RESEARCH ARTICLE

Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage

Rika E. Anderson, William J. Brazelton & John A. Baross

School of Oceanography and Astrobiology Program, University of Washington, Seattle, WA, USA

Correspondence: Rika E. Anderson, School of Oceanography and Astrobiology Program, University of Washington, Box 357940, Seattle, WA 98195, USA. Tel.: +1 206 543 0546; fax: + 1 206 543 6073; e-mail: rikander@u.washington.edu

Received 20 December 2010; revised 4 March 2011; accepted 7 March 2011.
Final version published online 11 April 2011.

DOI:10.1111/j.1574-6941.2011.01090.x

Editor: Julian Marchesi

Keywords
viral ecology; metagenome; hydrothermal vent ecology.

Abstract

Metagenomic analyses of viruses have revealed widespread diversity in the viriosphere, but it remains a challenge to identify specific hosts for a viral assemblage. To address this problem, we analyze the viral metagenome of a northeast Pacific hydrothermal vent with a comprehensive database of spacers derived from the clustered regularly interspaced short palindromic repeat (CRISPR) putative immune system. CRISPR spacer matches to the marine vent virome suggest that viruses infecting hosts from diverse taxonomic groups are present in this vent environment. Comparative virome analyses show that CRISPR spacers from vent isolates and from thermophiles in general have a higher percentage of matches to the vent virome than to other marine or terrestrial hot spring viromes. However, a high percentage of hits to spacers from mesophilic hosts, combined with a moderately high modeled alpha diversity, suggest that the marine vent virome is comprised of viruses that have the potential to infect diverse taxonomic groups of multiple thermal regimes in both the bacterial and the archaeal domains.

Introduction

Viruses play important ecological, biogeochemical and evolutionary roles throughout the world's ecosystems, particularly in the oceans (Suttle, 2005). Several viral metagenomes, or viromes, have been published from a wide range of marine and terrestrial environments (Breitbart *et al.*, 2002; Angly *et al.*, 2006; Bench *et al.*, 2007; Desnues *et al.*, 2008; Schoenfeld *et al.*, 2008; Williamson *et al.*, 2008b; Lopez-Bueno *et al.*, 2009; Santos *et al.*, 2010). These reports have demonstrated the mobility of viral genes between environments as well as the tremendous diversity of genes encoded by the global viriosphere (Breitbart & Rohwer, 2005; Dinsdale *et al.*, 2008; Kristensen *et al.*, 2009).

These metagenomic analyses face several challenges, however. First, in most viromes published to date, the vast majority of reads have no match to existing databases, while the majority of the rest have matches to bacterial or archaeal genes (see e.g. Angly *et al.*, 2006). This high percentage of unknown sequences renders further identification of viral types or viral genes more challenging. Moreover, the isolation of viral particles independent of their hosts, as is done for most viromes, makes it difficult to identify which hosts are targeted by the viral assemblage. However, identification of viral hosts is crucial to understanding the role of viruses in an ecosystem. Identifying viral hosts would aid in determining how viruses impact the microbial diversity of a given ecosystem, for example, or which microorganisms may be sharing genes through virally mediated horizontal gene transfer. To address this problem, we have analyzed a viral metagenome from a diffuse flow hydrothermal vent with a comprehensive database created from the clustered regularly interspaced palindromic repeat (CRISPR) immune system.

The CRISPR system is a putative antiviral immunity mechanism found in both archaea and bacteria (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Sorek *et al.*, 2008; van der Oost *et al.*, 2009; Hovarth & Barrangou, 2010; Labrie *et al.*, 2010; Marraffini & Sontheimer, 2010). CRISPR loci generally consist of a series of short repeats, each approximately 20–50 bp in length, interspersed by spacers about 25–75 bp in length (Grissa *et al.*, 2007a). CRISPR loci are thought to create immunity when short sequences derived from invaders such as viruses or plasmids are incorporated as spacers between the repeat sequences by genes involved in the CRISPR response, known as CRISPR-associated (*cas*) genes. When introduced genetic elements, such as viruses or plasmids, have a 100% match to a pre-existing CRISPR spacer sequence in the host genome, these elements are recognized as pathogenic invaders (Makarova *et al.*, 2003; Bolotin *et al.*, 2005; Haft *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005; Marraffini & Sontheimer, 2008; Hale *et al.*, 2009). In response, the CRISPR/Cas system cleaves the invading nucleic acid (Garneau *et al.*, 2010).

CRISPR loci effectively act as libraries of previous viral infection; thus, analyses of CRISPR spacers across species have great potential for microbial and viral ecology. Previous studies have examined CRISPRs in an ecological context, focusing on variability and distribution in acid mine drainage biofilms (Andersson & Banfield, 2008; Tyson & Banfield, 2008) as well as in terrestrial hot springs (Heidelberg *et al.*, 2009; Held & Whitaker, 2009; Held *et al.*, 2010). These studies have found a high degree of variability within CRISPR spacer sequences, implying a rapid rate of host–virus coevolution. These studies have also demonstrated a clear biogeographic distribution in CRISPR spacers. Additionally, Snyder *et al.* (2010) have designed microarrays using CRISPR spacer sequences to detect viruses in environmental samples.

CRISPR spacers provide a means to analyze and compare viral sequences for which we have some host genomic context (i.e. the complete genomes of isolates), whereas metagenomics provides information about the genetic content of a viral assemblage at a particular location and point in time. Here, by comparing a database of CRISPR spacers from all published archaeal and bacterial genomes with reads from a viral metagenome, we are able to infer what types of hosts might be infected by the viruses in the viral assemblage, even if their sequences have no close BLAST matches in available databases.

The environment we have chosen as our focus for this analysis is a diffuse flow hydrothermal vent system in the Main Endeavour Field in the northeast Pacific Ocean. As in other marine environments, the virus to cell ratio is approximately 10 to one in vents at the Main Endeavour Field (Ortmann & Suttle, 2005), yet induction experiments have shown that vent communities of the East Pacific Rise display a higher incidence of lysogeny than other marine environments (Williamson *et al.*, 2008a). While it is evident that viruses play a prominent role in the vent environment, until now, the diversity, structure and genomic content of vent viral communities have not been assessed.

The dynamic, gradient-dominated nature of the vent environment makes it a particularly attractive site for studies of viral ecology and evolution. In these environments, ambient seawater mixes with high-temperature hydrothermal fluid enriched in reduced compounds, creating gradients in pH, temperature, chemical composition, and mineralogy both above and below the seafloor (Baross & Hoffman, 1985). These gradients set up a series of microenvironments, providing niches for diverse communities of microorganisms (Huber *et al.*, 2003; Schrenk *et al.*, 2003). Continuous circulation of the hydrothermal fluid both above and below the seafloor enables potentially frequent contact among these microbial communities and their accompanying viral assemblages. In such an environment, viruses of a diverse array of hosts could also potentially come into frequent contact with each other. As viruses are known vectors of horizontal gene exchange, the presence of a wide diversity of viruses and their hosts could facilitate widespread gene transfer. This analysis, with an emphasis on identifying potential viral hosts, provides a new perspective on a viral assemblage whose unique signature mirrors the dynamic yet extreme environment it inhabits.

Materials and methods

Diffuse flow hydrothermal fluid virus sampling

Hydrothermal vent fluid (170 L) was collected with a barrel sampler from diffuse flow at the base of Hulk vent in the Main Endeavour Field on the Juan de Fuca ridge (approximately 450 km west of Washington state in the Pacific Ocean). The sample funnel was placed atop a clump of tubeworms on a sulfide structure venting diffuse flow hydrothermal vent fluid (Supporting Information, Fig. S1). Chemical and physical metadata from Hulk vent are summarized in Table S1. The minimum temperature of the sample was 13 °C, as measured through a temperature probe on a hydrothermal fluid sampler (HFS) at the sample site. However, the average chemistry-derived temperature of the barrel sample, calculated based on dissolved silica content, was much higher. The measured silica content of hydrothermal fluid from a 300 °C black smoker about 10 m away from the sample site was 15 199 μM, whereas background seawater silica content was 185 μM at 1.8 °C. From this, our sample temperature was calculated to be approximately 125 °C. While this is much higher than the diffuse flow temperatures recorded by the HFS, it is possible that this is because diffuse flow measured by the intake nozzle of the HFS retained much higher amounts of seawater than that taken in by the intake funnel of the barrel sampler, which may have had a better seal on the sulfide structure and therefore pulled in higher temperature fluid.

