Diverse viruses have restricted biogeography in deep-sea hydrothermal vent fluids

Elaina Thomas[†1,2], *Rika Anderson*[*†1], *Viola Li*[1], *Jenni Rogan*[1], *Julie A. Huber*[3]

Affiliations:
[1] Biology Department, Carleton College. Northfield, MN
[2] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology. Cambridge, MA
[3] Marine Chemistry & Geochemistry, Woods Hole Oceanographic Institution. Woods Hole, MA

[*] To whom correspondence should be directed. randerson@carleton.edu
[†] These authors contributed equally to this work.

Running title: Viruses in deep-sea hydrothermal vents

Abstract

In the ocean, viruses impact microbial mortality, regulate biogeochemical cycling, and alter the metabolic potential of microbial lineages. At deep-sea hydrothermal vents, abundant viruses infect a wide range of hosts among the archaea and bacteria that inhabit these dynamic habitats. However, little is known about viral diversity, host range, and biogeography across different vent ecosystems, which has important implications for how viruses manipulate microbial function and evolution. Here, we examined viral diversity, viral and host distribution, and viral-host interactions in venting fluids from two geographically distant hydrothermal systems, the Mid-Cayman Rise in the Caribbean Sea and Axial Seamount in the Pacific Ocean. Analysis of viral sequences and CRISPR spacers revealed highly diverse viral assemblages and abundant lysogenic viruses, with 40% of metagenome-assembled genomes encoding a putative prophage. Network analysis revealed that viral host range was relatively narrow, with very few viruses infecting multiple microbial lineages. Viruses were largely endemic to individual vent sites, indicating restricted dispersal, and in some cases viral assemblages persisted over time. Thus, while the viruses at deep-sea hydrothermal systems play an important role in driving the evolution and ecology of resident vent microbial communities, their influence is highly localized to specific regions and taxa.

Introduction

Deep-sea hydrothermal vents are regions on the seafloor where high-temperature hydrothermal vent fluid is created from water-rock reactions deep within the oceanic crust, mixing with seawater beneath and at the seafloor to create a dynamic, gradient-dominated habitat that supports diverse microbial communities. These low-temperature diffuse vent fluids are hotspots of primary production in the deep ocean, dominated by diverse chemolithoautotrophic bacteria and archaea carrying out a variety of metabolisms utilizing hydrogen, sulfur compounds, nitrate, and methane [1–7]. An important, though understudied, driver of microbial diversity and evolution in hydrothermal systems is viruses. Viruses are major sources of microbial mortality in marine systems and play key roles in mediating biogeochemical cycling and shaping microbial community structure [8–11]. At deep-sea hydrothermal vents, viruses are abundant [12] and infect a wide range of microbial hosts [13]. Lysogenic viruses, which integrate into the genomes of their microbial hosts, are particularly abundant in hydrothermal fluids compared to other marine habitats [14, 15]. The high abundance of lysogenic viruses suggests that there are unique attributes to the vent environment that influence viral infection strategies. Viruses can "metabolically reprogram" their microbial hosts via the introduction and expression of auxiliary metabolic genes (AMGs) [16], which has profound impacts on the ecology and evolution of microbial populations. Like horizontally transferred genes, integrated prophage with AMGs can alter the functional potential of a given organism, allowing it to adapt to changing conditions or expand to new ecological niches. Viruses in hydrothermal habitats have been found to encode AMGs [17–19], and thus can manipulate the metabolic potential of microbial populations in hydrothermal vents.

Viruses also function as vectors of horizontal gene transfer via transduction. Previous work has suggested that horizontal gene transfer is particularly prevalent among microbes inhabiting high-temperature environments [20–22]. Transposases, which catalyze the movement

of mobile genetic elements within and between genomes [23, 24], are especially abundant in hydrothermal vent sites [18, 25] Given the abundance of viruses and the inferred high rates of transduction at deep-sea vents, it is likely that viruses are an important vector for horizontal gene transfer in these dynamic systems. Therefore, a clear understanding of which viruses infect which hosts, and what genes those viruses carry, can provide insight into highways of gene sharing in the deep sea.

