THE ROYAL SOCIETY FirstCite® e-publishing

Received 5 May 2003
Accepted 29 October 2003
Published online

Diversity and population structure of a near-shore marine-sediment viral community

Mya Breitbart[1], Ben Felts[2], Scott Kelley[1], Joseph M. Mahaffy[2], James Nulton[2], Peter Salamon[2] and Forest Rohwer[1,3*]

[1]*Department of Biology,* [2]*Department of Mathematical Sciences, and* [3]*Center for Microbial Sciences, San Diego State University, San Diego, CA 92182-4614, USA*

Viruses, most of which are phage, are extremely abundant in marine sediments, yet almost nothing is known about their identity or diversity. We present the metagenomic analysis of an uncultured near-shore marine-sediment viral community. Three-quarters of the sequences in the sample were not related to anything previously reported. Among the sequences that could be identified, the majority belonged to double-stranded DNA phage. Temperate phage were more common than lytic phage, suggesting that lysogeny may be an important lifestyle for sediment viruses. Comparisons between the sediment sample and previously sequenced seawater viral communities showed that certain phage phylogenetic groups were abundant in all marine viral communities, while other phage groups were under-represented or absent. This 'marineness' suggests that marine phage are derived from a common set of ancestors. Several independent mathematical models, based on the distribution of overlapping shotgun sequence fragments from the library, were used to show that the diversity of the viral community was extremely high, with at least 10^4 viral genotypes per kilogram of sediment and a Shannon index greater than 9 nats. Based on these observations we propose that marine-sediment viral communities are one of the largest unexplored reservoirs of sequence space on the planet.

Keywords: sediment; phage; viruses; diversity; metagenomics

1. INTRODUCTION

Marine sediments constitute one of the largest biospheres in the world (Whitman *et al.* 1998). There are 10^9–10^{13} viruses per kilogram of sediment; most of these viruses are phage that infect and kill prokaryotes (Paul *et al.* 1993; Steward *et al.* 1996; Drake *et al.* 1998; Danovaro & Serresi 2000; Danovaro *et al.* 2001, 2002; Hewson *et al.* 2001; Middelboe *et al.* 2003; Hewson & Fuhrman 2003). Viruses in sediments are rapidly produced and account for a significant percentage of bacterial mortality (Hewson & Fuhrman 2003). The majority of sediment viruses are produced autochthonously, rather than arriving through sedimentation of particles from the water column (Danovaro *et al.* 2002; Hewson & Fuhrman 2003). The abundances of bacteria in the water column and in sediment are not correlated, suggesting independence of microbial processes in these two environments (Paul *et al.* 1993).

Transmission electron microscopy (TEM) studies have shown that viruses in marine sediments have high morphological diversity (Danovaro & Serresi 2000; Middelboe *et al.* 2003). RNA and single-stranded DNA (ssDNA) phages, which are rare in the water column, may have been observed in sediments (Middelboe *et al.* 2003; M. Middelboe, personal communication). Using plaque assays, Paul *et al.* (1993) showed that there are different phage populations in the sediments and in the water column. These methods of measuring phage diversity are limited because of the biases associated with culturing the hosts (Fuhrman & Campbell 1998) and the low taxonomic resolution of viral morphology. Analysing phage diversity directly through nucleic-acid sequencing avoids these limitations.

The Phage Proteomic Tree is a recently proposed taxonomic system based on the complete genomic sequences of phage (Rohwer & Edwards 2002). This system groups phage into taxa that are predictive of many aspects of phage biology. The phylogenetic groups on the Phage Proteomic Tree can be used to determine the relationships of phage based on their DNA sequences, as opposed to the physical characteristics of the phage particles (e.g. the system sponsored by the International Committee on Taxonomy of Viruses (ICTV); Murphy *et al.* 1995). A sequence-based taxonomic system for phage also allows statistical comparisons of viral communities based on phylogeny, as have been carried out for bacteria (Martin 2002).

Previously we analysed uncultured near-shore marine-water-column viral communities using partial shotgun sequencing (Breitbart *et al.* 2002). The types of double-stranded DNA (dsDNA) viruses in the communities were identified and the population structure was modelled based on analyses of the contig spectra (i.e. the distribution of overlapping shotgun sequence fragments; see § 3e). Here we present, to our knowledge, the first metagenomic analysis of the diversity and population structure of a near-shore marine-sediment viral community. Uncultured viral diversity of the marine-sediment sample was extremely high and the viruses were predominantly novel. The sediment community was dominated by phage with the potential for temperate lifestyles. Various population models predicted a very even population with more than 10^4 viral genotypes in a sample of 10^{12} individuals.

* Author for correspondence (forest@sunstroke.sdsu.edu).

Comparisons between the sediment and water-column samples showed that certain phage groups were abundant in all marine communities, while other groups were never observed, suggesting a common phylogenetic origin of marine phage populations.

2. MATERIAL AND METHODS

(a) *Isolation of viral-community DNA*

A sample of *ca*. 1 kg of surface sediment was collected from the channel side of Fiesta Island in Mission Bay, San Diego, CA, USA, in June 2001. The core included the upper 25 cm and consisted of fine particles. This sample, designated MBSED, was harvested at the same time as the previously described overlying water-column sample (MB; Breitbart *et al.* 2002). The sediment sample was immediately transported to the laboratory where the pore water was extracted by centrifuging at *ca*. 8000*g* for 10 min. The sediments were shaken vigorously with autoclaved 100 kD filtered sea water and pelleted and the supernatant was poured off. This process was repeated five times. The pore water and supernatants were combined, poured through a Nitex filter (*ca*. 100 μm pore size) and centrifuged at *ca*. 8000*g* for 10 min to remove larger particles. After centrifugation, the supernatant was filtered through a Whatman GF/C filter and a 0.2 μm Sterivex to remove bacteria. Viruses in the filtrate were concentrated to *ca*. 100 ml using a 100 kD tangential flow filter (Pall Filtron), loaded onto a caesium chloride (CsCl) step gradient and ultracentrifuged, and the 1.35–1.5 g ml^{-1} fraction was collected. This fraction contains the majority of the marine viral particles, as shown using epifluorescent microscopy (method of Noble & Fuhrman 1998; data not shown), and marine viral DNA, as determined by pulse-field gel electrophoresis (Steward *et al.* 2000). After CsCl purification, the viruses were lysed using a formamide extraction and the DNA was recovered by an isopropanol precipitation and a hexadecyltrimethylammonium bromide (CTAB)-based extraction (Sambrook *et al.* 1989).