Upon recovery, several 20-mL fluid subsamples were collected for cell and virus counts. Samples were fixed with 10% paraformaldehyde and stored at 4 °C for 2 weeks until counted. Cell and viral counts were conducted by filtering 1 mL of a 1/10 diluted sample onto a 25 mm 0.02-μm Anodisc filter (Whatman Inc., Kent, UK) backed by a GF/F nitrocellulose filter at < 20 kPa pressure. Filters were placed on a drop of 1–5 × SYBR Gold and allowed to sit for 15 min before mounting on slides with a filtered phosphate-buffered saline/glycerol/ascorbate solution. At least 200 cells and viruses were counted in a minimum of 20 fields of view.

For pyrosequencing, the sample was filtered with a 0.22-μm Steripak filter unit (Millipore, MA) on ice to remove cells. The filtrate was concentrated through tangential flow filtration (30 kDa cutoff) to approximately 400 mL (Thurber *et al.*, 2009) in a 4 °C cold room. Samples were stored in 50-mL fractions and frozen at − 80 °C. Upon thawing, 10% w/v PEG 8000 was added to one 50-mL fraction and incubated at 4 °C overnight. Each sample was pelleted by centrifugation at 13 000 ***g*** for 50 min, resuspended in TE and incubated for 15 min with 0.7 volume of chloroform to lyse any remaining cellular contamination. After centrifugation for 10 min at 4 °C to remove chloroform, the aqueous fraction was incubated with 10% DNase I for 2 h at 37 °C to eliminate any free DNA in solution. DNase was inactivated by adding EDTA to a final concentration of 0.02 M. Viral DNA was extracted using the QIAamp MinElute Virus Spin Kit (Qiagen Inc., CA). Samples were sent to the Broad Institute for 454 Titanium pyrosequencing (454 Life Sciences, Branford, CT).

Bioinformatics

Phylogenetic assignments of reads in the marine vent virome were carried out through the MG-RAST pipeline (Meyer *et al.*, 2008). Reads were compared with the SEED database with TBLASTX with a maximum *e*-value cutoff of 10^{-5}. Reads with a significant match to a viral sequence according to these parameters were categorized into families as defined by the International Commission on Taxonomy of Viruses 2009 release of Virus Taxonomy (http://www.ictvonline.org/virusTaxonomy.asp?bhcp=1). Marine vent virome contigs were assembled and analyzed using GENEIOUS (Drummond *et al.*, 2009) (http://www.geneious.com). Contigs were assembled using the 'Medium Sensitivity' method with a word length of 14, a maximum gap size of 2, maximum gaps per read of 15, and maximum mismatches of 15. Contig taxonomy for each read was defined according to the consensus taxonomy as defined by the taxonomy of the majority of reads within each contig. Read taxonomy was assigned through the MG-RAST pipeline by comparing with the SEED database, with a maximum *e*-value cutoff of 10^{-5}. Metagenomic reads were deposited to the CAMERA database (http://camera.calit2.net/) under accession number CAM_SMPL_A0003 under the "Moore Marine Phage/Virus Metagenomes" project.

Modeling uncultured viral assemblage diversity

The alpha diversity of each virome was estimated using the PHAGE COMMUNITIES FROM CONTIG SPECTRUM (PHACCS) online tool (http://biome.sdsu.edu/phaccs), described in previous publications (Angly *et al.*, 2005, 2006). Briefly, 10 000 random sequences were assembled using MINIMO (98% identity over at least 35 bp overlap). CIRCONSPECT (Angly *et al.*, 2006) (http://sourceforge.net/projects/circonspect/) was used to calculate a

Table 1. Diversity indices for seven viral metagenomes as calculated by PHACCS (http://biome.sdsu.edu/phaccs)

Virome	Number of reads	Average genome size (bp)	Average read length for CIRCONSPECT	Rank-abundance model	Model error	Richness (number of genotypes)	Evenness	Most abundant genotype (% of the community)	Shannon–Wiener index	References
Hulk hydrothermal vent (a)	228 698	56 013	100	Logarithmic	2.62	1730	0.970	2.81	7.23	This study
Hulk hydrothermal vent (b)	216 966	53 838	100	Logarithmic	1.22	1840	0.973	2.51	7.32	This study
Yellowstone hot springs – Bear Paw	8352	35 340	650	Power	313	1610	0.855	6.27	6.32	Schoenfeld *et al.* (2008)
Yellowstone hot springs – Octopus	22 272	29 086	650	Power	281	2340	0.94	2.03	7.29	Schoenfeld *et al.* (2008)
Arctic Ocean	688 590	57 927	100	Lognormal	1.58	886	0.888	3.26	6.03	Angly *et al.* (2006)
Bay of British Columbia	416 456	65 922	100	Logarithmic	35.3	2020	0.952	4.45	7.25	Angly *et al.* (2006)
Gulf of Mexico	263 907	60 921	100	Logarithmic	159	9020	0.906	8.31	8.25	Angly *et al.* (2006)
Sargasso Sea	399 343	65 104	100	Logarithmic	89.4	2990	0.890	10.1	7.12	Angly *et al.* (2006)

Hulk hydrothermal vent diversity was modeled twice: (a) modeled diversity from all reads and (b) modeled diversity only from reads in contigs shorter than 3 kbp. See text for details. Model error describes the difference between the modeled contig spectrum from each diversity model and the actual contig spectrum.

contig spectrum by calculating the number of contigs of each size, using a minimum metagenome coverage of 2, a minimum dinucleotide entropy of 2.0, low-complexity filter window length of 21, and with varying trim and discard sizes depending on the average read length of the metagenome (Table 1). The average viral genome length was estimated using GAAS (Angly *et al.*, 2009) through a CAMERA 2.0 alpha diversity workflow (http://camera.calit2.net/).

CRISPR spacer analyses

All genomes analyzed in this study (1083 archaea and bacteria) were downloaded from the NCBI ftp server on April 20, 2010. CRISPRs were identified in each of these genomes with the CRISPR RECOGNITION TOOL (CRT) (Bland *et al.*, 2007) using default parameters, and the number of CRISPR loci and total CRISPR spacers per genome were tabulated. CRISPR spacers were compiled into a single database and categorized by genome. The CRISPR spacer database has been deposited to the publicly accessible Data Dryad repository at http://dx.doi.org/10.5061/dryad.8826. All spacer comparisons were conducted with BLASTN (Altschul *et al.*, 1990). Spacer 'matches' were defined as matches of 100% identity along at least 20 base pairs of the spacer sequence. To compare the spacers in the database with each other, we performed a BLASTN comparison of the set of spacers within an individual genome against the set of spacers in each of the other genomes and then compiled these results into a resemblance matrix (Table S2). From this, we determined what proportion of all CRISPR spacers between the two genomes was shared. To compare the CRISPR spacer database with metagenomic reads, a FASTA file containing all spacers in the database was compared with the raw metagenomic reads of each metagenome using BLASTN, where a 'match' was again defined as 100% identity across 20 base pairs. To calculate the percentage of reads with a match to a spacer, only unique queries (reads) were counted. For the analysis in which we identified the taxonomy of potential hosts, we included matches of multiple reads to the same spacer, as well as multiple spacers to the same read.

For analysis of CRISPR spacer matches from each temperature regime, each genome in the NCBI database was sorted according to thermal regime as defined by a genome properties list downloaded from the NCBI ftp server. 'Vent isolates' were characterized as all strains, both thermophilic and mesophilic, that had been isolated from either a shallow or a deep-sea hydrothermal vent. These are listed in Table S3.

