LETTERS

Biodiversity and biogeography of phages in modern stromatolites and thrombolites

Christelle Desnues[1], Beltran Rodriguez-Brito[1,2], Steve Rayhawk[1,2], Scott Kelley[1,3], Tuong Tran[1], Matthew Haynes[1], Hong Liu[1], Mike Furlan[1], Linda Wegley[1], Betty Chau[1], Yijun Ruan[4], Dana Hall[1], Florent E. Angly[1], Robert A. Edwards[1,2,3,5], Linlin Li[1], Rebecca Vega Thurber[1], R. Pamela Reid[6], Janet Siefert[7], Valeria Souza[8], David L. Valentine[9], Brandon K. Swan[9], Mya Breitbart[10] & Forest Rohwer[1,3]

Viruses, and more particularly phages (viruses that infect bacteria), represent one of the most abundant living entities in aquatic and terrestrial environments. The biogeography of phages has only recently been investigated and so far reveals a cosmopolitan distribution of phage genetic material (or genotypes)[1–4]. Here we address this cosmopolitan distribution through the analysis of phage communities in modern microbialites, the living representatives of one of the most ancient life forms on Earth. On the basis of a comparative metagenomic analysis of viral communities associated with marine (Highborne Cay, Bahamas) and freshwater (Pozas Azules II and Rio Mesquites, Mexico) microbialites, we show that some phage genotypes are geographically restricted. The high percentage of unknown sequences recovered from the three metagenomes (>97%), the low percentage similarities with sequences from other environmental viral ($n = 42$) and microbial ($n = 36$) metagenomes, and the absence of viral genotypes shared among microbialites indicate that viruses are genetically unique in these environments. Identifiable sequences in the Highborne Cay metagenome were dominated by single-stranded DNA microphages that were not detected in any other samples examined, including sea water, fresh water, sediment, terrestrial, extreme, metazoan-associated and marine microbial mats. Finally, a marine signature was present in the phage community of the Pozas Azules II microbialites, even though this environment has not been in contact with the ocean for tens of millions of years. Taken together, these results prove that viruses in modern microbialites display biogeographical variability and suggest that they may be derived from an ancient community.

Microbialites are organosedimentary structures accreted by sediment trapping, binding and *in situ* precipitation due to the growth and metabolic activities of microorganisms[5]. Stromatolites and thrombolites are morphological types of microbialites classified by their internal mesostructure: layered and clotted, respectively[5]. Microbialites first appeared in the geological record ~3.5 billion years ago, and for more than 2 billion years they are the main evidence of life on Earth[6,7]. Whether modern microbialites are proxies of ancient ecosystems is a major outstanding question[6].

Viruses, and more specifically phages, are the most abundant biological entities in the world's oceans[8]. Phages influence microbial growth rates, genetic exchange, diversity and adaptation, and thus evolution[8]. Current biogeographical studies of phages suggest that they are cosmopolitan in distribution, unlike some examples of highly endemic populations of bacteria and archaea[9–12]. Metagenomic analysis of viral communities from four major ocean regions using the same pyrosequencing technology has shown that essentially all marine viruses are spread widely throughout the oceans[1]. Identical phage-encoded exotoxin genes, T7-like DNA polymerase genes and T4-like structural genes are found in disparate terrestrial, aquatic and extreme environments[2–4]. Phages from soil, sediments and fresh water can productively infect marine microbes[13,14], showing that viruses move between major biomes.

Our metagenomic analysis of viral communities associated with a marine stromatolite (Highborne Cay, Bahamas) and two neighbouring (30 km) freshwater thrombolites and stromatolites (Pozas Azules II and Rio Mesquites, Mexico; Supplementary Fig. 1) showed that most of the sequences (98.8, 99.3 and 97.7% for Highborne Cay, Pozas Azules and Rio Mesquites, respectively) were unique when compared with the sequences in the non-redundant GenBank/SEED databases (BLASTx, E-value $<10^{-2}$). This proportion is much higher than any other previously sequenced viral metagenome (70–90% unknowns[1,15]). A comparison of microbialite metagenomic sequences with 42 viral and 36 microbial metagenomic libraries generated using the same pyrosequencing technology (Tables 1 and 2, respectively; Supplementary Tables 1 and 2 for details), showed that they were less than 5% similar (BLASTn, E-value $<10^{-3}$), further confirming that these are largely unrelated viral communities.