(b) *Linker-amplified shotgun library*

A linker-amplified shotgun library (LASL) was created from the Mission Bay sediment viral-community DNA. This method circumvents problems associated with modified nucleotides and deadly genes in viral genomes, as well as low DNA content in environmental samples. The LASL was constructed as described previously (Breitbart *et al.* 2002). Briefly, the total sediment viral-community DNA was randomly sheared (HydroShear, GenMachine, San Carlos, CA, USA) and end-repaired, and dsDNA linkers were ligated to the ends. The fragments were then amplified using the high-fidelity Vent DNA polymerase, ligated into the pSMART vector and electroporated into MC12 cells (Lucigen, Middleton, WI, USA). This method has been checked to ensure randomness and lack of chimeric sequences as described earlier (Rohwer *et al*. 2001; www.sci.sdsu.edu/PHAGE/LASL/index.htm).

(c) *Analysis of sequences: composition analyses*

A total of 1156 clones from the Mission Bay sediment library were sequenced with the AmpL2 forward primer (Lucigen, Middleton, WI, USA). These sequences were compared against GenBank using TBLASTX (Altschul *et al.* 1990, 1997). A hit was considered significant if it had an *E*-value of less than 0.001. Sequences with significant hits in GenBank were classified as phage, viruses, mobile elements, repeat elements, bacteria, archaea or eukarya. Transposons, plasmids, insertion sequences, retrotransposons, instable genetic elements and pathogenicity islands were grouped together as mobile elements. Bacterial, archaeal and eukaryotic hits were examined manually to identify repeat elements and potential prophage hits according to a list provided by Sherwood Casjens (University of Utah; similar to the list in Casjens (2003)).

(d) *Analysis of sequences: statistical comparisons of communities*

Phylogenetic methods adapted by Martin (2002) were used to determine whether there were differences in the phylogenetic compositions of the marine-sediment uncultured viral community (MBSED) and the two previously described marine-water-column viral communities—Mission Bay (MB) and Scripps Pier (SP) (Breitbart *et al.* 2002). Since the viral sequences were not from a single genetic locus, it was not possible to perform the multiple sequence alignments needed for standard phylogenetic analyses. To circumvent this limitation, we determined the phylogenetic relationships by mapping the shotgun-cloned phage sequences to known phage genomes whose phylogenetic relationships had been previously determined on the Phage Proteomic Tree (Rohwer & Edwards 2002). For this approach, sequences with a significant BLAST hit (*E*-value of less than 0.001) to a genome on the Phage Proteomic Tree were considered to be closely related to that particular phage. If none of the significant phage hits for a given sequence were present on the Phage Proteomic Tree, the sequence was not used. The sequences were assigned character states based on the environment from which they were obtained (i.e. their community association—MB, MBSED or SP). If sequences from two communities were most similar to the same genome on the tree, these were considered as two closely related phage sequences instead of as one multistate character. The environmental phage sequences were mapped to the positions of their closest relatives on the Phage Proteomic Tree using MACCLADE (Maddison & Maddison 2002). PHYLOGENETIC ANALYSIS USING PARSIMONY (PAUP) was then used to perform the permutation tail probability (PTP) test, where the character states were randomized 10 000 times while holding the tree topology constant (Maddison & Slatkin 1991; Swofford 2000). The parsimony criteria were used to determine the number of evolutionary changes (i.e. steps) between community types. If sequences from one environment are more closely related to each other than to those from another environment, then the observed number of changes as measured by steps on the tree would be small. The PTP test yielded a frequency distribution of the number of steps of the randomly permuted character against which the observed number of steps was compared. Viral phylogeny was significantly correlated with the environment if the observed number of steps was fewer than expected given a random distribution of character states. The PTP tests were also performed using randomized trees and the results agreed with those obtained using the randomized character states.

(e) *Analysis of sequences: contig spectrum*

All sequences were analysed using SEQUENCHER 4.0 (Gene Codes, Ann Arbor, MI, USA) to identify contigs (i.e. contiguous or overlapping sequences) based on a minimum overlap of 20 bp with 98% minimal match percentage (MM%). These assembly parameters were sufficient to differentiate between very closely related phage such as *E. coli* ϕ T7 and ϕ T3 (F. Rohwer and Y. Yu, unpublished results, and protocol presented in Breitbart *et*