To compare the average growth temperature with CRISPR abundance, archaea and bacteria were grouped as thermophiles (optimal growth temperature of 60 °C or above; includes hyperthermophiles) or mesophiles (optimal growth temperature between 25 and 60 °C), and some bacteria were designated as psychrophiles (optimal growth temperature < 25 °C).

Results and discussion

The structure of our analysis focused first on ensuring virome quality through contig analysis. BLAST analyses were conducted on virome reads to gain an overall picture of both the structure and the content of the marine vent viral assemblage, and to determine which viral families were present. Next, we modeled the richness and evenness of the viral assemblage and compared this with previously sequenced marine and hot springs viromes. Finally, to provide host context for these results, we queried the marine vent virome with a comprehensive CRISPR spacer database to identify potential microbial hosts of viruses in the marine vent viral assemblage.

Matches of marine vent virome reads to known sequences

Of 228 698 reads, the majority (67.14%) of reads in the marine vent virome yielded no matches to the SEED database (*e*-value cutoff of 10^{-5}) (Fig. 1). This viral metagenome has a smaller percentage of unknown reads than found in some previous marine viral metagenomes (Table S4). However, this may be an artifact of read length: this metagenome, sequenced with 454 Titanium technology, had an average read length of 334 bp, whereas marine viral metagenomes sequenced with 454 FLX technology averaged approximately 100 bp (Angly *et al.*, 2006). Longer reads are more likely to have significant matches to existing databases. Similar percentages of unknown reads have been found in viral metagenomes with longer read lengths (Bench *et al.*, 2007; Schoenfeld *et al.*, 2008), although this is not true for all cases (Lopez-Bueno *et al.*, 2009). It is possible that contamination with cellular sequences may contribute to the relatively low percentage of unknown sequences; however,

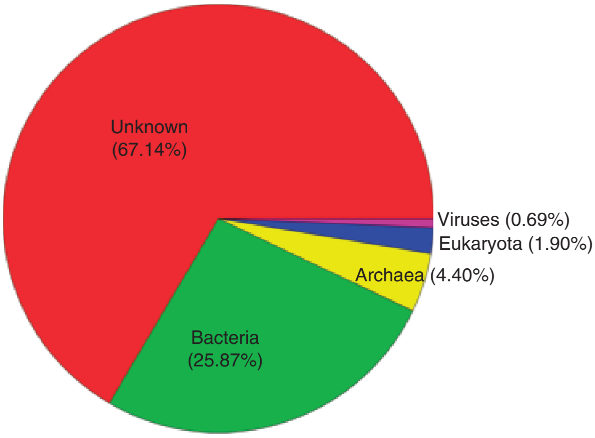

Fig. 1. Distribution of reads in the hydrothermal vent virome with matches to the SEED database, maximum *e*-value 10^{-5}. All analyses were performed through the MG-RAST pipeline.

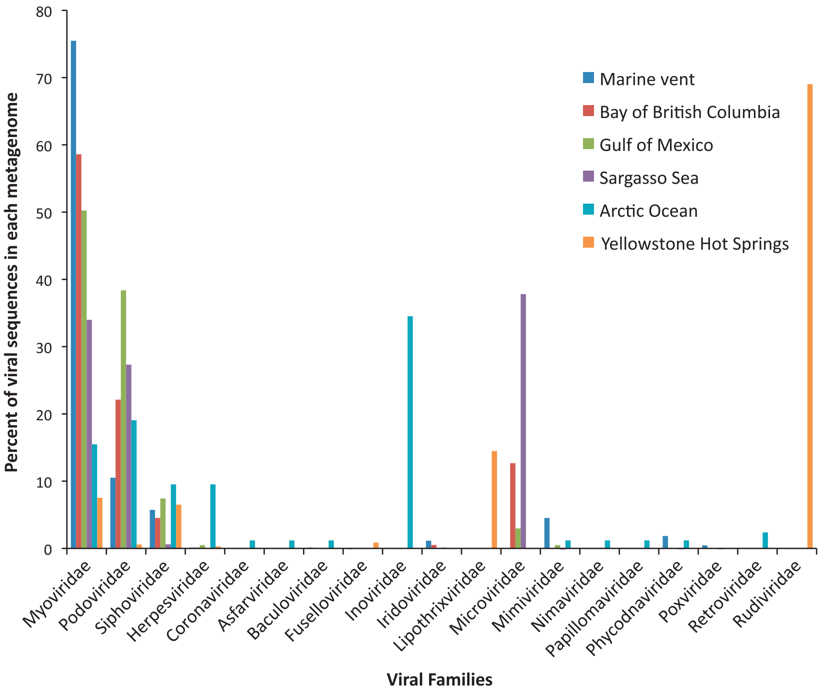

Fig. 2. Comparison of viral family types present in the hydrothermal vent viral assemblage as well as that of four marine biomes and terrestrial hot springs. Marine viral metagenomes from Angly *et al*. (2006), and Yellowstone hot springs viral metagenomes from Schoenfeld *et al*. (2008).

we believe that a significant portion of the metagenome was viral. This is discussed in greater detail below.

Of the reads in the marine vent virome with a significant database match (where a match corresponded to an *e*-value cutoff of 10^{-5}), 25.87% matched bacterial sequences, 4.40% matched archaeal sequences, and only 0.69% matched known viral sequences. Similar proportions have been found in previously sequenced viromes. In general, the abundance of bacterial and archaeal matches may be explained by the larger number of archaeal and bacterial sequences in the database and possibly also by a high rate of horizontal gene transfer between viruses and their hosts, resulting in the presence of microbial genes in viral genomes and vice versa (Angly *et al.*, 2006).

We next examined the presence of specific viral families based on reads with matches to known viral sequences. The results, shown in Fig. 2, suggest that the viral assemblage at marine vents is more similar to other marine viral assemblages than to those in terrestrial hot springs. As only DNA was sequenced, this analysis would necessarily miss RNA viruses or retroviruses, but the presence of DNA viruses among different biomes can be compared. The majority of viral reads in the marine vent virome belonged to the *Myoviridae* family, as is the case with many other marine viromes (Fig. 2). Other tailed viruses common to marine viromes, the *Podoviridae* and *Siphoviridae*, were also relatively common in the marine vent virome. Recent studies have shown that single-stranded DNA (ssDNA) viruses such as *Microviridae* are predominant in temperate marine waters such as the Sargasso Sea and the Bay of British Columbia (Angly *et al.*, 2006), and yet sequences matching the *Microviridae* family were largely absent from the marine vent virome. However, unlike other viromes, our sample was not amplified with Phi29 polymerase, which is biased toward the amplification of ssDNA viruses, and may explain the relative lack of ssDNA viruses in this virome (Kim *et al.*, 2008).

Viruses known to infect archaea such as the *Rudiviridae*, *Fueselloviridae*, and *Lipothrixviridae*, commonly found in hot spring viral assemblages (Prangishvili *et al.*, 2006; Schoenfeld *et al.*, 2008), were largely absent from the marine vent virome. The abundance of archaea in marine vents would suggest that it is unlikely that archaeal viruses are absent from the marine vent assemblage, and therefore, this implies that archaeal viruses present in the marine vent assemblage were unlike any sequenced strains found in terrestrial hot springs. Therefore, marine vent systems may play host to novel archaeal viruses not yet discovered.

In total, 11 different virus families were found in the marine vent assemblage, which is higher than any of the other viromes compared in this analysis, with the exception of the Arctic Ocean (Fig. 2). This supports the notion that a wide range of viral types is present in the marine vent viral assemblage.