One large gap in our understanding of eco-evolutionary dynamics in the deep sea is how broadly viruses are distributed within and between vent systems, and how viral interactions shift across hydrothermal vent types. Previous work has shown that microbial communities exhibit high endemism locally [4, 7, 26, 27], but some cosmopolitan species are distributed globally [28–30]. However, less is known about the geographic distribution of viral populations. If viruses exhibit restricted biogeographic distribution, this would limit their role as vectors of gene flow across and between hydrothermal systems. Microbial populations and their viruses are also limited by environmental conditions: hydrothermal systems hosted in basalt rocks are characterized by metal-enriched, low pH fluids up to 400˚C [31]. In contrast, hydrothermal systems hosted in peridotite are influenced by serpentinization and feature organic carbon-enriched, high pH fluids with slightly lower temperature [32]. Microbial populations in basalt-hosted and peridotite-hosted vents exhibit distinct patterns of genomic variation [33], suggesting that microbes are subject to different selection pressures depending on the vent type, but we know little about the role viruses play in driving those differences, nor how viral diversity varies across hydrothermal systems.

To gain further insight into viral diversity, biogeography and host range across hydrothermal systems, we compared viral sequences recovered from microbial metagenomes collected from two hydrothermal regions in two distinct ocean basins: Axial Seamount, a submarine volcano located on the Juan de Fuca Ridge in the Pacific Ocean at ~1520 meters depth, and the Mid-Cayman Rise, an ultraslow spreading ridge in the Caribbean Sea. Axial Seamount is a basalt-hosted, magma-driven system with fluids that are low in pH and high in carbon dioxide and hydrogen sulfide [34]. In contrast, the Mid-Cayman Rise hosts two geologically and geochemically distinct hydrothermal vent fields in close proximity to each other: Piccard hydrothermal field, located in basalt rocks along the ridge axis, is the deepest hydrothermal field discovered to date (~4950 meters depth) and is characterized by fluids that are acidic with high hydrogen and hydrogen sulfide content, whereas Von Damm vent field, located approximately 20 km away on a nearby massif at ~2350 meters depth, is influenced by serpentinization and is characterized by fluids that are high in hydrogen, methane, and small carbon compounds [35–38]. Previous work found distinct community composition but similar functional potential in the microbial communities at Piccard and Von Damm [7, 27, 33, 39, 40] spatially restricted but temporally stable at individual vent sites over time [4, 27]. Here, we compare viral sequences across venting fluids from both the Mid-Cayman Rise and Axial Seamount in a comparative survey of viral diversity and gene content across geographically distinct hydrothermal vent sites. We show that viral populations have severely restricted biogeographic distribution and host range in hydrothermal systems, limiting the capacity of viruses to act as vectors of gene flow between disparate hosts and vent locations.

Methods

Sample collection and DNA preparation and sequencing

Low-temperature diffuse flow fluid samples were collected from the Mid-Cayman Rise vent fields Piccard and Von Damm in January 2012 and June 2013 during research cruises on the R/V Atlantis and R/V Falkor, respectively. Sample locations, depth, and metagenomic data are described in Supplementary Table 1. The ROV Jason II and Mat Sampler were used to collect the 2012 Mid-Cayman Rise samples, as previously described in [41]. The 2013 Mid-Cayman Rise samples were collected using the SUPR version 2 sampler and HROV Nereus [42]. For microbial DNA collection, approximately 3 to 6 L of diffuse flow fluid were pumped through 0.22 µm Sterivex filters (Millipore). Shipboard, the filters were flooded with RNALater (Ambion), sealed with Luer Caps, stored in sterile Falcon tubes, and frozen at −80 °C. Sample collection and preservation are further described in Anderson et al. [33] and Reveillaud et al. [40]. Genomic DNA was extracted and metagenomic libraries constructed as described in Anderson et al. [33]. Sequencing was done on an Illumina Hi Seq 1000 at the W.M. Keck Facility in the Josephine Bay Paul Center at the Marine Biological Laboratory.