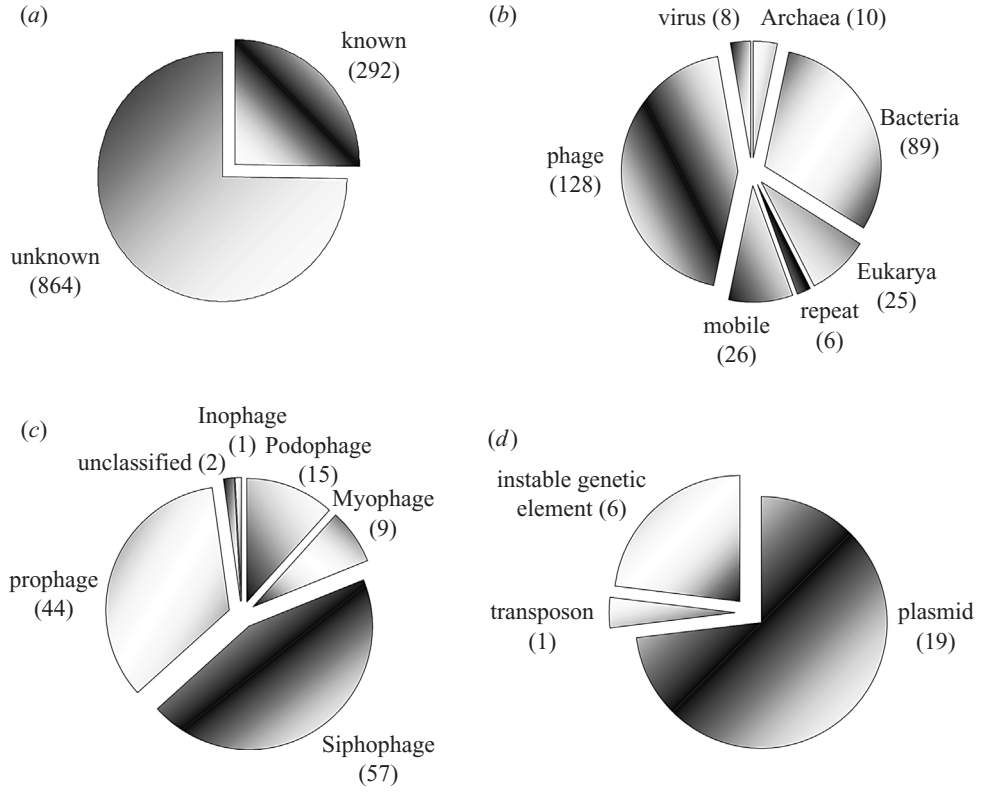

Figure 1. Genomic overview of the uncultured sediment viral community from Mission Bay, CA, USA, based on sequence similarities. (*a*) Number of sequences with a significant hit (*E*-value < 0.001) to GenBank. (*b*) Distribution of significant hits among the major classes of biological entity. (*c*) Families of phage represented in the sediment library. (*d*) Types of mobile element identified in the library.

al. (2002)). Based on the distribution of overlapping sequences, the population structure of the sediment viral community was determined using the contig-spectrum model described previously (Breitbart *et al.* 2002), as well as with a refinement of this model (model 2; details presented in § 3e). These two contig-spectrum models were supplemented with a Monte Carlo simulation, described in § 3e, to predict the population structure of the viral community.

3. RESULTS

(a) *Composition of uncultured sediment viral community*

The majority (75%) of the 1156 fragments obtained from the MBSED LASL (GenBank accession numbers CC821301–CC822456) were not significantly similar (*E*-values of greater than 0.001) to anything in GenBank, demonstrating that most sediment viral diversity remains uncharacterized (figure 1*a*). Sequences with significant hits to GenBank entries were classified into broad biological groups based on GenBank annotation. Phage were the most common known hits (44%), while viruses that infect eukaryotes represented only 3% of the significant hits (figure 1*b*). Sequences with significant similarity to Podo-, Myo- and Siphophage were observed in the sediment library, with Siphophage being the most common (45%; figure 1*c*). The most common phage hits were to *Pseudomonas aeruginosa* ϕ D3 and *Burkholderia thailandensis* ϕ E125, both of which are λ-like Siphophage. Prophage and phage-like proteins identified within bacterial genomes accounted for 34% of the significant phage hits. Together, prophage and Siphophage accounted for 79% of the phage hits, suggesting that the MBSED community is dominated by phage with the potential for temperate lifestyles. In total, 66% of the significant phage hits were similar to genes of known function, including DNA and RNA polymerases, helicases, exonucleases, integrases, terminases and a variety of structural proteins (table 1). Most of the mobile-element sequences from the Mission Bay sediment library were similar to bacterial plasmids (73%; figure 1*d*).

(b) *Compositional comparison of uncultured marine viral communities from sediment and water column*

The types of viruses in the Mission Bay sediment community were compared with those of two previously characterized marine-water-column viral communities (MB and SP; Breitbart *et al.* 2002). The majority of sequences from all three marine viral communities showed no significant similarity to previously reported sequences in GenBank. Interestingly, TBLASTX comparisons of the three viral communities showed that many of the unknown sequences found in each sample were related to each other (*E*-value < 0.001).

The MBSED and MB samples were collected at the same time and location, allowing comparison between the uncultured viral communities from a sediment sample and the overlying water column. Several basic differences were observed between the two communities. The MB water-column sample appeared to be more eukaryotic in origin, with 43% of the significant hits to eukaryotes, their viruses and repeat sequences, which are common within eukaryotic

Table 1. Categories of phage proteins observed among significant phage hits in the Mission Bay sediment viral community. (Overall, 66% of the phage hits were similar to previously described proteins, with the majority of these hits being most similar to terminases or structural proteins.)

protein category	number observed in MBSED LASL
unknown	43
structural	26
terminase	18
DNA polymerase	5
helicase	5
exonuclease	4
methylase	3
endolysin	2
endonuclease	2
integrase	2
lytic enzyme	2
prophage regulation	2
ribonucleotide reductase	2
RNA polymerase	2
DNA end-binding	1
holin	1
host-interacting	1
host specificity	1
methyltransferase	1
packaging	1
primase	1
protease	1
superinfection exclusion	1
transposase	1
total	128

genomes (Breitbart *et al.* 2002). The sediment community from the same site did not share this feature (eukaryotes, eukaryotic viruses and repeats accounted for only 13% of the significant hits; figure 1*b*). Mobile elements observed in the communities also differed in composition. More than half of the mobile elements observed in the water-column community were similar to retrotransposons, while not a single retrotransposon was observed in the sediment community. Overall, the viral communities from the sediment and the overlying water column appeared to be very different in composition.

(c) *Prevalence of temperate phage in the marine-sediment viral community*

The MBSED viral community consisted mostly of Siphophage and prophage (79% of the significant hits). This is significant because these two groups are potentially temperate. These groups were much less abundant in the marine-water-column communities (44% in SP; 52% in MB). This trend of more temperate (Siphophage + prophage) phage in the sediment than in the marine water column was confirmed statistically using a *G*-test of independence (MBSED versus MB: $\chi^2 = 12.5$, d.f. = 2, $p < 0.005$, and MBSED versus SP: $\chi^2 = 29.5$, d.f. = 2, $p < 0.001$). There was no significant difference, however, between the abundances of lytic and temperate phage groups in the two seawater samples (SP versus MB: $\chi^2 = 2.0$, d.f. = 2, $p > 0.05$). These results suggest that production of temperate phage is more prevalent in the marine-sediment community. The short time-scale of processing and the fact that the water and sediment samples were treated similarly make it unlikely that prophage were induced during processing.

(d) *Phylogenetic comparisons of the sediment and water-column marine viral communities*

Figure 2 shows that certain phage groups were more abundant in the marine viral communities than others. The two major subgroups of Podophage are a good example of this phenomenon. T7-like Podophage were common in all three marine communities, while the PZA-like Podophage were observed only a few times. Similarly, while most groups of Siphophage were rare in the marine environment, the λ-like Siphophage were abundant. A χ^2-test showed that the distribution of top phage hits observed in each of the marine viral communities was not similar to that expected from a random sampling of the groups present on the Phage Proteomic Tree ($p < 0.001$).