Using the approach developed by Angly *et al.*[1], random subsets of 10,000 sequences from each virome were assembled against each other to identify cross-contigs (that is, sequence overlaps between two samples). A read from one metagenome that assembled with a read from another metagenome indicated an overlap between these two metagenomes[1]. Only contigs produced by sequences from different metagenomes were taken into account to assess how many species were common to the two communities (percentage shared)[1]. Comparisons between Highborne Cay and Pozas Azules II and between Highborne Cay and Rio Mesquites did not produce any cross-contigs, indicating that none of the viruses was shared between these microbialites. The Pozas Azules II-Rio Mesquites comparison produced a very small average cross-contig spectrum, again indicating that essentially nothing is shared between these samples, even though they were taken from microbialites located 30 km from each other. A Monte Carlo analysis of the cross-contig spectra showed that the percentage of genome shared between Pozas Azules II, Highborne

[1]Department of Biology, [2]Computational Sciences Research Center, [3]Center for Microbial Sciences, San Diego State University, San Diego, California 92182, USA. [4]Genome Institute of Singapore, Singapore 138672, Singapore. [5]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [6]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida 33149, USA. [7]Department of Statistics, Rice University, Houston, Texas 77251, USA. [8]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México AP 70-275 Coyoacán, 04510 Mexico D.F., Mexico. [9]Department of Earth Science, University of California Santa Barbara, Santa Barbara, California 93106, USA. [10]College of Marine Science, University of South Florida, St Petersburg, Florida 33701, USA.

Table 1 | Similarity among the microbialite viral metagenomes and other environmental viral metagenomes

	Average percentage similarity (BLASTn, E-value $<10^{-3}$)*		
	Highborne Cay viral metagenome	Pozas Azules II viral metagenome	Rio Mesquites viral metagenome
Highborne Cay	100	1.140	0.910
Pozas Azules II	4.020	100	1.100
Rio Mesquites	0.970	0.700	100
Freshwaters ($n = 4$)	1.154 ± 0.240	0.477 ± 0.031	0.916 ± 0.278
Coral reef waters ($n = 4$)	1.462 ± 0.285	0.840 ± 0.032	0.808 ± 0.043
Marine waters ($n = 4$)	1.770 ± 0.573	0.585 ± 0.116	0.543 ± 0.098
Fish ($n = 4$)	0.701 ± 0.156	0.279 ± 0.015	0.387 ± 0.061
Mosquito ($n = 1$)	0.731	0.273	0.683
Coral ($n = 6$)	0.735 ± 0.150	0.290 ± 0.027	0.243 ± 0.024
Human ($n = 2$)	0.881 ± 0.336	0.377 ± 0.019	0.375 ± 0.019
Saltern waters ($n = 11$)	0.690 ± 0.145	0.439 ± 0.059	0.445 ± 0.058
Marine sediments ($n = 3$)	0.654 ± 0.079	0.568 ± 0.057	0.401 ± 0.089

* Average percentage similarity ± s.e.m.

Table 2 | Similarity among the microbialite viral metagenomes and other environmental microbial metagenomes

	Average percentage similarity (BLASTn, E-value $<10^{-3}$)*		
	Highborne Cay viral metagenome	Pozas Azules II viral metagenome	Rio Mesquites viral metagenome
Highborne Cay	47.104	0.400	0.230
Pozas Azules II	4.310	3.742	0.410
Rio Mesquites	1.021	0.637	0.541
Freshwaters ($n = 4$)	1.853 ± 0.609	0.466 ± 0.083	0.559 ± 0.091
Coral reef waters ($n = 4$)	0.903 ± 0.256	0.340 ± 0.050	0.276 ± 0.022
Fish ($n = 4$)	0.288 ± 0.015	0.252 ± 0.007	0.331 ± 0.038
Coral ($n = 7$)	0.805 ± 0.167	0.255 ± 0.016	0.252 ± 0.031
Saltern waters ($n = 11$)	0.655 ± 0.122	0.419 ± 0.034	0.398 ± 0.037
Subterranean ($n = 2$)	0.959 ± 0.377	0.442 ± 0.045	0.470 ± 0.122
Marine sediments ($n = 1$)	1.168	0.432	0.321

* Average percentage similarity ± s.e.m.