Marine vent virome assembly

Assembly of the marine vent virome yielded several large contigs. Figure 3 shows the mean coverage and length of

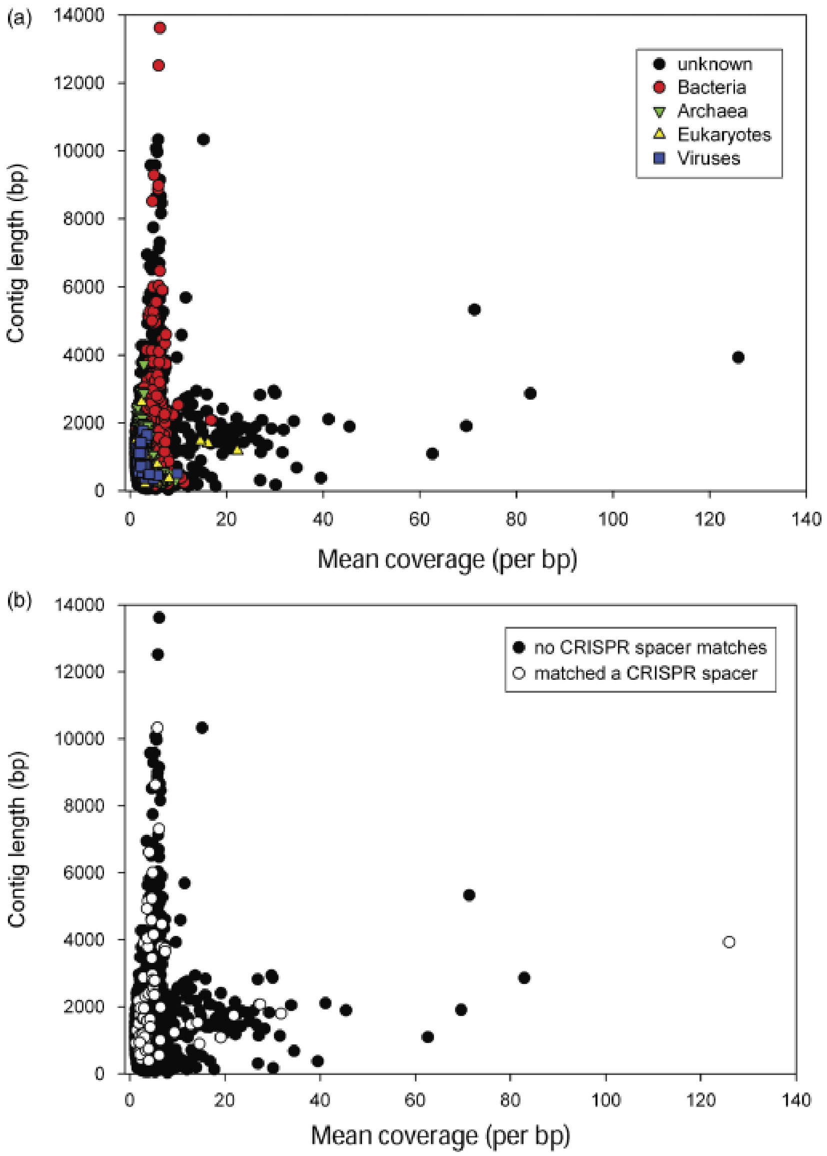

Fig. 3. Assembly of marine hydrothermal vent virome reads. Contigs were assembled using GENEIOUS (see Materials and methods). Only contigs containing three or more reads are shown. (a) Contigs labeled according to domain. Reads were assigned taxa by comparison with the SEED database; contigs were labeled according to the most common taxonomic grouping among constituent reads. (b) Contigs labeled according to whether the contig contained a read with a 100% identity alignment of at least 20 bp to a CRISPR spacer.

each contig. Many of the longest contigs in the vent virome contained reads matching bacterial genes (Fig. 3a). However, the longest contigs in the virome had a relatively low mean coverage. Reads with the highest mean coverage tended to be slightly shorter and contained reads with no matches to the SEED database. One possible explanation for this pattern is that shorter reads with higher coverage were derived from viral genomes, whereas the longer reads with a low mean coverage were derived from bacterial or archaeal genomes. Contigs with high coverage tended to contain reads with no matches to existing databases (Fig. 3a). A BLAST search of the contig with the highest coverage revealed hits only to short segments at each end of the contig, most of which corresponded to DNA ligases, suggesting that these high-coverage contigs were viral.

Additionally, sequenced genomes from a range of viral types have been found to contain sequences with high similarity to archaeal or bacterial genes (see e.g. Mann *et al*., 2003; Filée *et al*., 2007; Geslin *et al*., 2007; Fischer *et al*., 2010), and therefore, some contigs that were assigned to bacterial or archaeal taxa may actually lie within viral genomes.

Modeling richness and evenness of the viral assemblage

We modeled the alpha diversity of the marine vent virome and compared it with the diversity of six previously sequenced viromes: Bear Paw and Octopus Spring from Yellowstone National Park (Schoenfeld *et al*., 2008), for a high-temperature comparison, and four marine viromes: the Sargasso Sea, the Gulf of Mexico, the Bay of British Columbia, and the Arctic Ocean (Angly *et al.*, 2006), for a marine comparison. We remodeled the alpha diversity of each virome to maintain consistent parameters in the CIRCONSPECT and PHACCS models to enable a direct comparison. The modeled diversity values thus differ from original published results due to changes in both the CIRCONSPECT and the PHACCS software (F.E. Angly, pers. commun.). In each of the metagenomes sequenced with 454 technology (resulting in over 100 000 reads ranging from 100 to 300 bp), the trim size and discard size were set to 100, and the sample size in CIRCONSPECT was set to 10 000 reads. For the metagenomes sequenced using Sanger technology (resulting in only 8000–22 000 reads of about 1000 bp long), the trim size and discard size were set to 650 due to longer read lengths, but the 10 000 read sample size was only possible for one of the metagenomes. Therefore, comparison of richness across viromes sequenced by different technologies must be carried out cautiously, as the different read lengths and number of reads alter the output values. We sought to minimize error in the analysis while retaining reasonable read lengths and sample sizes, given the sequencing technology.

Our results (Table 1) indicate that the richness of the marine vent virome is comparable to that of other high-temperature or marine environments. Our results also show that the evenness of the marine vent virome is higher than that of any other virome, and thus the viral assemblage is not dominated by any single genotype. We also modeled the alpha diversity of the marine vent virome after removing all reads contained within contigs longer than 3000 bp in order to test whether the presence of long, low-coverage contigs (possibly derived from archaea and bacteria) influenced the results. The results, labeled (b) in Table 1, were not significantly altered.

Using the CRISPR spacer database to identify potential hosts

The analyses described above, which describe the viral types and the overall diversity of the viral assemblage and have been used in previous viral metagenomics studies (i.e. Angly *et al.*, 2006; Lopez-Bueno *et al.*, 2009), are nevertheless unable to provide specific information about what types of hosts are infected by the viral assemblage. To address this, we created a database of the CRISPR spacers contained within all sequenced organisms in the NCBI database, consisting of 81 260 spacers from 1083 genomes (http://dx.doi.org/10.5061/dryad.8826). As each CRISPR spacer is thought to be derived from a viral (or plasmid) sequence, this database serves as a repository of sequences from viruses that have infected these organisms. Moreover, because each of these spacer sequences is derived from the genome of a particular organism, we can match the viral sequence to the host. A similar CRISPR spacer search was conducted by Garrett *et al.* (2010) to query hyperthermophilic viral enrichments; however, rather than targeting specific hosts, this CRISPR spacer database was designed to identify potential hosts for the viruses represented by our assembled metagenomic sequences.

For our initial analysis, we conducted a BLASTN search between the CRISPR spacer database we generated and the marine vent virome, searching for 98% identity across the entire spacer sequence. Zero matches were found with these parameters, which attests to the diversity of viral sequences and the speed at which they mutate.

However, phage genomes are known to be mosaic in nature (Hendrix *et al.*, 2000; Hendrix, 2003), and it is thought that viruses can evade the CRISPR system by scrambling their sequences through the process of recombination (Andersson & Banfield, 2008). Thus, we searched for 100% alignments of CRISPR spacers across a portion of the spacer sequence rather than the full sequence, choosing as our cutoff 100% alignment across at least 20 base pairs. This 20 base pair cutoff was chosen to be lenient enough to find matches that are significant, but stringent enough to preclude false matches to CRISPR spacers.