The PTP test was then used to determine whether there were statistical differences in taxonomical composition between phage samples. When the different marine viral communities were compared using PTP, no significant differences were found (figure 3*a–c*). By contrast, when the phage from all three marine communities were combined and compared with all of the completely sequenced dsDNA phage on the Phage Proteomic Tree, there was a significant phylogenetic difference (figure 3*d*, $p < 0.05$). This shows that marine viral communities have a shared phylogenetic quality.

(e) *Monte Carlo and analytical modelling of the marine-sediment viral population*

The sample used to make the sediment shotgun library contained *ca.* 10^{12} individual viral particles. This is very similar to the number of viral particles in each of the two water-column samples (Breitbart *et al.* 2002). The higher the diversity of a viral community, the lower the chances of sequencing overlapping fragments from the same viral genome.

For convenience in discussing the distribution of overlapping fragments, we have introduced the notion of the contig spectrum of a sample $[c_1, c_2, \ldots]$, where c_1 is the number of sequences that do not overlap with any other sequences (1-contigs), c_2 is the number of pairs of contiguous sequences (2-contigs), etc. Using this notation, the contig spectrum of the MBSED sample was [1152, 2, 0, …], in contrast to the contig spectra of the water columns, which were [1021, 17, 2, 0, …] and [841, 13, 2, 0, …] for SP and MB, respectively (Breitbart *et al.* 2002). The MBSED library had many fewer overlapping fragments than the MB and SP libraries, which suggests that viral diversity is higher in the sediment than in the sea water.

Previously we have described a methodology for modelling viral populations based on contig spectra obtained from partial shotgun sequencing. The original contig-spectrum model, henceforth called model 1, extends the Lander–Waterman equation to describe entire viral populations (Lander & Waterman 1988; Breitbart *et al.* 2002). Model 1 predicts a frequency n_i for the ith most frequent viral genome. In the original derivation of model 1, whole numbers for n_i were replaced by their expected values, which were not integers. Model 1 yields good

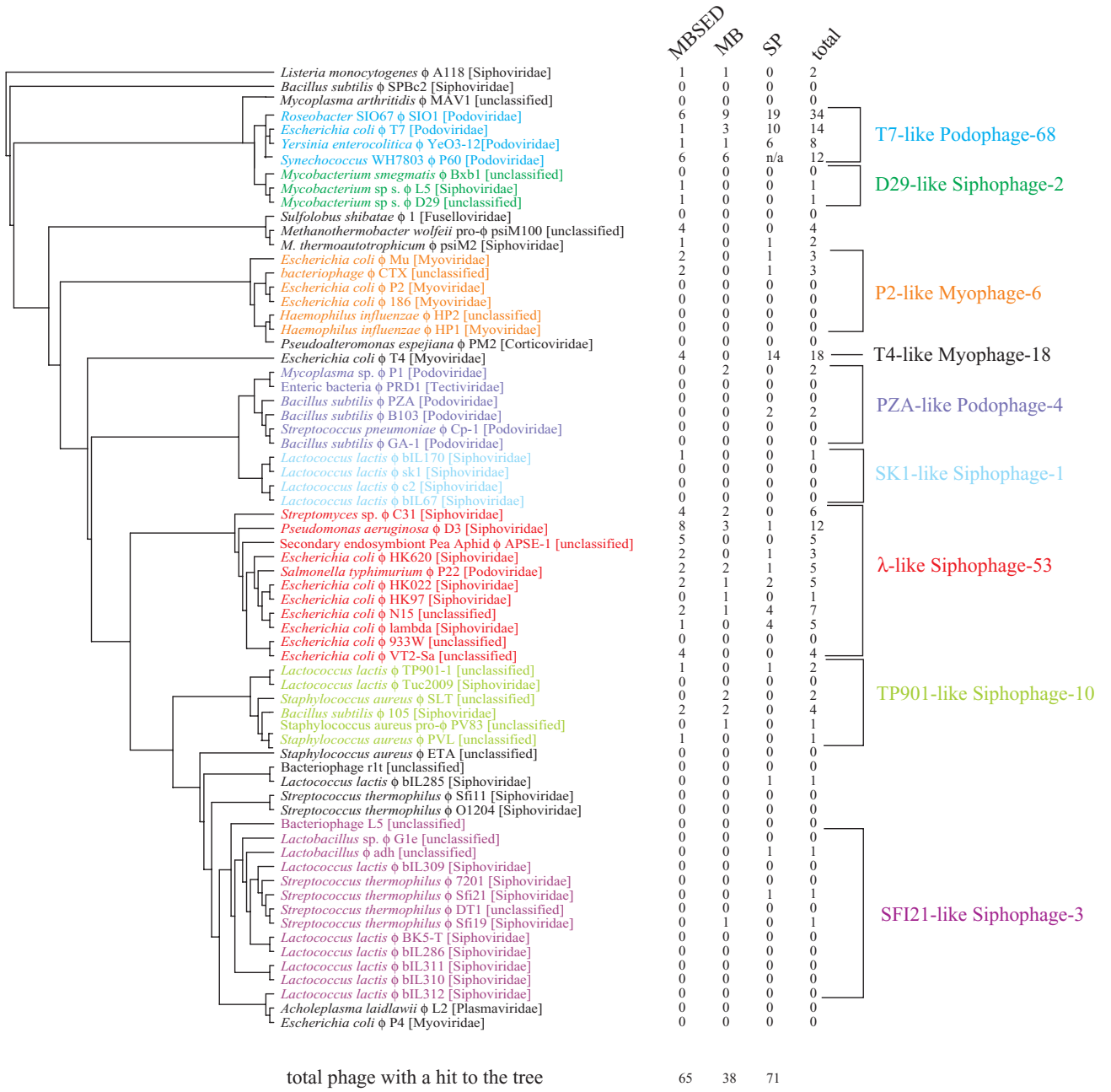

Figure 2. Phage Proteomic Tree showing the number of top significant hits in the uncultured marine phage libraries to each of the completely sequenced dsDNA phage genomes. Certain phage genomic groups (e.g. T7-like Podophage, λ-like Siphophage and T4-like Myophage) were more prevalent in the marine environment.

approximations provided that several of the n_i values are greater than one. For the MBSED sample, however, the expected frequencies of even the most abundant viral genotypes were less than one. This raised concern about the applicability of model 1 for predicting the population structure of a community with so few contigs. Therefore, a Monte Carlo simulation (described in more detail below) and a more complicated, albeit more exact, analytical method, called model 2, were developed and compared with model 1. As discussed below, these methods confirmed that the predictions of model 1 were approximately correct (table 2).