Cay and Rio Mesquites was zero (Supplementary Fig. 5) and therefore that the viruses are genetically unique in all three microbialites.

The small number of 'known' phage sequences in the microbialite metagenomes was assigned taxonomical designations based on the top BLAST similarities (Fig. 1, right panel). Their relative abundances were plotted onto the Phage Proteomic Tree[16] (PPT; Fig. 1, left panel). Microphages (icosahedral single-stranded DNA phages infecting *Escherichia coli, Bdellovibrio, Chlamydia* and *Spiroplasma* species[17], Supplementary Fig. 3) were the most common phages in the Highborne Cay and Pozas Azules II phage communities, representing 93.1% and 13.5% of the known phage sequences, respectively. In contrast, microphages were absent in Rio Mesquites, and the phage community was dominated by *Shewanella oneidensis* prophages (MuSo2 and LambdaSo) and *Burkholderia cepacia* phage sequences (54.6% of the total number of phage reads). At the taxonomic resolution of the PPT, the Highborne Cay and Pozas Azules II viral communities resembled each other and a previously described marine virome from the Sargasso Sea, which also contained high

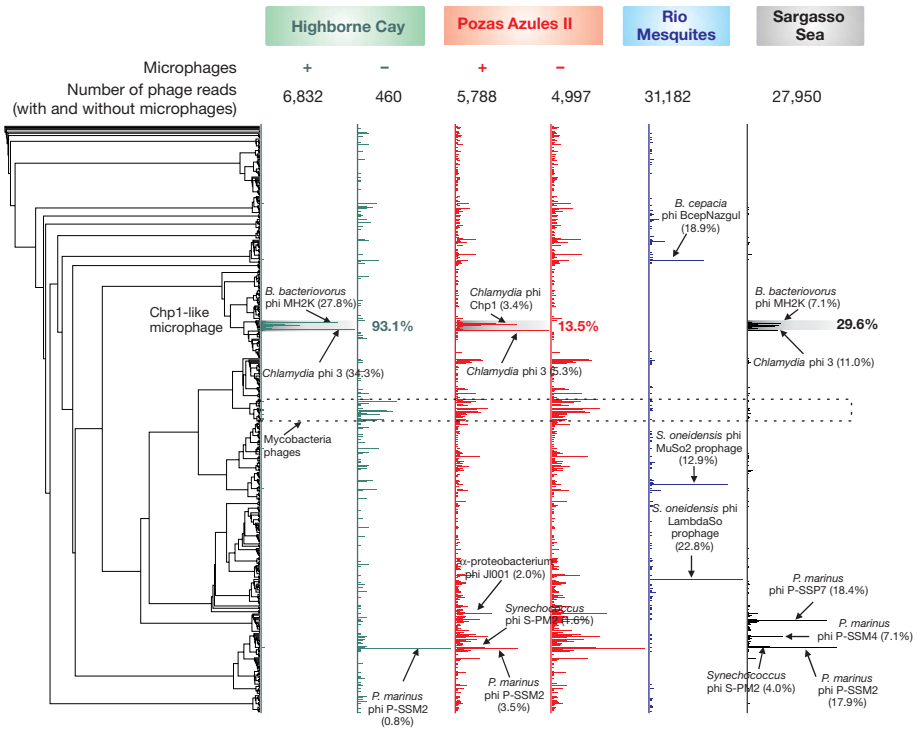

Figure 1 | The phage proteomic tree. The tree (left) shows the similarities of the viral metagenomic sequences to completely sequenced phage genomes. The presence and abundance of phage reads (right; abundance is proportional to line length) are presented in green for Highborne Cay, red for Pozas Azules II, blue for Rio Mesquites and grey for the Sargasso Sea samples. The total number of reads with significant similarity to phages (plus and minus microphages) is also indicated for Highborne Cay and Pozas Azules II. The name of the phage associated with the most abundant reads of each metagenome is given as well as the percentage of the total represented by these reads.

abundances of microphages (29.6%), *Prochlorococcus* phages P-SSM2 and P-SSM4 and *Synechococcus* phage S-PM2 (ref. 1) (Fig. 1).