Control dataset: comparing spacers within the database

To test the significance of these parameters, spacers from all of the sequenced bacteria and archaea in our CRISPR spacer database were compared with each other with BLASTN, searching for matches of 100% identity across 20 bp. The set of CRISPR spacers within each of the 578 CRISPR-containing archaeal and bacterial genomes was compared with the set of CRISPR spacers in each of the other 578 genomes. Of the 166 753 unique genome comparisons, 262 had one or more matching spacers at 100% identity across 20 bp. Of these, 249 (95%) were between spacers from genomes of the same genus, and of these, 155 (63%) were spacer matches between spacers from genomes of the same species. This provides strong evidence that a sequence with a match to a CRISPR spacer at this level (100% identity across 20 bp) is most likely derived from a virus that infected a host of the same genus or species as the CRISPR spacer it matches. These results are summarized in a resemblance matrix in Table S2.

Querying the marine vent virome with the CRISPR spacer database

When comparing the CRISPR spacer database with the marine vent virome, a total of 290 different spacers out of 81 260 spacers in the database (0.36%) had a match to the marine vent virome at 100% identity across 20 base pairs. Three hundred and eighty-two different reads out of 228 698

(0.167%) in the marine vent virome contained a match to one of these CRISPR spacers. While these reads represent a low percentage of the total, the conservative parameters were retained to minimize the possibility of false matches. At this stringency level, there is a $(0.25)^4$, or 9.09×10^{-13}, chance that a random sequence would match, and therefore out of 228 698 reads, one would expect 2.08×10^{-7} reads to have a match. Thus, the result of 382 different virome reads with a match to a spacer cannot be due to random sequence similarity.

To compare this result with that of cellular metagenomes, we conducted the same BLAST search of the CRISPR spacer database against several other cellular metagenomes taken from the MG-RAST database. These metagenomes represent a range of GC content, read length, and number of reads. The results are shown in Table S5 in terms of the ratio of spacer matches to base pairs to normalize for differences in read length and number. The results show that the average ratio of matches to base pairs is 4.027×10^{-6} for the cellular metagenomes, whereas it is 6.27×10^{-6} for the marine vent virome. This suggests that there was a higher proportion of spacer hits to this virome than to these cellular metagenomes, despite the presence of CRISPR loci and possible viral contamination in the cellular metagenomes. To demonstrate the presence of CRISPR loci in the cellular metagenomes, the numbers of CRISPR-associated (*cas*) genes identified in each metagenome are also listed in this table. While our results indicate that the number of *cas* genes identified in a given metagenome can vary widely, these results do indicate that CRISPRs were present in the cellular metagenomes and may have contributed to the total number of spacer matches. While some *cas* genes were found in the marine vent virome as well, the reads matching these *cas* genes fell on only five contigs consisting of three or more reads; of these, each of these *cas*-gene-containing contigs had a relatively low coverage (maximum 3.2).

To further test for the presence of contaminating bacterial or archaeal reads in the marine vent virome that may have contained CRISPR loci, we searched for evidence of CRISPR direct repeats in the vent virome to act as a proxy for CRISPR loci derived from cellular genomes. CRISPR direct repeat sequences, unlike spacer sequences, do not correspond to viral sequences and are much more highly conserved among loci and among taxa (Kunin *et al*., 2007). The marine vent virome contained only 58 reads with a match to a CRISPR repeat sequence, compared with 382 reads with a match to a CRISPR spacer. Here, a ‘match’ is again defined as 100% identity over at least 20 base pairs. If the vent virome had a high proportion of contaminating CRISPR loci from bacterial or archaeal genomes, we would have expected a relatively higher number of matches to CRISPR direct repeats.

Figure 3b shows which contigs from the marine vent virome contained a read with a match of 100% identity across 20 bp to a CRISPR spacer in our database. While some were found within reads assigned to bacteria, 76% of the contigs with matches to CRISPR spacers contained a majority of reads with no match to the SEED database (where a ‘match’ here is defined as a TBLASTX hit with an *e*-value cutoff of 10^{-5}). Additionally, nearly half of the reads with matches to the spacer database belonged to these unidentified contigs, which contain only about one-third of the total virome reads. This supports the notion that the reads with matches to the CRISPR spacer database represent viral sequences.

Identification of potential hosts for the marine vent viral assemblage

To identify potential hosts for the marine vent viral assemblage, we grouped all of the spacers matching the virome according to the taxonomic group of the strain from which they were derived. Table 2 depicts the distribution of BLAST hits between the CRISPR spacers of each group and the marine vent virome. Most notable about the results is the wide range of both archaeal and bacterial taxonomic groups that had CRISPR spacers matching the marine vent virome,

Table 2. CRISPR spacer database matches in the marine vent virome

Group	Number of matches in vent virome to group	Number of spacers from the group in the database
Firmicutes	109	7796
Bacteroidetes/Chlorobi	78	1556
Gammaproteobacteria	64	5076
Euryarchaeota	63	4195
Crenarchaeota	33	4038
Chloroflexi	24	3188
Thermotogae	21	1392
Cyanobacteria	14	1935
Aquificae	13	519
Actinobacteria	12	3662
Betaproteobacteria	11	1301
Deinococcus-Thermus	7	453
Deltaproteobacteria	5	2365
Alphaproteobacteria	3	1368
Dictyoglomi	3	245
Epsilonproteobacteria	2	302
Fusobacteria	2	47
Nanoarchaeota	2	41
Nitrospirae	2	182
Thermobaculum	2	206
Deferribacteres	1	19
Planctomycetes	1	30
Spirochaetes	1	88

The first column lists the groups with spacers having a match to the marine vent virome; the second column lists the number of hits in the vent virome to spacers in that group; and the third column lists the total number of spacers from that group in the CRISPR spacer database.

with no single taxonomic group dominating. This suggests that the viruses in the vent assemblage have the potential to infect a wide range of taxonomic groups. The groups with the most matches between their CRISPR spacers and the marine vent virome were the *Firmicutes*, the *Bacteroidetes*/*Chlorobi*, and the *Gammaproteobacteria*; however, the high number of hits from these groups may be attributed partially to the high number of CRISPR spacers from these groups in the database. Interestingly, a relatively small percentage of spacers from the *Proteobacteria* (particularly *Alpha*-, *Beta*-, *Gamma*-, and *Deltaproteobacteria*) had matches to the marine vent virome, despite the large number of spacers from *Proteobacteria* in the spacer database, and despite the prevalence of these taxa at this site (Huber *et al*., 2007). Therefore, this result may reflect a surprising lack of viruses infecting *Proteobacteria* in our sample. Matches to archaeal CRISPR spacers are common within the marine vent virome (Table 2), despite the relative absence of known archaeal virus families in the virome (Fig. 2). However, known archaeal virus families have predominantly been cultured from terrestrial hot springs. Because the archaeal domain is known to be well represented in marine hydrothermal fluids (Huber *et al*., 2007), these data suggest that the archaeal viruses present in the marine vent virome are unlike those found in terrestrial hot springs and were therefore undetectable with traditional BLAST searches, but may have been detected by our CRISPR spacer analysis.

To more closely examine species known to be endemic to marine hydrothermal vent ecosystems, we determined the relative numbers of matches between the marine vent virome and the CRISPR spacers from genomes of vent isolates. While these spacer matches do not necessarily indicate that viruses infecting these specific species have been identified, we can state that sequences similar to those from viruses that have infected these species in the past are present in this virome. It is interesting to note that the results (Table 3) show that two-thirds of the vent isolate spacer hits were from *Methanocaldococcus* species, despite the fact that *Methanocaldococcus* strains only comprise about 15% of sequenced vent isolates.