A power-law function was used to describe the rank–abundance relationship in the population. The power law is one of the basic functional forms observed in biological populations (Ulrich 2001) and we found that this function best described the MB and SP marine viral communities (Breitbart *et al.* 2002). Based on this information, a normalized power-law function ($f_i = ai^{-b}$, $M \geqslant i \geqslant 1$, $\Sigma_{i=1}^{M} f_i = 1$) was used for the relative frequencies f_i, making the expected n_i equal to the sample size times f_i. This distribution used two independent parameters of the sediment viral community: (i) M, the number of viral types in the sample, and (ii) a, the relative frequency of the most abundant virus.[1] The evenness parameter b, which controls the rate at which f_i decreases, was then determined from the normalization condition $\Sigma_{i=1}^{M} f_i = 1$. The maximum-likelihood values for these parameters predicted by model 1 are $a = 0.00012$, $b = 0.00$ and $M = 8600$ (table 2). This represents a completely even population in which all viral types have relative frequencies of 0.012%. Model 2 gave very similar values

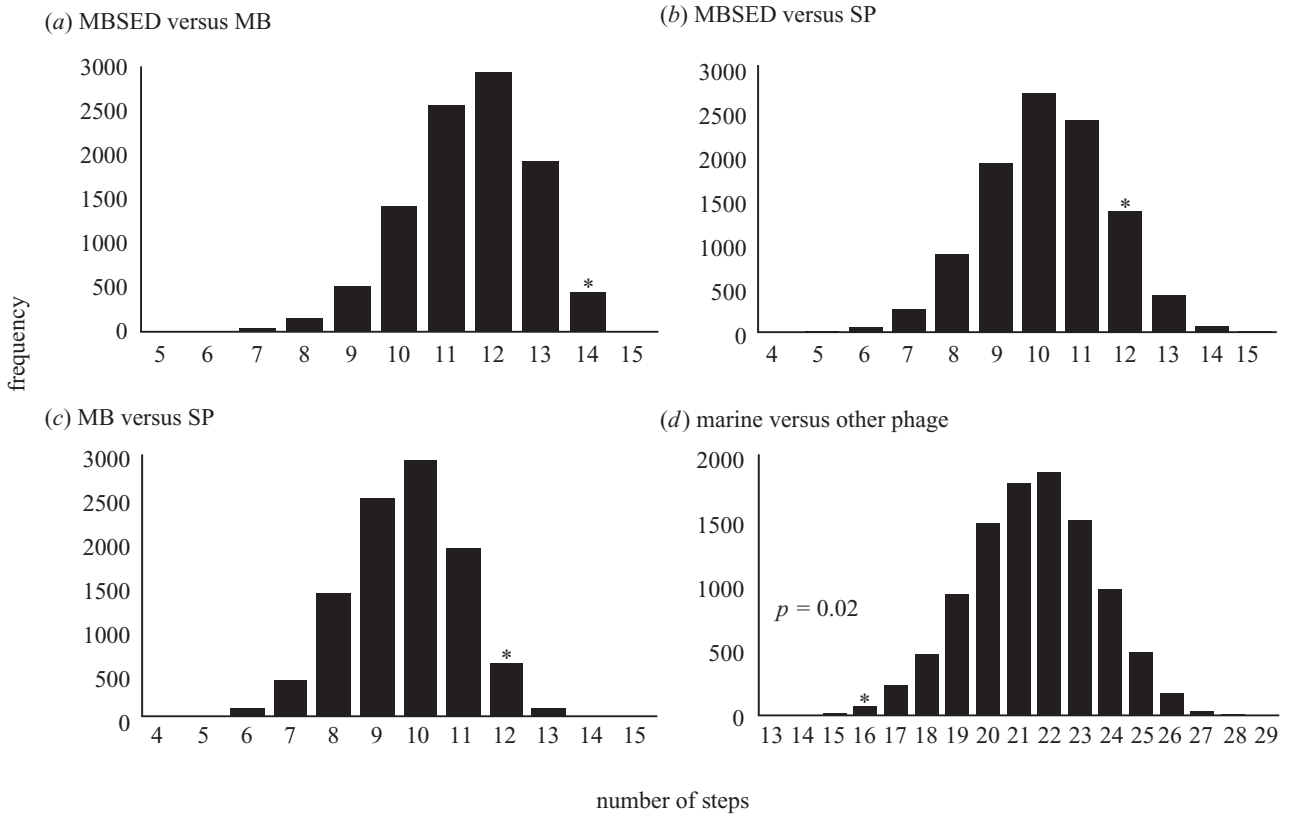

Figure 3. Phylogenetic statistical comparisons of viral sequences from different communities using the PTP test. The asterisks represent the actual number of steps observed when the two communities were compared. This value was compared with the histogram produced by randomizing the character states 10 000 times on a fixed tree (i.e. a random model). If the observed number of steps between communities was significantly less than expected using a random model, then there was a significant association between environment and phage phylogeny. (*a–c*) In pairwise comparisons between the MBSED, MB and SP viral communities, the observed tree lengths were no shorter than expected using a random model, revealing complete phylogenetic overlap between the marine water-column and sediment viral communities. (*d*) There is a significant phylogenetic difference between the groups of phage found in the marine communities and the other phage sequences available on the Phage Proteomic Tree (i.e. the marine phage represent only a subset of the sequences on the tree; $p < 0.05$).

Table 2. Maximum-likelihood values of the parameters obtained through contig-spectra modelling of the three marine viral populations.
(The parameters for all three populations are based upon a rank–abundance curve described by a power-law function. The errors were computed using the inverse of the second derivative of the negative log likelihood. For MBSED, models 1 and 2 had maximum-likelihood values indicating completely even populations; therefore, no second derivative was available at these endpoint optima and thus no error values are shown. Please note that *a* is expressed as a percentage.)

	relative percentage of the most abundant virus *a* (×100)	evenness *b*	richness *M*	Shannon index *H* (nats)
Monte Carlo				
MBSED	0.1 ± 0.4	0.28 ± 0.45	10000 ± 6400	9.2
MB	2.5 ± 0.5	0.70 ± 0.05	5100 ± 2100	7.8
SP	1.9 ± 0.5	0.61 ± 0.06	2600 ± 800	7.4
model 1				
MBSED	0.012	0	8600	9.0
MB	2.7 ± 5.5	0.73 ± 0.11	7000 ± 12000	8.0
SP	2.0 ± 4.5	0.64 ± 0.98	3300 ± 3000	7.6
model 2				
MBSED	0.012	0	8500	9.0
MB	2.9 ± 1.4	0.76 ± 0.25	9000 ± 37000	8.1
SP	2.2 ± 1.1	0.66 ± 0.19	3500 ± 6300	7.6