Genetic distances of the microphages in Highborne Cay, Pozas Azules II and the Sargasso Sea were calculated using global alignments of the viral capsid protein (Vp1) reconstructed from the metagenomes (Fig. 2). The microphages from these three environments clustered together and were branched to the group of phages infecting *Chlamydia*. However, cross-assembly of the microphage nucleic-acid sequences did not produce a single cross-contig, indicating that amino-acid-level functionality is maintained but the nucleic acids have significantly diverged. On the basis of each consensus sequence recovered from the Highborne Cay, Pozas Azules II and Sargasso Sea metagenomes (Supplementary Information part 2), primers targeting the *Vp1* genes were designed (Supplementary Table 4). The capsid genes were successfully amplified from these metagenomes. No polymerase chain reaction (PCR) products were obtained when one sample was tested with the two other primer sets (for example, PCR of Highborne Cay viral DNA with the Pozas Azules II or the Sargasso Sea primer sets). Phylogenetic analysis of PCR products from the Highborne Cay sample showed that the similarity between clones and cultured microphage capsid sequences ranged from 47.5 to 61.2% at the nucleic-acid level and from 37.2 to 69.3% at the protein level, respectively (Supplementary Figs 8A and 8B).

We previously recovered cosmopolitan, essentially identical, T7-like podophage DNA polymerase sequences in the major biomes on Earth, including: marine, freshwater, sediment, terrestrial, extreme and metazoan-associated[3]. These environmental samples, as well as other marine microbial mats from different parts of the world (11 samples—from France, Israel, Bahamas, Puerto Rico and Connecticut, USA), were tested for the presence of the Highborne Cay microphages (Supplementary Table 5). No such microphages were detected in all the environmental samples tested, even though our PCR was sensitive enough to amplify fewer than 100 copies of the *Vp1* gene (Supplementary Fig. 6). New Highborne Cay stromatolite samples (July 2007) tested positive for the presence of the microphages, further confirming that these phages are native to the Highborne Cay stromatolites and persistent across time. To our knowledge, this is the first evidence of endemism in phages.

A 'marine signature' of the microbes from the Cuatro Ciénegas Basin was recently described by Souza *et al.*[18], implying that the whole ecosystem may be derived from an ancient marine community. Similarly, weighted and unweighted Unifrac analyses of the PPT (Supplementary Figs 4A, B) showed a genetic overlap between the Gulf of Mexico, the Sargasso Sea and the Pozas Azules II phage communities, even though these environments have not been in contact since the late Jurassic. This observation supports the hypothesis that phages in modern microbialites may be relicts from an ancient community. An alternative hypothesis that we cannot exclude is that there was a recent marine phage introduction, possibly through aerial vectors such as birds or airborne particles. However, the observation that these microbialite phages are extremely diverged from the global virome and from its nearest neighbour is more congruent with our ancient phage hypothesis.

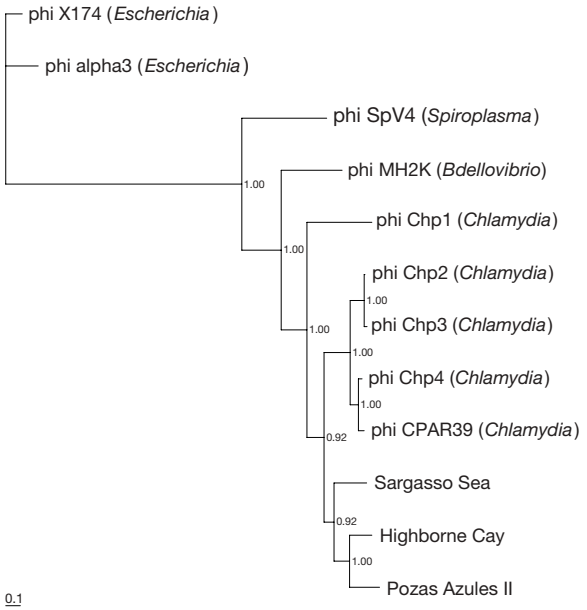

Figure 2 | Phylogenetic relationships among viral capsid amino-acid sequences of microphages. The Bayes values represent the proportion of sampled trees in which those sequences are clustered together.