Table 3. CRISPR spacer database matches in the marine vent virome, focusing only on species endemic to hydrothermal vents

Species	Number of matches in vent virome to species	Number of spacers from species in the database
Methanocaldococcus vulcanius M7	18	219
Methanocaldococcus sp. FS406-22	5	238
Hyperthermus butylicus DSM 5456	3	94
Methanocaldococcus jannaschii DSM 2661	3	177
Thermococcus kodakarensis KOD1	3	75
Methanocaldococcus fervens AG86	2	77
Nanoarchaeum equitans Kin4-M	2	41
Persephonella marina EX-H1	2	38
Thermococcus onnurineus NA1	2	118
Pyrobaculum aerophilum str. IM2	1	131
Pyrococcus horikoshii OT3	1	149

The first column lists the vent species with spacers having a match to the marine vent virome; the second column lists the number of hits in the vent virome to spacers in that species; and the third column lists the total number of spacers from that species in the CRISPR spacer database.

CRISPR spacers in *Methanocaldococcus* genomes

The high abundance of CRISPR spacer matches from *Methanocaldococcus* species can be attributed in part to the high number of CRISPR spacers in individual *Methanocaldococcus* genomes. For example, the genome of *Methanocaldococcus* sp. FS406-22, a hyperthermophilic methanogen that fixes nitrogen at 92 °C (Mehta & Baross, 2006), has the highest number of CRISPR loci of all sequenced isolates to date: 23 were identified using the CRISPFINDER application (Grissa *et al*., 2007a, b), and 20 were identified using the CRT (Bland *et al*., 2007). *Methanocaldococcus vulcanius* M7 and *Methanocaldococcus jannaschii* DSM 2661, also isolated from marine hydrothermal vents, contain the second and the third highest numbers of CRISPR loci of all sequenced isolates, respectively. As described above, no spacers were shared among genomes. Interestingly, nearly all spacers (94–99%) were also unique within each thermophilic methanogen genome, even in those containing high numbers of CRISPR loci. In other words, almost none of these genomes contained a duplicate CRISPR spacer. It seems unlikely that typical recombination and mutation events could cause this level of diversity in the spacer sequences, but not in the CRISPR repeats, all of which are identical or nearly identical within each CRISPR locus. Instead, it is likely that these methanogens gained their multitude of CRISPR spacers through distinct infection events. It is not clear whether spacer diversity correlates with infecting viral diversity, however, because of the apparent semi-random nature of the CRISPR mechanism. Protospacer adjacent motifs (PAMs) are thought to act as recognition sequences for CRISPR genes. Most PAMs are two to three nucleotides long, resulting in a large number of potential spacer sites on a viral genome (Mojica *et al*., 2009). Therefore, the lack of duplicate sequences indicates a large number of distinct infection events, but it is unclear whether it also implies a high diversity of infecting viruses. While the reasons for the abundance of CRISPR loci in thermophilic methanogens are unknown, it is part of a larger trend in thermophiles that is discussed further below.

Nevertheless, while the abundance of CRISPRs in *Methanocaldococcus* genomes is striking, it does not fully explain the large percentage of matches between the marine vent virome and *Methanocaldococcus* spacers. CRISPR spacers from *Methanocaldococcus* species represent 28% of the CRISPR spacers from vent isolates (Table S3), yet they represented over 50% of the vent isolate spacer matches to the marine vent virome. This suggests that viruses of *Methanocaldococcus* species were particularly prevalent in this diffuse flow sample.

CRISPR spacers as a probe of the host thermal regime

We next performed a BLASTN search of our CRISPR spacer database with five other previously published viromes, with a particular emphasis on the thermal regime. We compared the CRISPR spacer database with four marine viromes and two Yellowstone hot springs viromes (combined together for this analysis). Only a single match with 100% similarity over the full length of the spacer sequence was found: a spacer from *Synechococcus* sp. JA-2-3B'a(2-13), isolated from Octopus Spring in Yellowstone, had a match to the virome from the same site. No other perfect matches between the CRISPR spacer database and any of these viromes were found.

We next searched for 100% alignments of 20 bp or above, as before. A total of 901 out of 81 260 spacers, or 1.11% of the spacers in the CRISPR database, had a match to one or more of the six viromes. We grouped these hits according to the thermal regime of the host from which the spacer was derived (Fig. 4). Our results show that 1.84% of spacers specific to vent isolates had a match to the marine vent virome. This constitutes a higher percentage than to viromes in other environments, suggesting that there is a unique vent virus 'signature,' perhaps due to a particular sequence or set of sequences that is shared among vent viruses. Notably, the vent isolates included in this analysis were isolated from marine hydrothermal vents around the globe, indicating that this vent 'signature' is not unique to a particular marine vent location or depth.

Thermophilic strains also had a high percentage of spacer matches to the vent virome (0.98%) relative to other viromes. Several *Sulfolobus* spacers, for example, had matches to the marine vent virome despite being endemic to terrestrial hot springs. Again, this is an interesting contrast to the relative lack of marine vent virome read matches to archaeal virus families found in terrestrial hot springs (Fig. 2). It is possible that the *Sulfolobus* spacers with matches to the marine vent virome are derived from viruses that have not been isolated or sequenced, and therefore, had no matches in existing databases. As a natural 'library' of viral infection, the CRISPR spacer dataset does not rely on isolation of individual virus–host systems and is therefore able to identify potential hosts for viruses in the assemblage with no cultured relatives.

Finally, the marine vent virome had a relatively high proportion of matches to spacers from nonvent and nonthermophilic organisms (Fig. 4). This result highlights the multitude of microenvironments present in marine diffuse flow hydrothermal systems. Because gradients in temperature, pH, chemical composition, and mineralogy are known to dominate vent systems (Baross & Hoffman, 1985), our vent

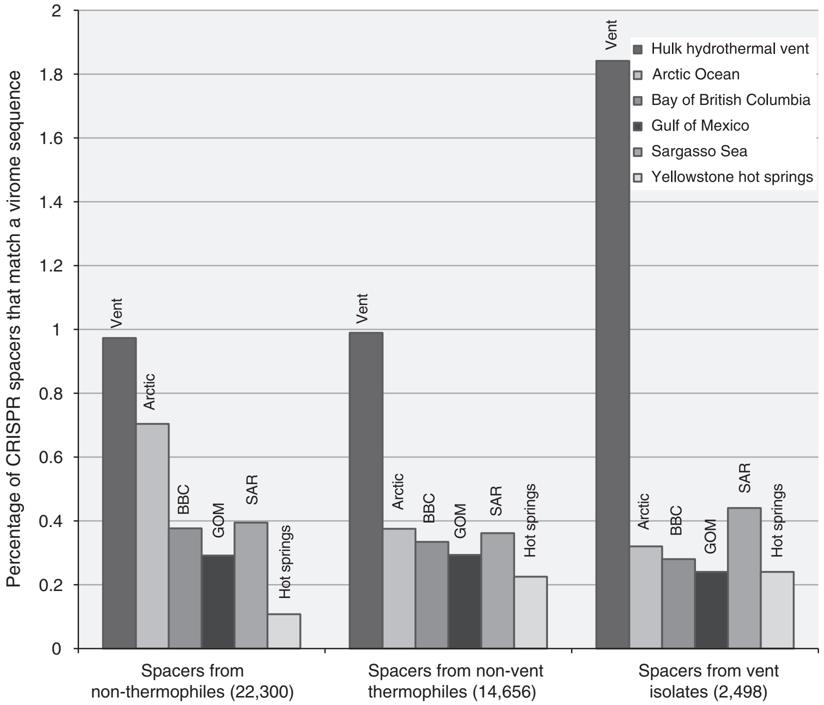

Fig. 4. CRISPR spacer matches to other marine or hot springs viromes. CRISPR spacers were grouped as derived from vent isolates, nonvent thermophiles, and all other isolates. Sequence similarity searches were performed with BLASTN, and a 'match' was defined as a 100% match across an alignment of 20 base pairs or greater. Numbers under each category on the *x*-axis indicate the number of spacers in each group.

fluid sample likely was a composite of fluids that were exposed to a variety of environmental conditions in the subsurface. These fluids may have been exposed to temperatures ranging from that of ambient seawater, at around 2 °C, up to 135 °C or possibly even higher. The pH could have ranged from that of ambient seawater, between pH 7 and 8, to much lower pH values typical of high-temperature hydrothermal fluids, at around pH 2 or 3. Therefore, it is not surprising that the diverse microbial communities inhabiting vents play host to diverse viral communities as well.