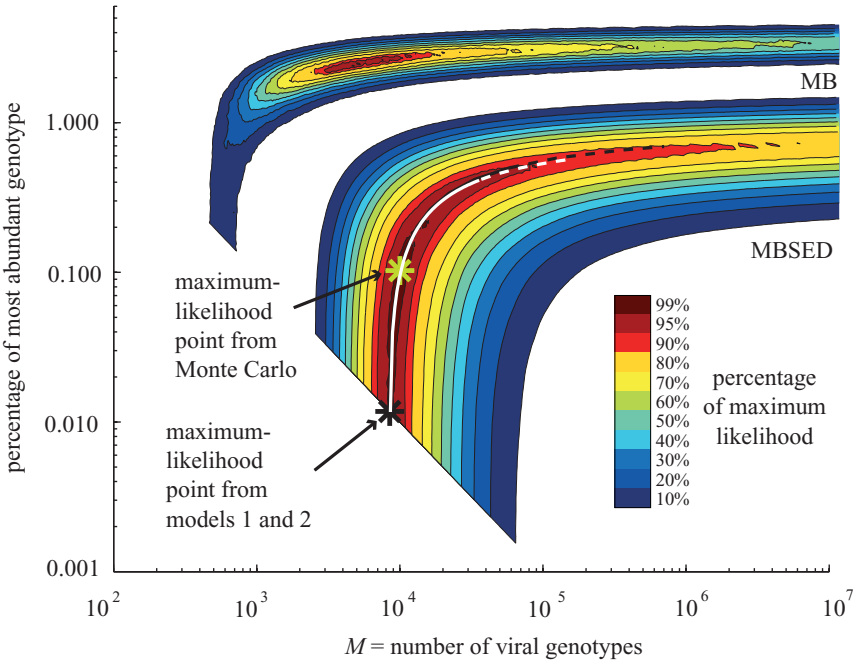

Figure 4. Contour plots of the likelihood functions for MB and MBSED as percentages of the maximum likelihood observed in direct Monte Carlo simulations. The Monte Carlo simulations count the number of exact matches to the observed contig spectrum as a function of the number of viral genotypes M and the percentage abundance of the most abundant viral genotype a. For the MBSED sample, the maximum-likelihood points are shown (Monte Carlo: yellow asterisk; models 1 and 2: black asterisk; the values for the two models are the same). The ridge line identifying the maximum-likelihood a for each M as predicted from model 1 (white curve) and model 2 (black curve) are also shown. The range of M-values shown for these two curves corresponds to model predictions above 90% of the maximum likelihood. The solid portions of the lines show the 95–100% range while the dashed portions show the 90–95% range.

with $a = 0.00012$, $b = 0.00$ and $M = 8500$. Since such an even population was predicted, the results are relatively independent of the initial assumption about the shape of the rank–abundance curve.

To verify the predictions from models 1 and 2, a Monte Carlo simulation was also performed. In figure 4, the maximum-likelihood parameter values for the Monte Carlo simulation are shown as a yellow asterisk and the maximum-likelihood values for models 1 and 2 are indicated with a black asterisk (these values are virtually the same). These values are also compared in table 2. For the Monte Carlo simulation, 1156 fragments were selected from a mixture of M different genomes with relative frequencies f_i defined according to a power-law distribution, and the resulting contig spectrum was determined. This simulation was performed 7.5 billion times, 150 000 times for each pair of M and a values, covering a complete 500×100 grid in the region shown in figure 4. These computations, performed in MatLab, required more than 250 h on a 1 GHz Pentium 4 processor. For each pair of M and a values, the number of times that the simulated contig spectrum exactly matched the observed spectrum [1152, 2, 0, …] was counted. These counts provided an empirical likelihood function, whose contours are plotted in figure 4 with the counts rescaled as fractions of the maximum value. Note that there is a long ridge of approximately equal maximum-likelihood (M, a) values. In the spirit of equifinality, one should consider all those cases with nearly equal likelihood (Beven 1993). Within the 90% of maximum-likelihood contour, the total number of viral genotypes M in the sample (richness) varies between *ca.* 10^4 and 10^7 (figure 4). In addition, along the maximum-likelihood ridge within the 90% contour, the population ranges from a completely even population ($b = 0$, where the most abundant viral type is only 0.01%) to a less even population ($b = 0.74$, where the most abundant viral type is 0.7%).

Figure 4 also shows curves of optimal values from Models 1 and 2 for a range of fixed M values. Model 1 predicts a curve (shown in white) that closely follows the ridge of the Monte Carlo results. Model 2 is an improvement on model 1 in several respects, including the incorporation of a covariance matrix. Model 2 predicts a curve (shown in black) that is similar to that predicted by model 1, but which matches more closely the contour lines of the Monte Carlo simulation. Both models 1 and 2 give maximum (unconstrained M) likelihoods for the sediment viral community at completely even populations, as represented by the black asterisk with a richness value approaching 10 000.

The Monte Carlo simulation predictions for the population structure of the MB water-column sample are also shown in figure 4. The SP population is not shown because the SP and MB viral communities almost completely overlap when plotted in this manner. There is no overlap between the high-likelihood regions of the Mission Bay sediment and water-column viral populations.