Diffuse flow fluid samples were collected from Axial Seamount in September 2013, August 2014, and August 2015 (approximately five months after the eruption of Axial Seamount) during research cruises on the R/V Falkor and R/V Thompson in 2013, R/V Brown in 2014, and R/V Thompson in 2015 (Supplementary Table 1). Diffuse flow samples were collected from four different vent fields at Axial (ASHES, International District, North Rift Zone, and Dependable) using the ROVs ROPOS and Jason. For collection of microbial DNA, 3 L of diffuse fluid was pumped through 0.22 µm, 47 mm GWSP filters (Millipore) and the filters were flooded with RNALater (Ambion) on the seafloor, as described in Fortunato et al. [4]. Fluids from a hydrothermal plume above Anemone and background seawater were collected using a Seabird SBE911 CTD and 10 L Niskin bottles and 3 L of the plume and seawater fluid was filtered through 0.22 µm Sterivex filters (Millipore). DNA was extracted from the filters using a phenol-chloroform method adapted from Crump et al. [43] and Zhou et al. [44]. The Ovation Ultralow Library DR multiplex system (Nugen) was used to prepare metagenomic libraries following the manufacturer instructions. DNA extraction and metagenomic library construction is further described in Fortunato and Huber [45]. The 2013 and 2014 samples were sequenced on an Illumina HiSeq 1000 and the 2015 samples on a NextSeq 500. Sequencing was done at the W.M. Keck sequencing facility at the Marine Biological Laboratory.

For all metagenomes, paired-end partially overlapping reads were merged and quality filtered using the illumina-utils package [46] using the iu-quality-filter-minoche flag, then assembled using idba-ud [47] v1.1.2 with default settings. All data from both sites is available in the European Nucleotide Archives archives under study accession number PRJEB15541 for the Mid-Cayman Rise and under study accession numbers PRJEB7866, PRJEB12000, and PRJEB19456 for Axial samples in the years 2013, 2014 and 2015 respectively (Supplementary Table 1).

Identification of CRISPR loci, viral clusters, and spacer clusters

Crass [48] v1.0.1 was used to identify CRISPR loci (i.e., unique direct repeat types) and CRISPR spacers in the metagenomic reads. Viral contigs in the metagenomic assemblies were identified with VirFinder [48, 49] v1.1 using a p-value threshold of 0.05. To cluster the CRISPR

spacers, we performed an all-v-all blast using BLASTn v2.5.0 (E-value threshold of 10^{-08}) and then clustered using Markov Cluster Algorithm (MCL) [50] v14-137 (inflation 1.2 and scheme 7) based on bitscore. We used ClusterGenomes (https://bitbucket.org/MAVERICLab/stampede-clustergenomes) v1.1.3 (95% identity, 80% coverage) to cluster viral contigs.

Recovery of MAGs

Metagenome assembled genomes (MAGs) were recovered from Mid-Cayman Rise assemblies using anvi'o [51]. Supervised clustering was used to recover bins from Mid-Cayman Rise contigs using anvi'o v2.1.0. MAGs from Axial Seamount assemblies were recovered using unsupervised binning with CONCOCT [50, 52, 53] within the anvi'o v4.0 pipeline, followed by manual refinement within anvi'o. Only bins with >=70% completeness and <=10% redundancy were retained as MAGs for this analysis. The MAGs were assigned taxonomies using PhyloSift [50, 52] v1.0.1 as described in Anderson et al. [33]. Only MAGs for which taxonomies could be identified were used in analyses. MAG coverage was normalized by the number of merged reads in the metagenome. All of the Mid-Cayman Rise MAGs were previously described in Anderson et al. [33].