Correlation of CRISPR locus abundance per genome and growth temperature

Our analyses indicated that CRISPR spacers from thermophiles are common in all viromes, which may be attributed to the abundance of CRISPR spacers from thermophiles in our database. Closer examination of this trend shows that thermophilic strains, on average, have higher numbers of CRISPR loci in their genomes than mesophiles. Early literature on CRISPR loci made a brief note of this trend (Makarova *et al*., 2003, 2006), but it has not yet been given extensive treatment. This is an important consideration when using CRISPR spacers for metagenomic analysis, however, because this indicates that CRISPR spacers are not distributed evenly among bacteria and archaea. Any attempts to quantify viral hosts using the CRISPR spacer database must bear this in mind.

To examine this trend more explicitly, we calculated numbers of CRISPR loci per genome (as determined by CRT) and binned the isolates according to growth temperature. The genomes of bacteria and archaea isolated from high-temperature environments contain higher numbers of CRISPR loci, on average, than mesophilic or psychrophilic organisms (Fig. 5a). The trend is evident in both the bacteria and the archaea.

However, the number of spacers contained within each CRISPR locus is not constant: while most CRISPR loci contain an average of 30–40 spacers, some contain as few as one or two, while others, such as a CRISPR locus in *Haliangium ochraceum*, contain as many as 600 spacers in a single locus. We calculated the total number of spacers encoded within all CRISPR loci for each genome and correlated this with growth temperature, as before. The trend of increased CRISPR locus abundance in thermophiles (Fig. 5a) held for CRISPR spacers as well (Fig. 5b).

This trend is not an artifact of high CRISPR abundance in specific taxonomic groups. For example, this trend can be seen across thermal groups in the methanogens (Fig. 5c). The trend also holds within single genera: for example, of the 11 sequenced *Synechococcus* isolates, only three possess CRISPR loci. Two of these three CRISPR-possessing strains are the only thermophilic *Synechococcus* isolates with sequenced genomes.

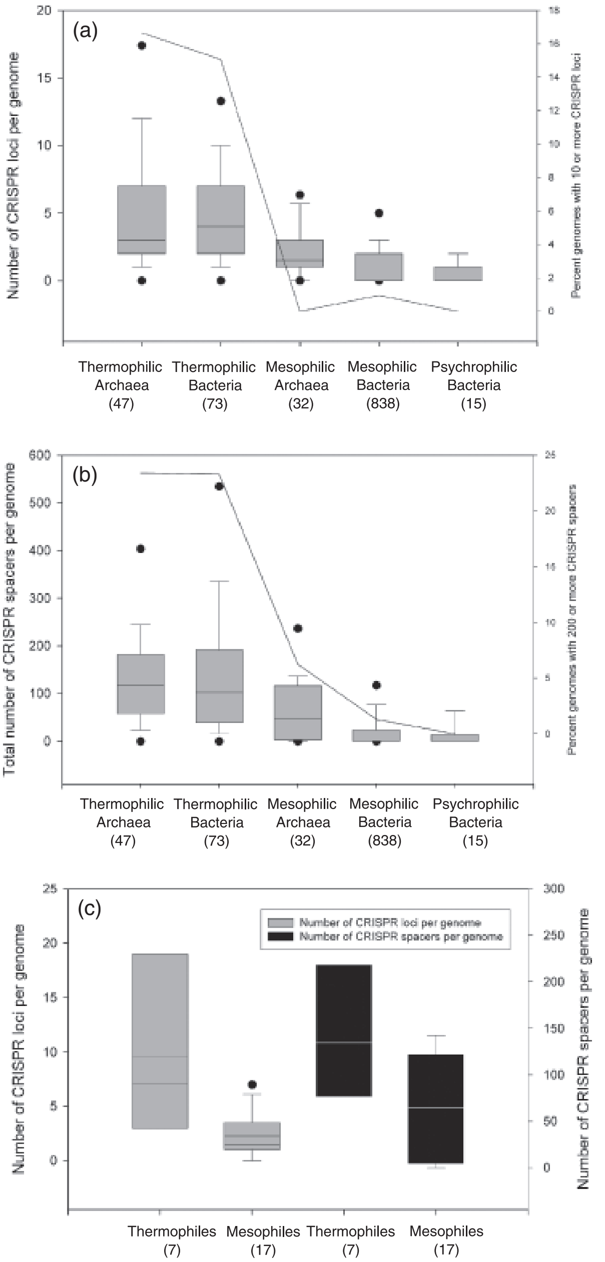

Fig. 5. Abundances of CRISPR loci and spacers in different thermal groups. Numbers below temperature categories list the number of genomes in that category. Box boundaries represent the 25th and 75th percentiles; a line within the box marks the median. Error bars above and below the box indicate the 90th and 10th percentiles. Outlying points represent the fifth and 95th percentiles. The dashed line shows the percent of genomes within each group containing 10 or more CRISPR loci or 100 or more CRISPR spacers. (a) Number of CRISPR loci per genome; (b) number of CRISPR spacers per genome; and (c) number of CRISPR loci and spacers per genome in methanogens only.

Outliers do exist in each category (Table S6). Most notably, a small number of mesophilic bacteria contain relatively large numbers of spacers in their genomes. While these cases are unusual and should be studied in further detail, they constitute a small minority of mesophilic bacteria: 85% of the over 800 sequenced mesophilic bacteria have between 0 and 2 CRISPR loci.

The reasons for this temperature trend are not yet clear. It is unlikely that the CRISPR overabundance in thermophiles is due to a higher diversity among viruses infecting thermophilic hosts, as our diversity modeling results indicate that this is not universally the case (Table 1). We also do not expect that the high abundance of CRISPR loci and spacers in thermophiles can be attributed to higher rates of infection in high-temperature environments, as studies thus far indicate that virus-to-cell ratios are not necessarily higher in the vent and hot spring environments than in other environments (Srinivasiah *et al.*, 2008). It is possible that CRISPRs are the predominant immunity system in thermophiles, whereas mesophiles favor other types of immunity mechanisms; alternatively, it is possible that the abundance of CRISPR loci in thermophiles is the result of high rates of horizontal gene transfer at high temperatures. However, our current understanding of viral immune systems across the bacteria and archaea and of horizontal gene transfer in different thermal regimes is not thorough enough to distinguish between these possibilities at present.

Conclusion

Our results indicate that the 1840 genotypes present in the viral assemblage of this marine diffuse flow hydrothermal vent represent a range of viruses with the potential to infect mesophilic and thermophilic hosts across both the archaeal and the bacterial domains. The high evenness of the vent viral assemblage indicates that each of the viral types is fairly equally represented. Therefore, it is likely that viruses infecting a diverse range of hosts are relatively evenly represented in the viral assemblage. This is reflective of the dynamic hydrothermal vent environment, which enables potentially frequent interactions among diverse and extreme microbial communities and their associated viral communities. No other environment possesses the range of physiochemical gradients that characterizes the subsurface vent system, nor the means by which to bring such a wide range of taxonomic groups into close contact. Moreover, the abundance of CRISPR spacers in thermophiles, especially in vent methanogens, suggests that viruses play a unique role in the vent environment; yet, the sheer diversity of these spacers attests to the rapid rates of virus–host evolution in these environments. This study, by pairing traditional metagenomic analyses with a novel comparison to a comprehensive CRISPR spacer database, has provided the first insight into the infection potential of the viral assemblage at vents, and paves the way for further studies into how these viruses impact the ecology and evolution of their microbial hosts.