(f) *Shannon index for marine viral communities*

Figure 5 shows the high-likelihood (greater than 90%) regions predicted by the Monte Carlo simulation overlaid on a contour plot of the Shannon index (H, in nats) for a population with a power-law distribution. These regions are shown for the sediment and the water-column

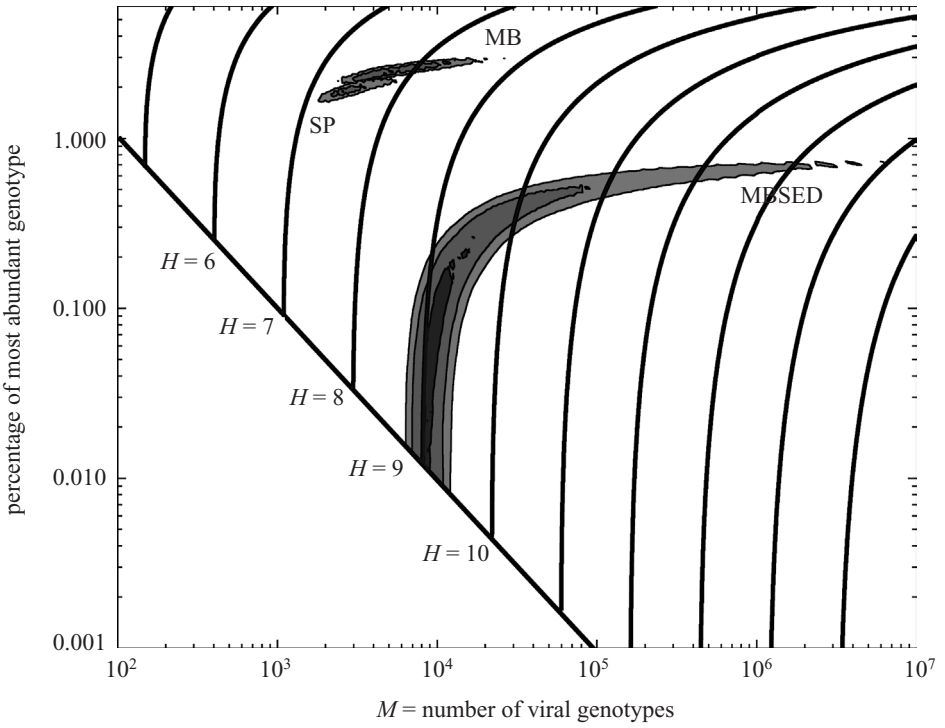

Figure 5. Isodiversity contours of the Shannon index (H, expressed in nats) overlaid on a plot of the regions having more than 90% of the maximum likelihood for the MBSED, MB and SP samples. The dark-grey region represents more than 99% of the maximum likelihood. The lighter regions are successively the 95–99% and 90–95% maximum-likelihood regions from the Monte Carlo simulations.

samples, revealing clear differences between the populations. Although the evenness (b) and richness (M) of the MBSED viral community vary along the ridge, within the 99% contour the diversity of the community, as described by the Shannon index, is consistently between 9 and 9.5 nats (figure 5; table 2). The ridge passes through higher Shannon indices beyond the 99% contour.

4. DISCUSSION

The majority of the sequences in the uncultured marine-sediment library were unrelated to anything in GenBank, which is consistent with the idea that most viral diversity is currently unsampled (Breitbart *et al.* 2002, 2003; Pedulla *et al.* 2003; Rohwer 2003). As we found in the previously analysed marine-water-column samples (Breitbart *et al.* 2002), viruses were the most common significant hits to GenBank sequences (figure 1). Among the viral hits, fragments most closely related to Siphophage and prophage dominated the sediment sample (79%). These two phage groups include most of the classically temperate phage (i.e. phage with the ability to integrate into the bacterial host's chromosome). Siphophage and prophage also dominated (81%) a recently analysed uncultured viral community from human faeces (Breitbart *et al.* 2003). By contrast, Siphophage and prophage comprised less than 50% of the total hits in the two marine-water-column samples that have been analysed (Breitbart *et al.* 2002). This observation suggests that production of temperate phage may be more prevalent in the environments with higher microbial densities. The higher occurrence of temperate phage in the sediments may be explained in several ways. Lysogeny is thought to be more important for survival under conditions where host densities or resources are low (Wilson *et al.* 1998; McDaniel *et al.* 2002). High nutrient availability and bacterial densities in sediments may encourage phage with the ability to be temperate to reproduce in a purely lytic manner, explaining their presence in the free-phage fraction. Alternatively, lysogens in the sediment sample may be experiencing stresses that lead to induction. Since the phage in this environment are less exposed to UV light, induction may be triggered by other signals such as quorum factors expressed as a result of the high bacterial densities.

Even though phage with the potential to be temperate were more common in the sediment community, statistical analyses (PTP; Martin 2002) demonstrated that there was almost complete phylogenetic overlap between the marine water-column and sediment viral communities (figure 3). There was, however, a significant phylogenetic difference between the groups of phage that were found in the marine communities and the phage genomes available in GenBank. The phylogenetic composition of the sediment and water-column viral communities also differed from that of the faecal sample (Breitbart *et al.* 2003). Most notably, phage known to infect Gram-positive bacteria (e.g. SFI21- and TP901-like Siphophage) were common in the faecal sample (Breitbart *et al.* 2003), but almost absent from the marine viral communities (figure 2). In addition, the faecal community had very few hits to the groups of phage that dominated the marine samples. This means that marine phage represented only a subset of the total known phage, most notably the T7-like Podophage, λ-like Siphophage and T4-like Myophage. Further studies will be necessary to determine why certain phylogenetic groups are more successful in the marine environment.

The population structure (i.e. total number of viral genotypes, abundance of the most abundant genotype, evenness of the population) of the sediment community was very different from those of the water-column and faecal samples. Owing to the small number of contigs in the sediment library, the applicability of some of the assumptions of model 1 was called into question (Breitbart *et al.* 2002). Model 2 and the Monte Carlo simulation (described in § 3e) were used to verify the predictions of model 1. All three population models predicted an extremely even population with approximately 10^4 viral genotypes in the sediment viral community (asterisks in figure 4; table 2). There were no dominant viral genotypes in the sediment community, and, in the maximum-likelihood predictions, the most abundant virus comprised only 0.01–0.1% of the total population. Model 2 and the Monte Carlo simulation were also used to reanalyse the population structure of the water-column samples. The water-column viral populations predicted by all three models were very similar (table 2). The viral population structures of the two marine-water-column samples were nearly identical. However, there was absolutely no overlap between the predicted viral population in the sediment and those of the two water-column samples, indicating that environment determines how individual viruses are distributed within a sample.

The diversity of the marine-sediment viral community is extremely high, with the most likely Shannon index (H) being greater than 9 nats and possibly as high as 14 nats (figure 5). This Shannon index is the highest value currently reported in the literature and is significantly higher than the values obtained from marine-water-column ($H = 7.4$–8.1; table 2) and faecal viral communities ($H = 6.4$; Breitbart *et al.* 2003). Assuming an absolutely uniform distribution ($b = 0$), which gives the highest possible Shannon index for a given M, there needs to be more than 8000 species in a population to obtain a Shannon index of 9 nats. Currently, there are only 4200 species of amphibians and 6300 species of reptiles known to science (Wilson 1999). Therefore, as measured by the Shannon index, all the amphibians or reptiles known on the planet are less diverse than the viruses in only 1 kg of near-shore marine surface sediment.

