence to F.R.
e-mail: forest@sunstroke.sdsu.edu*

1. Sanger, F. *et al.* Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature* **265**, 687–695 (1977).
2. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14250–14255 (2002).
3. Wommack, K. E., Ravel, J., Hill, R. T., Chun, J. & Colwell, R. R. Population dynamics of Chesapeake Bay virioplankton: total-community analysis by pulsed-field gel electrophoresis. *Appl. Environ. Microbiol.* **65**, 231–240 (1999).
4. Steward, G., Montiel, J. & Azam, F. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol. Oceanogr.* **45**, 1697–1706 (2000).
5. DeFlaun, M., Paul, J. & Jeffrey, W. Distribution and molecular weight of dissolved DNA in subtropical estuarine and oceanic environments. *Mar. Ecol. Prog. Ser.* **38**, 65–73 (1987).
6. DeFlaun, M. F., Paul, J. H. & Davis, D. Simplified method for dissolved DNA determination in aquatic environments. *Appl. Environ. Microbiol.* **52**, 654–659 (1986).
7. Wang, I., Smith, D. & Young, R. Holins: the protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* **54**, 799–825 (2000).
8. Warren, R. Modified bases in bacteriophage DNAs. *Annu. Rev. Microbiol.* **34**, 137–158 (1980).
9. Breitbart, M. *et al.* Diversity and population structure of a nearshore marine sediment viral community. *Proc. Biol. Sci.* **271**, 565–574 (2004).
10. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **85**, 6220–6223 (2003).
11. Cann, A., Fandrich, S. & Heaphy, S. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**, 151–156 (2005).

36. Casjens, S. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* **49**, 277–300 (2003).

37. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).

38. Angly, F. *et al.* PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41 (2005).

39. Bak, P. *How Nature Works: The Science of Self-Organized Criticality* (Springer–Verlag, New York, 1996).

40. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114. (2000).

41. Paul, J., Jeffrey, W. & DeFlaun, M. Dynamics of extracellular DNA in the marine environment. *Appl. Environ. Microbiol.* **53**, 170–179 (1987).

42. Paul, J., Jiang, S. & Rose, J. Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. *Appl. Environ. Microbiol.* **57**, 2197–2204 (1991).

43. Simon, M. & Azam, F. Protein content and protein synthesis rates of planktonic marine bacteria. *Mar. Ecol. Prog. Ser.* **51**, 201–213 (1989).

44. Rohwer, F. *Construction and analyses of linker-amplified shotgun libraries (LASLs)* [online], <http://www.sci.sdsu.edu/PHAGE/LASL/index.htm> (2002).

45. Blaisdell, B. E., Campbell, A. M. & Karlin, S. Similarities and dissimilarities of phage genomes. *Proc. Natl Acad. Sci. USA* **93**, 5854–5859 (1996).

🌐 Online links

DATABASES
The following terms in this article are linked online to:
Entrez: http://www.ncbi.nlm.nih.gov/Entrez
ΦX147 | *Streptococcus pyogenes* strain MGAS315 | *Xylella fastidiosa* strain Temecula1

FURTHER INFORMATION
Forest Rohwer's laboratory: http://phage.sdsu.edu
BLAST: http://www.ncbi.nlm.nih.gov/BLAST
FASTA: http://www.ebi.ac.uk/fasta33
GenBank: http://www.ncbi.nlm.nih.gov/Genbank/index.html
PHACCS: http://phage.sdsu.edu/phaccs
Phrap computer program: http://www.phrap.com
Phred computer program: http://www.phrap.com/phred
Sequencher: http://www.genecodes.com/sequencher
Access to this interactive links box is free online.

OUTLOOK

Metagenomics and industrial applications

Patrick Lorenz and Jürgen Eck

Abstract | Different industries have different motivations to probe the enormous resource that is uncultivated microbial diversity. Currently, there is a global political drive to promote white (industrial) biotechnology as a central feature of the sustainable economic future of modern industrialized societies. This requires the development of novel enzymes, processes, products and applications. Metagenomics promises to provide new molecules with diverse functions, but ultimately, expression systems are required for any new enzymes and bioactive molecules to become an economic success. This review highlights industrial efforts and achievements in metagenomics.

Metagenomics[1] has the potential to substantially impact industrial production. The dimensions of the enormous biological and molecular diversity, as shown by Torsvik[2], Venter[3] and their co-workers, are truly astonishing. A pristine soil sample might contain in the order of 10^4 different bacterial species. More than one million novel open reading frames, many of which encode putative enzymes, were identified in a single effort that sampled marine prokaryotic plankton retrieved from the Sargasso Sea.

An industrial perspective

In this perspective, the discussion is limited to prokaryotes, as their genomes are most easily targeted by the functional screening tools available in metagenomics and because it is assumed, based on published literature, that the largest biodiversity occurs in the bacterial lineages[4–6]. Different industries are interested in exploiting the resource of uncultivated microorganisms that has been identified through large-scale environmental genomics for several reasons detailed below.

The ideal biocatalyst. For any industrial application, enzymes need to function sufficiently well according to several application-specific performance parameters (FIG. 1). With the exception of yeasts and filamentous fungi, access to novel enzymes and biocatalysts has largely been limited by the comparatively small number of cultivable bacteria. A corollary of this limitation is, however, that any application has to be designed with enzymatic constraints in mind, leading to suboptimal process and reaction conditions. Instead of designing a process to fit a mediocre enzyme, it is conceivable that the uncultivated microbial diversity, together with *in vitro* evolution technologies, might be used to find a suitable natural enzyme(s) that can serve as a backbone to produce a designer enzyme that optimally fits process requirements that are solely dictated by substrate and product properties[7].

Novelty. For industries that produce bulk commodities such as high-performance detergents, a single enzyme backbone with superior functionality that has an entirely new sequence would be useful to avoid infringing competitors' intellectual property rights. This problem is illustrated by the fact that substitutions at nearly every position in the mature 275 amino acid BPN (bacillus protease Novo type, from *Bacillus amyloliquefaciens*) subtilisin have been claimed in patents[8].

Maximum diversity. The pharmaceutical and supporting fine-chemicals industries often seek entire sets of multiple diverse biocatalysts to build in-house toolboxes for biotransformations[9]. These toolboxes need to be rapidly accessible to meet the strict timelines of a biosynthetic-feasability evaluation in competition with traditional synthetic chemistry.

Elusive metabolites. Many pharmacologically active secondary metabolites are produced by bacteria that live in complex CONSORTIA (see Glossary) or by bacteria that inhabit niches that are difficult to reconstitute *in vitro*[10]. So, although there are reports on how to circumvent this general problem of microbial cultivation either by mimicking natural habitats[11] or by allowing for interspecies communication after single cell micro-encapsulation[12], the cloning and heterologous expression of biosynthetic genes that encode secondary metabolites (usually present as gene clusters) is the most straightforward and reproducible method of accessing their biosynthetic potential.

Industrial enzyme applications

Enzymes are used in a wide range of applications and industries[13]. They are required in only minute quantities to synthesize kilograms of stereochemically challenging chiral SYNTHONS that are used as building blocks to produce highly active pharmaceuticals[14], and at a kiloton/year scale as active ingredients for bulk products such as high-performance