Identification of reads matching 16S rRNA genes

The reads from the Mid-Cayman Rise and Axial Seamount metagenomes were mapped to the Silva SSU and LSU Parc databases [54, 55] (release 132) with bowtie2 [56] v2.2.9 using default settings and local alignment. Matching reads were then mapped to the Greengenes 13_5 16S rRNA database [57] and reads that mapped were retained for downstream analysis. Mapped reads were assigned taxonomies using the classify.seqs function in mothur [57, 58] v1.38.1 with the Silva 16S rRNA database (release 119) and a cutoff of 50. Reads mapping to 16S rRNA gene sequences that were classified as eukaryotes were excluded from analyses.

Viral, spacer, and host diversity

Rarefaction curves for viral clusters, CRISPR spacer clusters, and reads matching 16S rRNA genes categorized at the class level were created using the Vegan R package [59] v2.4-5. The number of viral clusters per contig and spacer clusters per read (paired reads) were used as proxies for viral diversity, and the number of different taxa at the class level matching 16S rRNA genes reflected microbial diversity, calculated on a per read basis (merged reads).

Viral, spacer, and CRISPR relative abundance

To calculate relative viral abundance, the reads in each of the Mid-Cayman Rise and Axial Seamount metagenomes were mapped against all of the viral contigs from the corresponding geographic region using bowtie2 [56] v2.2.9. The reads from each sample were mapped against all of the viral contigs from the corresponding geographic region rather than solely the viral contigs in the sample because there were viral reads in samples that did not assemble into viral contigs. This method therefore allowed the identification of more viral sequences. The number of reads that mapped to viral contigs was normalized by the number of merged reads as a measure of relative viral abundance. This measure of relative viral abundance reflects only the proportion of viruses that were retained on the filter as viral capsids or prophages. We used the number of spacers per read and CRISPR direct repeat types per read as measures of spacer and CRISPR relative abundance, respectively. Paired rather than merged

reads were used for these analyses. Relative abundance and diversity of viruses, microbes, CRISPR spacers and CRISPR loci were visualized using the Seaborn library within Python [60].

Taxonomic distribution of viral clusters, spacer clusters, and hosts

The relative compositions of the microbial community, the viral assemblage, and CRISPR spacers were compared between vent sites within the Mid-Cayman Rise and Axial Seamount. To calculate the relative abundance of each viral cluster in each sample, the number of reads in the sample that mapped to the viral contigs in the viral cluster was determined using bowtie2 [56] v2.2.9. The number of reads in each metagenome that mapped to the viral cluster was normalized by the total length of the viral contigs in the viral cluster and the number of merged reads in the metagenome. We defined the most common viral clusters as the six clusters with the highest relative abundance in each sample. The relative abundance of each CRISPR spacer cluster in each sample was calculated as the percent of spacers in the sample that were part of the spacer cluster. For spacer clusters, we defined the most common as the three) clusters with the highest relative abundance in each sample. To compare the compositions of viral contigs and CRISPR spacers, all of the spacers were blasted against all of the viral contigs using blastn (E-value $<=10^{-05}$, <=1 mismatch, as per Emerson et al. [61]. The relative abundance of each microbial host was measured as the number of reads that mapped to 16S rRNA gene sequences of the given taxon, normalized by the number of 16S rRNA gene reads in the sample. Analyses of microbial taxa were done at either the class level or the lowest taxonomic level available, depending on the analysis.

Microbiome datasets are compositional in nature because sequencing instruments impose an arbitrary total [62]. Therefore, to conduct hierarchical clustering of samples based on viral, spacer, and host composition, we used the protocol outlined by Gloor et al. [62] and Gloor and Reid [63] for computing distances between samples containing compositional data. For hierarchical clustering, we did not normalize; we performed analyses on the number of reads that mapped to each viral cluster, the number of CRISPR spacers in each spacer cluster, and the number of reads that mapped to 16S rRNA gene sequences of each microbial host in each sample; for microbial hosts, we did not include reads that mapped to unclassified sequences or sequences classified as eukaryotes. We replaced zero counts with estimates using the count zero multiplicative method for viral clusters and hosts and the Bayes-Laplace Bayesian multiplicative method for spacer clusters via the zCompositions R package [64] v1.2.0. Using the CoDaSeq R package (https://github.com/ggloor/CoDaSeq) [62] v0.99.3, we applied a centered log-ratio (clr) transformation to the count data (that lacked zero counts), thereby capturing the ratios between parts. To calculate distances between samples for hierarchical clustering, we used the ward.D2 method on the transformed counts.