The protocols we used to isolate and clone the viruses will underestimate the viral diversity of the sediment community. No chemical detergents or sonication were used in the viral-isolation protocol because of concerns that these treatments would selectively remove certain viral types (e.g. lipid-containing phage such as *E. coli* ϕ PRD1). It may be possible, for example, to fix the sample, sonicate (e.g. Danovaro *et al.* 2001) and then reverse the fixation before cloning. However, the problems with efficiently reversing the fixation and the effects of the fixation on the CsCl purification made this a much less attractive approach for the initial analyses presented here. At this juncture it is not known whether various phage types differentially attach to sediment particles. If this does occur, certain phage types may have been lost owing to preferential adhesion. In addition, all ssDNA and RNA viruses will be lost during our cloning step. Middelboe *et al.* (2003) recently observed these viral types in sediment communities. These observations show that we have underestimated sediment viral diversity and our numbers should be considered as lower bounds. Taken together, these results strongly suggest that viruses in marine sediments are one of the most diverse biological groups on the planet.

The authors thank David Bangor, Jed Fuhrman, Ian Hewson and Beltran Rodriguez Brito for helpful suggestions. Mya Breitbart was funded by the National Center for Environmental Research (NCER) STAR Program, Environmental Protection Agency. This project was supported by NSF DEB03-16518 and NSF DEB-BE0221763.

ENDNOTE

[1]Note that this usage differs from that in Breitbart *et al.* (2002) where *a* was used to denote the frequency rather than the relative frequency.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Beven, K. 1993 Prophecy, reality and uncertainty in distributed hydrological modeling. *Adv. Water Resources* **16**, 41–51.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. 2002 Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99**, 14 250–14 255.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J., Nulton, J., Salamon, P. & Rohwer, F. 2003 Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **85**, 6220–6223.

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

Danovaro, R. & Serresi, M. 2000 Viral density and virus-to-bacterium ratio in deep-sea sediments of the eastern Mediterranean. *Appl. Environ. Microbiol.* **66**, 1857–1861.

Danovaro, R., Dell'Anno, A., Trucco, A., Serresi, M. & Vanucci, S. 2001 Determination of virus abundance in marine sediments. *Appl. Environ. Microbiol.* **67**, 1384–1387.

Danovaro, R., Manini, E. & Dell'Anno, A. 2002 Higher abundance of bacteria than of viruses in deep Mediterranean sediments. *Appl. Environ. Microbiol.* **68**, 1468–1472.

Drake, L. A., Choi, K. H., Haskell, A. G. E. & Dobbs, F. C. 1998 Vertical profiles of virus-like particles and bacteria in the water column and sediments of Chesapeake Bay, USA. *Aquat. Microb. Ecol.* **16**, 17–25.

Fuhrman, J. A. & Campbell, L. 1998 Microbial microdiversity. *Nature* **393**, 410–411.

Hewson, I. & Fuhrman, J. A. 2003 Viriobenthos production and virioplankton sorptive scavenging by suspended sediment particles in coastal and pelagic waters. *Microb. Ecol.* **46**, 337–347.

Hewson, I., O'Neil, J. M., Fuhrman, J. A. & Dennison, W. C. 2001 Virus-like particle distribution and abundance in sediments and overlying waters along eutrophication gradients in two subtropical estuaries. *Limnol. Oceanogr.* **46**, 1734–1746.

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

McDaniel, L., Houchin, L. A., Williamson, S. J. & Paul, J. H. 2002 Lysogeny in marine *Synechococcus*. *Nature* **415**, 496.

Maddison, D. R. & Maddison, W. P. 2002 *MacClade*. Sunderland, MA: Sinauer.

Maddison, W. P. & Slatkin, M. 1991 Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution* **45**, 1184–1197.

Martin, A. 2002 Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* **68**, 3673–3682.

Middelboe, M., Glud, R. N. & Finster, K. 2003 Distribution of viruses and bacteria in relation to diagenetic activity in an estuarine sediment. *Limnol. Oceanogr.* **48**, 1447–1456.

Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A. & Summers, M. D. (eds) 1995 *Virus taxonomy: sixth report of the international committee on taxonomy of viruses*. New York: Springer.

Noble, R. T. & Fuhrman, J. A. 1998 Use of SYBR green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb. Ecol.* **14**, 113–118.

Paul, J. H., Rose, J. B., Jiang, S. C., Kellogg, C. A. & Dickson, L. 1993 Distribution of viral abundance in the reef environment of Key Largo, Florida. *Appl. Environ. Microbiol.* **59**, 718–724.

Pedulla, M. L. (and 20 others) 2003 Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171–182.

Rohwer, F. 2003 Global phage diversity. *Cell* **113**, 141.

Rohwer, F. & Edwards, R. 2002 The phage proteomic tree: a genome based taxonomy for phage. *J. Bacteriol.* **184**, 4529–4535.

Rohwer, F., Seguritan, V., Choi, D. H., Segall, A. M. & Azam, F. 2001 Production of shotgun libraries using random amplification. *BioTechniques* **31**, 108–118.

Sambrook, J., Fritsch, E. F. & Maniatis, T. 1989 *Molecular cloning: a laboratory manual*. New York: Cold Spring Harbor Laboratory Press.

Steward, G. F., Smith, D. C. & Azam, F. 1996 Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. *Mar. Ecol. Prog. Ser.* **131**, 287–300.

Steward, G. F., Montiel, J. & Azam, F. 2000 Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol. Oceanogr.* **45**, 1697–1706.

Swofford, D. L. 2000 *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*. Sunderland, MA: Sinauer.

Ulrich, W. 2001 Models of relative abundance distributions I: model fitting by stochastic models. *Polish J. Ecol.* **49**, 145–157.

Whitman, W., Coleman, D. & Wiebe, W. 1998 Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583.

Wilson, E. O. 1999 *The diversity of life*. New York: Norton.

Wilson, W. H., Turner, S. & Mann, N. H. 1998 Population dynamics of phytoplankton and viruses in a phosphate-limited mesocosm and their effect on DMSP and DMS production. *Estuarine Coastal Shelf Sci.* **46**(Suppl. A), 49–59.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.