METHODS SUMMARY

Microbialites were collected from the Pozas Azules II (PAII) pool and the Rio Mesquites (RM) River located in the Cuatro Ciénegas Basin (Mexico) and from the Highborne Cay (HC) marine waters (Bahamas). The viral particles were resuspended and purified using a combination of filtration and caesium chloride density gradient centrifugation[15]. Viral DNA was isolated by a formamide/CTAB extraction[19] and amplified with GenomiPhi (GE Healthcare) following the manufacturer's recommendations. Approximately 10 μg purified DNA was sequenced using pyrosequencing technology[20] (454 Life Sciences).

The sequences from each metagenome were compared to the SEED non-redundant database, our in-house phage database and 78 other metagenomes (using BLAST). The presence and the abundance of the sequences that have the phage databases were mapped onto the PPT (Fig. 1) using Bio-Metamapper (http://scums.sdsu.edu/Mapper). The diversity of the viral community and the percentage of viral genomes shared among samples were determined as previously described[1]. The genetic distances were calculated using the online UniFrac tool[21]. The Isolation by Distance web service[22] was used to test the correlation of the geographical distance and the genetic divergence between two viral communities.

Microphage capsid consensus sequences were reconstructed from the HC, PAII and Sargasso Sea[1] metagenomes and replaced onto a phylogenetic tree (Fig. 2). Primers were designed on the basis of these sequences (Supplementary Table 4) to retrospectively amplify the microphage capsid from the HC stromatolites. These sequences were cloned, sequenced (8 clones) and replaced in phylogenetic trees (Supplementary Figs 8A and 8B). PCR detection limit was defined (Supplementary Fig. 6) and optimal conditions were used to test the occurrence of the HC microphages in 63 different environmental samples (Supplementary Table 5).

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

Received 5 December 2007; accepted 23 January 2008.
Published online 2 March 2008.

1. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4,** e368 (2006).
2. Casas, V. *et al.* Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. *FEMS Microbiol. Lett.* **261,** 141–149 (2006).
3. Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236,** 249–256 (2004).
4. Short, C. M. & Suttle, C. A. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* **71,** 480–486 (2005).
5. Walter, M. R. *Stromatolites*. (Elsevier, Amsterdam, 1976).
6. Allwood, A. C., Walter, M. R., Kamber, B. S., Marshall, C. P. & Burch, I. W. Stromatolite reef from the Early Archaean era of Australia. *Nature* **441,** 714–718 (2006).
7. Schopf, J. W. Fossil evidence of Archaean life. *Phil. Trans. R. Soc. Lond. B* **361,** 869–885 (2006).
8. Suttle, C. A. Viruses in the sea. *Nature* **437,** 356–361 (2005).
9. Cho, J. C. & Tiedje, J. M. Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl. Environ. Microbiol.* **66,** 5448–5456 (2000).
10. Papke, R. T., Ramsing, N. B., Bateson, M. M. & Ward, D. M. Geographical isolation in hot spring cyanobacteria. *Environ. Microbiol.* **5,** 650–659 (2003).
11. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Geographic barriers isolate endemic populations of hyperthermophilic *Archaea*. *Science* **301,** 976–978 (2003).
12. Whitaker, R. J. Allopatric origins of microbial species. *Phil. Trans. R. Soc. Lond. B* **361,** 1975–1984 (2006).
13. Sano, E., Carlson, S., Wegley, L. & Rohwer, F. Movement of viruses between biomes. *Appl. Environ. Microbiol.* **70,** 5842–5846 (2004).

14. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13,** 278–284 (2005).
15. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl Acad. Sci. USA* **99,** 14250–14255 (2002).
16. Rohwer, F. & Edwards, R. A. The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.* **184,** 4529–4535 (2002).
17. Fane, B. *Microviridae*, in *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses*. (eds Fauquet, M. A. M. C., Maniloff, J., Desselberger, U. & Ball, L. A.) 289–299 (Elsevier Academic Press, San Diego, California, 2005).
18. Souza, V. *et al.* An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proc. Natl Acad. Sci. USA* **103,** 6565–6570 (2006).
19. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, New York, 1989).
20. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
21. Lozupone, C., Hamady, M. & Knight, R. UniFrac - An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).
22. Jensen, J., Bohonak, A. & Kelley, S. Isolation by distance, web service. *BMC Genet.* **6,** 13 (2005).