Networks of viral infection

Infection networks of viral-host interactions were created using CRISPR sequences to connect microbial hosts with clusters of viral contigs, adapted from the methods used by Daly et al. [65] and Emerson et al. [61]. First, MAGs were connected to CRISPR direct repeat types within each sample using BLASTn v2.5.0 (E-value $<=10^{-10}$, 100% nucleotide identity, as per Emerson et al. [61]. Then, each CRISPR direct repeat type was matched to a set of CRISPR spacers as identified by Crass (Skennerton et al. 2013) v1.0.1. Finally, the CRISPR spacers in each sample were matched to viral contigs in the corresponding region using BLASTn v2.5.0

with an E-value cutoff of 10^{-05} and a maximum of one mismatch, as per Emerson et al. [61]. Only one mismatch was allowed because resistance has been found to be lost by single nucleotide differences between bacterial spacers and target phage sequences [66]. In contrast, resistance by archaeal CRISPR systems can still be provided when there are up to three mismatches between spacers and target phage sequences [67, 68]. We did not allow for more mismatches between archaeal spacers and phage sequences because resistance is weakened by more mismatches [68]. Each direct repeat type provided by Crass v1.0.1 does not necessarily represent an individual CRISPR locus [48].

Viruses and prophages in MAGS and viral and prophage genes

VirSorter [69] v1.0.3 was used to extract putative viral and prophage sequences from metagenomic assemblies and from individual MAGs. Category one, category two, category four (prophage), and category five (prophage) VirSorter sequences were used in analyses. Putative viral and prophage sequences from all assembled contigs were annotated via Prokka [69, 70] v1.14 with the taxonomic identifier set to "bacteria" as well as to "viruses." We assigned categories to the ORFs that had been identified with the "bacteria" taxonomic identifier using the Clusters of Orthologous Groups of proteins (COG) database [71, 72]. We conducted a second round of annotation for ORFs identified by Prokka using the "viruses" taxonomic identifier using VirSorter; we classified these ORFs as having viral function if they were annotated with terms that included "capsid," "tail," "spike," "terminase large subunit," "portal," and "coat."

Data analysis software

The majority of analyses were conducted in RStudio [73] v1.0.136. R packages used were readr [74] v1.1.1, readxl [75] v1.0.0, tidyr [76] v0.7.2, stringr [77] v1.3.0, dplyr [78] v0.7.4, ggnetwork [79] v0.5.1, statnet [80] v2016.9, ggpubr [81] v0.1.7, ggplot2 [82] v2.2.1, and svglite [83] v1.2.1.

Results

Relative abundance of viral and microbial reads

In order to examine the distribution and abundance of viruses across individual vent sites, vent fields, and regions, we quantified the relative abundance of viral sequences, CRISPRs, and spacers in each of the metagenomes (Figure 1, Supplementary Table 2). It is possible to examine viruses in 0.22 µm-filtered fluids because this fraction includes integrated prophage, lytic infections in progress, and free viral particles captured on filters. However, our analysis misses free viral particles that were not retained on the filters, and we are only capturing the viral diversity and variation across sites based on the viral sequences identified in the metagenomes. It is also important to note that CRISPR loci vary in number across microbial genomes, and we did not distinguish between CRISPR loci with the same direct repeat type. Therefore, our measure of CRISPR relative abundance is not a direct proxy for the abundance of viruses but instead gives a sense of how microbes respond to viral infection across samples.