Acknowledgements

We thank Chief Scientist Jim Holden, Dave Butterfield, the captain and crew of the R/V *Atlantis*, the pilots and crew of DSV *Alvin*, and Sanjoy Som for assistance in virome sample collection. Tor Bjorklund and Paul Johnson assisted in designing and constructing the barrel sampler used for sample collection, and Tim Shanks provided an elevator for barrel sampler deployment. Florent Angly and Lawrence Gabriel assisted with CIRCONSPECT and PHACCS, and Clara Fuchsman provided a database of bacterial phenotypes. We would like to thank Jody Deming, Victoria Meadows and two anonymous reviewers for helpful comments on the manuscript. This research was funded in part by the Gordon and Betty Moore Foundation through a grant to the Broad Institute. Sample G2810 was sequenced at the Broad Institute. Funding was also provided by an NSF Graduate Research Fellowship, NSF IGERT, the ARCS program, and a NASA Astrobiology Institute grant through Cooperative Agreement NNA04CC09A to the Geophysical Laboratory at the Carnegie Institution for Science.

References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andersson AF & Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**: 1047–1050.

Angly FE, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J & Rohwer F (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**: 41.

Angly FE, Felts B, Breitbart M *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

Angly FE, Willner D, Prieto-Davó A *et al.* (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5**: e1000593.

Baross JA & Hoffman SE (1985) Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins Life Evol B* **15**: 327–345.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA & Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–1712.

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K & Wommack KE (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microb* **73**: 7629–7641.

Bland C, Ramsey T, Sabree F, Lowe M, Brown K, Kyrpides N & Hugenholtz P (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**: 209.

Bolotin A, Quinquis B, Sorokin A & Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**: 2551–2561.

Breitbart M & Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**: 278–284.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F & Rohwer F (2002) Genomic analysis of uncultured marine viral communities. *P Natl Acad Sci USA* **99**: 14250–14255.

Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV & Van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**: 960–964.

Desnues C, Rodriguez-Brito B, Rayhawk S *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452**: 340–343.

Dinsdale EA, Edwards RA, Hall D *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Drummond A, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T & Wilson A (2009) *Geneious v4.7*. Biomatters Ltd, Auckland.

Filée J, Siguier P & Chandler M (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet* **23**: 10–15.

Fischer MG, Allen MJ, Wilson WH & Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *P Natl Acad Sci USA* **107**: 19508–19513.

Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH & Moineau S (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**: 67–71.

Garrett RA, Prangishvili D, Shah SA, Reuter M, Stetter KO & Peng X (2010) Metagenomic analyses of novel viruses and plasmids from a cultured environmental sample of hyperthermophilic neutrophiles. *Environ Microbiol* **12**: 2918–2930.

Geslin C, Gaillard M, Flament D, Rouault K, Le Romancer M, Prieur D & Erauso G (2007) Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from *Pyrococcus abyssi*. *J Bacteriol* **189**: 4510–4519.

Grissa I, Vergnaud G & Pourcel C (2007a) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**: 172.

Grissa I, Vergnaud G & Pourcel C (2007b) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52–W57.

Haft DH, Selengut J, Mongodin EF & Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**: e60.

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM & Terns MP (2009) RNA-guided RNA cleavage by a CRISPR RNA–Cas protein complex. *Cell* **139**: 945–956.

Heidelberg JF, Nelson WC, Schoenfeld T & Bhaya D (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PloS One* **4**: e4169.

Held NL & Whitaker RJ (2009) Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ Microbiol* **11**: 457–466.

Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ & Planet PJ (2010) CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS One* **5**: e12988.

Hendrix RW (2003) Bacteriophage genomics. *Curr Opin Microbiol* **6**: 506–511.

Hendrix RW, Lawrence JG, Hatfull GF & Casjens S (2000) The origins and ongoing evolution of viruses. *Trends Microbiol* **8**: 504–508.

Hovarth P & Barrangou R (2010) CRISPR/Cas, the immune system of Bacteria and Archaea. *Science* **327**: 167–170.

Huber JA, Butterfield DA & Baross JA (2003) Bacterial diversity in a subseafloor habitat following a deep-sea volcanic eruption. *FEMS Microbiol Ecol* **43**: 393–409.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA & Sogin ML (2007) Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Kim K, Chang H, Nam Y, Roh SW, Kim M, Sung Y, Jeon CO, Oh H & Bae J (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microb* **74**: 5975–5985.

Kristensen DM, Mushegian AR, Dolja VV & Koonin EV (2009) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* **18**: 11–19.

Kunin V, Sorek R & Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**: R61.

Labrie SJ, Samson JE & Moineau S (2010) Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**: 317–327.

Lopez-Bueno A, Tamames J, Velazquez D, Moya A, Quesada A & Alcami A (2009) High diversity of the viral community from an Antarctic lake. *Science* **326**: 858–861.

Makarova K, Grishin N, Shabalina S, Wolf Y & Koonin E (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**: 7.

Makarova KS, Wolf YI & Koonin EV (2003) Potential genomic determinants of hyperthermophily. *Trends Genet* **19**: 172–176.

Mann NH, Cook A, Millard A, Bailey S & Clokie M (2003) Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**: 741.

Marraffini LA & Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**: 1843–1845.

Marraffini LA & Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**: 181–190.

Mehta MP & Baross JA (2006) Nitrogen fixation at 92 °C by a hydrothermal vent archaeon. *Science* **314**: 1783–1786.

Meyer F, Paarmann D, D'Souza M *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Mojica FJ, Díez-Villaseñor C, García-Martínez J & Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**: 174–182.

Mojica FJM, Diez-Villasenor C, Garcia-Martinez J & Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**: 733–740.

Ortmann AC & Suttle CA (2005) High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep-Sea Res Pt I* **52**: 1515–1527.

Pourcel C, Salvignol G & Vergnaud G (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**: 653–663.

Prangishvili D, Forterre P & Garrett RA (2006) Viruses of the Archaea: a unifying view. *Nat Rev Microbiol* **4**: 837–848.

Santos F, Yarza P, Parro V, Briones C & Antón J (2010) The metavirome of a hypersaline environment. *Environ Microbiol* **12**: 2965–2976.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M & Mead D (2008) Assembly of viral metagenomes from Yellowstone hot springs. *Appl Environ Microb* **74**: 4164–4174.

Schrenk MO, Kelley DS, Delaney JR & Baross JA (2003) Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. *Appl Environ Microb* **69**: 3580–3592.

Snyder JC, Bateson MM, Lavin M & Young MJ (2010) Use of cellular CRISPR (Clusters of Regularly Interspaced Short Palindromic Repeats) spacer-based microarrays for detection of viruses in environmental samples. *Appl Environ Microb* **76**: 7251–7258.

Sorek R, Kunin V & Hugenholtz P (2008) CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181–186.

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T & Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res Microbiol* **159**: 349–357.

Suttle CA (2005) Viruses in the sea. *Nature* **437**: 356–361.

Thurber RV, Haynes M, Breitbart M, Wegley L & Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**: 470–483.

Tyson GW & Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**: 200–207.

van der Oost J, Jore MM, Westra ER, Lundgren M & Brouns SJJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**: 401–407.

Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, Winget D & Wommack KE (2008a) Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *ISME J* **2**: 1112–1121.

Williamson SJ, Rusch DB, Yooseph S *et al.* (2008b) The Sorcerer II Global Ocean Sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3**: e1456.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Image of the sample intake funnel of the barrel sampler atop a sulfide structure on the side of Hulk vent in the Main Endeavour Field.

Table S1. Summary of physical, chemical, and biological attributes of Hulk vent in the Main Endeavour Field.

Table S2. Resemblance matrix used to compare the set of all CRISPR spacers in each genome in the NCBI database against the set of CRISPR spacers in every other genome.

Table S3. List of all sequenced strains categorized as hydrothermal vent isolates for this study.

Table S4. Percentages of sequences with matches to the SEED database, with a maximum *e*-value of 10^{-5}.

Table S5. Results of comparing the CRISPR spacer database against cellular metagenomes and against the marine vent virome.

Table S6. Top five organisms with the highest number of CRISPR loci per genome for each of the temperature categories listed in Fig. 1.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.