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

Acknowledgements Logistical field support was provided by the crew of the RV *Walton Smith*, Highborne Cay management and personnel of the Area de Proteccion de Flora y Fauna of Cuatro Ciénegas. This work was supported by an NSF grant to F.R. Support for B.K.S. and D.L.V. was provided by the NSF. M.B. was supported by a grant from the University of South Florida's Internal New Research Awards Program. V.S. was funded by the CONACYT 2002-C01-0237 project. The authors thank P. Visscher, K. Przekop, L. Rothschild, D. Rogoff, V. Michotey, P. Bonin, S. Norman and E. Bowlin for providing samples of marine microbial mats and M. Schaechter for a critical reading of the manuscript.

Author Contributions C.D. and F.R. designed the project. C.D. analysed most of the bioinformatic results, conducted the molecular biology and wrote the article. S.K. performed the bayesian analysis. S.R. implemented the cross-contig analyses. M.H. extracted viral DNAs. B.R.-B., H.L., F.E.A. and R.A.E. performed bioinformatic analyses. R.V.T. and D.H. helped with the interpretation of the bioinformatic results. V.S., M.B., J.S. and R.P.R. collected the samples. B.K.S., D.L.V., M.F., T.T., L.L., Y.R., L.W. and B.C. provided metagenomic data. F.R. supervised the project and helped with the writing. All authors edited and commented on the manuscript.

Author Information The microbialite viral metagenomes have been deposited into the ftp server of the SEED public database ftp://ftp.theseed.org/metagenomes under the project accession numbers 4440323.3 (Highborne Cay), 4440320.3 (Pozas Azules II) and 4440321.3 (Rio Mesquites). The metagenomes are also publicly accessible in the CAMERA metagenomic database (http://camera.calit2.net) under the project accession numbers HBCStromBahamasVir011105 (Highborne Cay), PAStromCCMexVir072205 (Pozas Azules II), and RMStromCCMexVir072205 (Rio Mesquites). The *Vp1* cloned sequences from the Highborne Cay sample have been deposited in GenBank under accession numbers EF679227 to EF679234. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.D. (cdesnues@yahoo.fr).

METHODS

Geographical sampling. Microbialites were collected in November 2005 from the Cuatro Ciénegas Basin in Mexico and the Highborne Cay Island in the Bahamas (Supplementary Fig. 1). In Mexico, thrombolite samples were collected from a spring, thermally heated pool (Pozas Azules II, site 1) and a free flowing river system (Rio Mesquites, site 2). These two spring sources are geographically isolated by 30 km. Multiple subsamples were combined from the Highborne Cay stromatolites (Highborne Cay, site 3) and used as one sample. The geologic characteristics for the sampling sites were previously described in detail[18,23].

Virus purification, viral DNA extraction and pyrosequencing. Approximately 5 g of microbialite were shaken in 30 ml of SM buffer (0.1 M NaCl, 1 mM MgSO4, 0.2 M Tris pH 7.5, 0.01% gelatin) for Pozas Azules II and Rio Mesquites samples and in 30 ml of 0.02 μm filtered seawater for Highborne Cay sample for 1 hour. The viral particles were then purified using filtration (0.22 μm) combined with caesium chloride density gradient centrifugation[15]. The absence of microbial and eukaryotic cells was verified under epifluorescence microscopy after SYBR-Gold staining[24] (Supplementary Figs 2A and 2B). For electron microscopy, viral particles were stained with 1.0% uranyl acetate and examined with a FEI Tecnai 12 transmission electron microscope (Supplementary Fig. 2C). Viral DNA was isolated by a formamide/CTAB extraction[19] and amplified with GenomiPhi (GE Healthcare) following the manufacturer's recommendations. The resulting DNA was purified on silica columns (Qiagen) and concentrated by ethanol precipitation. Approximately 10 μg DNA was sequenced using pyrosequencing technology[20] (454 Life Sciences). A total of 81,687,957 bp of DNA was generated from the three libraries (Pozas Azules II: 32 Mbp, Rio Mesquites: 35 Mbp and Highborne Cay: 15 Mbp). The 781,866 sequences had an average length of 104 bp. They have been deposited into the ftp server of the SEED public database ftp://ftp.theseed.org/metagenomes under the project accession numbers 4440323.3 (Highborne Cay), 4440320.3 (Pozas Azules II) and 4440321.3 (Rio Mesquites).