Our results indicate that the relative abundance of viral sequences was similar both across and within the Axial and Mid-Cayman Rise hydrothermal vent regions. Excluding the background seawater samples, there was a higher relative abundance of CRISPR loci in Axial

Seamount than the Mid-Cayman Rise (Figure 1A; p = 0.000389, t-test). However, there was no difference in the relative abundance of viral sequences between the Mid-Cayman Rise and Axial metagenomes (Figure 1B; p = 0.2392, t-test), nor in the number of spacers within CRISPR loci per read (Figure 1C; p = 0.071, t-test). Within the Mid-Cayman Rise, we compared samples from mafic-hosted (Piccard) versus ultramafic-hosted (Von Damm) hydrothermal systems. We did not observe significant differences in the relative abundance of viral sequences, CRISPR loci, or CRISPR spacers between Piccard and Von Damm (viral sequences: p = 0.66, t-test, CRISPR loci: 0.40, t-test, spacers: 0.37, t-test). Similar results emerged from comparisons among vent fields within Axial Seamount: we did not observe significant differences in the relative abundance of viral sequences, CRISPR loci, or CRISPR spacers between vent fields at Axial Seamount (viral sequences: p = 0.17, t-test, CRISPR loci: 0.41, t-test, spacers: 0.11, t-test). Finally, within Axial Seamount, we also compared the relative abundance of viruses in samples taken from plume and diffuse flow hydrothermal fluid at Anemone vent. The Anemone diffuse flow samples had a higher relative abundance of CRISPR spacers and CRISPR loci compared to the Anemone plume sample, but the relative abundance of viral sequences did not differ between the Anemone plume and diffuse flow samples.

Diversity of viral assemblage and microbial community

Overall, viral diversity analyses revealed that the viral assemblages within these vents had high richness and were not dominated by specific viral strains. With the exception of a few samples, the rarefaction curves for the viruses and microbes did not reach saturation (Supplementary Fig. 1), indicating relatively high viral and microbial diversity at this resolution across vent sites at both the Mid-Cayman Rise and Axial Seamount. In each sample, there were no dominant viral or spacer clusters, and each of the viral and spacer clusters present was relatively rare (Supplementary Figs. 2-3). Moreover, the viral and spacer clusters with the highest coverage did not correspond to each other across samples: in both the Mid-Cayman Rise and Axial Seamount, only one of the most common spacer clusters matched with one of the most common viral clusters by BLAST. Thus, there were no dominant viral sequences that were consistently found across all samples.

We observed a more diverse viral assemblage at Axial Seamount compared to the Mid-Cayman Rise. We observed higher richness of both viruses (Figure 1D; p = 0.002086, t-test) and their microbial hosts (Figure 1E; p = 0.00392, t-test) in samples from vents at Axial Seamount compared to the Mid-Cayman Rise. However, we did not observe significant differences in either viral or host diversity between the Piccard and Von Damm vent fields at the Mid-Cayman Rise (viral diversity: p = 0.05306071, t-test, host diversity: p = 0.6648, t-test). Finally, we did not observe a significant difference in the diversity of CRISPR spacers between the Mid-Cayman Rise and Axial Seamount (Figure 1F, p = 0.06592, t-test).

We observed significant differences in viral diversity between vent sites within Axial Seamount (p = 0.02165, t-test). However, no significant differences emerged in terms of the diversity of microbial hosts and CRISPR spacers (host diversity: p = 0.7444, t-test, spacer diversity: p = 0.1081, t-test). At Anemone vent, the diffuse flow samples had higher CRISPR spacer and microbial diversity than the plume sample, but viral diversity did not differ between these samples.

Biogeography and relative abundance of viral clusters, spacer clusters, and hosts

In order to characterize viral biogeography and community similarity across hydrothermal vent fluids, we evaluated the extent to which viral sequences and CRISPR spacers were distributed across samples, then compared these results to the host microbial community. As before, we clustered sequences based on similarity to compare across samples. On the whole, viral sequences and CRISPR spacers had fairly limited distributions. Only a few viral clusters (0.17%) appeared at both the Mid-Cayman Rise and Axial Seamount, while almost no CRISPR spacer clusters were identified in both regions (Table 1). The most cosmopolitan viruses and CRISPR spacers did not match each other: at the Mid-Cayman Rise, three of the 100 most widely distributed viral clusters matched with only four of the 100 most widely distributed CRISPR spacer clusters according to BLAST; and at Axial Seamount, one of the 100 most widely distributed viral clusters matched with two of the 100 most widely distributed CRISPR spacer clusters. In contrast, microbial taxa were much more cosmopolitan, with ~41% of taxa shared between the Mid-Cayman Rise and Axial Seamount according to 16S rRNA gene classification at the lowest taxonomic level available (Table 1). Viral and CRISPR spacer clusters were shared more widely among vent sites within the Mid-Cayman Rise compared to Axial Seamount, but microbial lineages were shared more widely among vent sites at Axial Seamount than the Mid-Cayman Rise (Table 1).

We created hierarchical dendrograms to assess the similarity of samples based on their viral content. Viral assemblages in samples from the Mid-Cayman Rise and Axial Seamount grouped separately (Figure 2a). Within Axial Seamount, samples taken in successive years from the same vent tended to have similar viral assemblage compositions (Figure 2a). In contrast, at the Mid-Cayman Rise, samples taken from the same site in two different years did not cluster together, and we observed weak clustering of samples by location (Figure 2a). At the Mid-Cayman Rise and Axial Seamount, grouping of microbial communities based on classification of 16S rRNA reads in the metagenomes showed similar patterns to the viral assemblages, with Axial and Mid-Cayman Rise communities grouping separately from each other, and with microbial communities at specific sites within Axial showing more similarity across years than at the Mid-Cayman Rise (Figure 2b).

In contrast to the viral and microbial assemblages, hierarchical dendrograms based on CRISPR spacer compositions showed little clustering by location (Figure 3). Based on spacer assemblages, samples from either Axial Seamount or the Mid-Cayman Rise did not cluster together, and samples taken from the same vent sites in different years did not cluster together. Very few CRISPR spacer clusters were found at multiple vent sites (Table 1).

Networks of viral infection

Networks of viral infection generated from viral sequences, CRISPRs, and MAGs were used to examine the distribution and host specificity of vent viruses. Networks show connections between CRISPR spacers, which represent a record of viral infection in the host, and viral sequences, which represent viral sequences present in the community at the time of sampling. Virus-host networks made for Axial Seamount (Figure 4) and Mid-Cayman Rise (Figure 5) indicate that vent viruses are restricted in both their host range and their biogeographic range. We did not observe any virus-host connections shared between the Mid-Cayman Rise and Axial Seamount. Within Axial Seamount, there were 89 viral clusters linked to 18 MAGs, and 105

connections between distinct pairs of viral clusters and MAGs (Figure 4). We did not observe any connections between MAGs and viral clusters in any of the non-diffuse flow samples (Background, CTD1200 and the Anemone plume sample). The number of viral connections appeared to be related to, but was not significantly correlated with, the microbial hosts' relative abundance (Supplementary Table 3): for example, a Clostridia MAG in the 2015 North Rift Zone sample (Clostridia_35) had high normalized coverage and was linked to nine viral clusters. We observed a high number of viruses shared among *Aquificae* MAGs sampled from the Anemone vent in 2013 and 2014 (Figure 4). Finally, we observed only a single viral cluster that was linked to MAGs of different taxonomic classes from different vent sites, with a viral cluster connected to an *Ignavibacteria* MAG (Ignavibacteria_15) in the 2014 Marker 33 sample and an *Aquificae* MAG (Aquificae_43) in the 2013 Anemone sample (Figure 4).