Bioinformatics. The sequences from each metagenome were compared to the SEED non-redundant (nr) database and environmental database using BLASTx[25] (E-value $<10^{-2}$). The SEED database contains annotated protein sequences from different databases such GenBank, Swiss-prot and KEGG. The environmental database contains, among other things, sequences from acid mine drainage, biofilm, soil or the Sargasso Sea. The best similarity for each sequence that matched an annotated protein in the SEED or environmental databases was automatically assigned as 'known' whereas 'unknown' describes sequences that did not have similarity to anything. To define the inter-library sequence similarities, the entire microbialite metagenomes were compared (BLASTn, E-value $<10^{-3}$) against each other and against other viral (Table 1) and microbial (Table 2) metagenomes from different environments (details are provided in Supplementary Tables 1 and 2, along with SEED accession numbers). All the metagenomes can be downloaded via the ftp server of the SEED database (ftp://ftp.theseed.org/metagenomes).

Structure of the viral communities. A set of 10,000 random sequences was extracted from each metagenome and assembled by the TIGR Assembler using a minimum overlap of 35 bp and 98% of sequence identity. Twenty repetitions were performed, leading to an average contig spectrum used to define the maximal likelihood community structure. Different rank-abundance models were calculated (Supplementary Table 3) using PHACCS (PHAge Communities from Contig Spectra) an online tool to analyse viral communities[26] (http://biome.sdsu.edu/phaccs/index.htm). As described previously[1], rank-abundance models as well as the cross-contig spectra generated between two metagenomes were used to define the percentage of genotypes that are shared between two communities (Supplementary Fig. 5). Even though the logarithmic rank-abundance model was not the best model for Rio Mesquites and Highborne Cay, it gave coefficients of errors close to those observed with the best models. To harmonize the analysis and to limit the possible bias during the simulation, the same model (logarithmic) was chosen for the three metagenomes (Supplementary Table 3).

Phage community taxonomy. The metagenome sequences from each library were compared to the phage and prophage genome database using tBLASTx (E-value $<10^{-3}$). This database contains sequences from 510 complete genomes of phages and prophages and was used to construct the Phage Proteomic Tree version 4 (PPT, http://phage.sdsu.edu/~rob/PhageTree/v4). A previous version of the tree detailing the construction steps was published in 2002 (ref. 16). The presence and the abundance of sequences that have significant similarities to those in the database were subsequently mapped onto the PPT (Fig. 1) using BioMetamapper, an online metagenome mapper to the Phage Proteomic Tree (http://scums.sdsu.edu/Mapper).

Genetic versus geographical distance of the phage community. UniFrac, an online tool[21], was used to measure the genetic differences in community composition between microbialites and marine environments. The UniFrac distance is calculated as the percentage of the branch length of the tree (in this case, the Phage Proteomic Tree) that leads to descendants from either one environment or the other, but not both. In this study, a weighted UniFrac distance metric that also takes account of the relative abundance of sequences in the different environments was used. Distances between the sets of sequences from each pair of environments (stromatolites and marine environments) were classified from lower quartile (red) to upper quartile (yellow); that is, a range from complete similarity to complete differentiation in the phylogenetic diversity of the samples (Supplementary Fig. 4). The Isolation by Distance Web Service (IBDWS) was used to test for a correlation between the geographical distance between two samples and the genetic divergence between viral communities[22]. This online software uses Mantel tests to determine whether phages in closer physical proximity have greater genetic similarity (as measured by UniFrac), than those separated by large geographical distances (Supplementary Fig. 4).

Genetic divergence of the microphage sequences. The sequences that had significant tBLASTx similarities (E-value $<10^{-3}$) to microphages in the Highborne Cay and the Pozas Azules II metagenomes were extracted into a sublibrary. These microphage libraries were cross-compared at the nucleic-acid level against themselves and against the microphages of the Sargasso Sea metagenome[1] using Circonspect, an online tool to build contig-spectra (http://biome.sdsu.edu/circonspect/index.php). The sublibraries were then assembled with Sequencher 4.0 (Gene Codes) using a minimal match percentage of 98% and a 35 bp minimum overlap. When the largest contigs were compared with tBLASTx against the nr database, most had similarities to the viral capsid protein (Vp1) of sequenced microphage. Multiple alignments of Vp1 amino-acid sequences from known microphages and from Pozas Azules II, Highborne Cay and Sargasso Sea viral reconstructed Vp1 consensus sequences were performed using CLUSTAL W[27]. The phylogenetic tree was generated using MrBayes 3.1 program[28] (Fig. 2). The protein evolutionary model (BLOSUM) used for this bayesian analysis was chosen from among seven different models because it had the highest posterior probability in an initial test of all models for the data. We ran four independent Monte Carlo Markov chains for 1 million generations and the chains converged after only 10,000 generations. To verify the assembly results, PCR primers were designed on the basis of the *Vp1* consensus sequences (Supplementary Table 4) and PCRs were performed on each sample. The reaction mixture (50 μl total) contained target DNA, 1x Taq Buffer, 0.2 mM dNTPs, 1 μM each primer, and 1 U Taq DNA polymerase. The thermocycler conditions were: 5 min at 94 °C; 30 cycles of 1 min at 94 °C, 1 min at 52 °C, 1 min at 72 °C; and 10 min at 72 °C. Amplification products were checked for size on a 1% agarose gel. No PCR product was obtained when one sample was tested with the two other primer sets (for example, PCR of Highborne Cay viral DNA with the Pozas Azules II or the Sargasso Sea primer sets; data not shown). PCR products from the Highborne Cay sample were cloned into a TOPO TA vector (Invitrogen) and transformed into Top 10 competent cells (Invitrogen). PCR was used to screen positive colonies using primers M13F and M13R provided by the TOPO TA cloning kit and following manufacturer's instructions. PCR products from eight clones were purified using a PCR clean-up kit (Mo Bio) and sequenced using the M13F and M13R primers (sequences are in the Supplementary Information part 3, accession numbers EF679227 to EF679234). Multiple sequence alignments of the clones and the known microphage Vp1 sequences were made using CLUSTAL W[27] (Supplementary Fig. 7). The nucleic-acid and protein-based phylogenetic trees (Supplementary Figs 8A and 8B, respectively) were constructed using the neighbour-joining method[29] and were plotted using the njplot program[30]. Plasmid purifications were completed using PureLink Quick Plasmid Miniprep Kit (Invitrogen).

Highborne Cay microphages in other environmental samples. The clone D4 was used to test the limit of the *Vp1* gene concentration for PCR detection. Serial dilutions were made to produce final concentrations ranging from 1 to 10^9 plasmid copies per microlitre (Supplementary Fig. 6). One microlitre of each dilution was then amplified with the Vp1HC-F and Vp1HC-R set of primers using touchdown PCR and a gradient of primer hybridization temperature ranging from 47 °C to 57 °C. The thermocycler conditions giving optimal PCR amplification (detection limit between 10 and 100 plasmid copies) were: 5 min at 94 °C, 20 cycles of (1 min at 94 °C, 1 min at 65–0.5 °C per cycle, and 1 min at 72 °C) followed by 15 cycles of (1 min at 94 °C, 1 min at 55 °C, and 1 min at 72 °C); and 10 min at 72 °C. These PCR conditions were then used to test the presence or absence of the Highborne Cay *Vp1* gene in 63 different environmental samples (Supplementary Table 5) including extreme, metazoan-associated, freshwater, marine, sediment, terrestrial, other marine mats and new viral DNA from the Highborne Cay stromatolites.

23. Reid, R. P., Macintyre, I. G. & Steneck, R. S. A microbialite/algal ridge fringing reef complex, Highborne Cay, Bahamas. *Atoll Res. Bull.* **465**, 1–18 (1999).

24. Chen, F., Lu, J., Binder, B. J., Liu, Y. & Hodson, R. E. Application of digital image analysis and flow cytometry to enumerate marine viruses stained with SYBR Gold. *Appl. Environ. Microbiol.* **67,** 539–545 (2001).
25. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215,** 403–410 (1990).
26. 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).
27. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
28. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
29. Saito, N. & Nei, M. The neighbour-joining method, a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **79,** 426–434 (1987).
30. Perriere, G. & Gouy, M. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie* **78,** 364–369 (1996).