The viral infection network from the Mid-Cayman Rise was larger than the Axial Seamount network, with 110 viral clusters linked to 16 MAGs, and 130 connections between distinct pairs of viral clusters and MAGs (Figure 5). While most viral clusters in the network represented relatively rare viruses, one viral cluster was among the top six most abundant in both X-19 and Shrimp Gully #2. This cluster linked to a Campylobacterales and a Desulfobacterales MAG, both from Shrimp Gully #2. This was the only viral cluster with high relative abundance in at least one sample that was present in either the Axial Seamount or the Mid-Cayman Rise viral infection network (Figures 4-5). Some MAGs had more viral connections than others: for example, a *Methanomicrobia* MAG (Methanomicrobia_41) from the 2012 Shrimp Hole sample was linked to 35 viral clusters (Figure 5). However, this is not a direct indication of the number of different viruses infecting a specific strain because some MAGs had more CRISPR spacers than others, increasing the possibility of finding a viral connection. As with Axial Seamount, there were instances of highly abundant MAGs with many viral connections (Supplementary Table 3): a *Sulfurovum* MAG in the 2012 X-19 sample (Sulfurovum_99) had the highest normalized coverage at the Mid-Cayman Rise and was linked to 18 viral clusters, making it the MAG with the third highest number of viral connections within the network. Another *Sulfurovum* MAG (Sulfurovum_37), had a high relative abundance in the 2012 Shrimp Gulley #2 sample, and had two viral connections. At the Mid-Cayman Rise, we observed several cases in which viruses were connected to multiple microbial hosts within the network (Figure 5). However, the MAGs were linked to these shared viral clusters through the same CRISPR direct repeat type and were in the same sample, suggesting that these connections may have been due to matching CRISPR direct repeat types rather than true cross-infection. For example, at the Mid-Cayman Rise, 9.62% of the CRISPR direct repeat types that were found in a MAG were found in multiple MAGs in the same sample. When direct repeat types are found in multiple MAGs in the same sample, it cannot be conclusively determined which MAG the CRISPR spacers associated with the direct repeat type came from. This is because Crass, the software used for CRISPR identification, creates networks of direct repeats and spacers based on reads rather than contigs. Therefore, all viral clusters connected to multiple MAGs in the Mid-Cayman Rise may either be true examples of viral-host cross-infection or may instead result from shared direct repeats.

Viruses and prophages in MAGs

To determine whether there were prophages integrated in the MAGs we recovered, we used VirSorter to identify putative viral contigs within each MAG. Of the 74 Mid-Cayman Rise

MAGs, 28 (38%) had at least one putative prophage sequence (VirSorter categories one, two, four, or five) (Supplementary Table 4). Of the 98 Axial Seamount MAGs, 41 (42%) had at least one putative viral or prophage sequence (VirSorter categories one, two, four, or five) (Supplementary Table 4).

Gene content of viral and prophage sequences

To determine whether viruses and prophage in these samples carried auxiliary metabolic genes (AMGs), we annotated and characterized viral and prophage gene content in the viral sequences recovered from the metagenomes. The vast majority of genes within the putative viral and prophage sequences were annotated as belonging to the COG categories replication and repair; nucleotide metabolism and transport; and post-translational modification, protein turnover, chaperone functions. Many genes were also annotated as belonging to the cell wall/membrane/envelope biogenesis category. Within the Mid-Cayman Rise, a total of 106 ORFs within putative viral or prophage sequences identified in 10 different metagenomes fell within the broad COG category of "metabolism" (Supplementary Fig. 4). This included genes categorized as energy production and conversion, amino acid transport and metabolism, nucleotide transport and metabolism, carbohydrate transport and metabolism, coenzyme transport and metabolism, lipid transport and metabolism, and inorganic ion transport and metabolism. Samples taken from X-19 and Shrimp Hole (2012) had a high number of metabolism genes (30 and 28, respectively) on putative viral or prophage sequences compared to the other Mid-Cayman Rise samples. A higher proportion of genes on putative viral or prophage sequences were related to metabolism at the Mid-Cayman Rise than Axial Seamount (Supplementary Fig. 4). At Axial Seamount, we identified a total of 185 ORFs on putative viral or prophage sequences from 14 different metagenomes that fell within the broad COG category of "metabolism." However, it is important to note that none of the observed metabolism genes were surrounded by confirmed virus genes (i.e., known viral ORFs both upstream and downstream of the metabolic ORF), and therefore we cannot rule out the possibility that some of these genes were microbial in origin.

Discussion

Viruses are important drivers of microbial mortality, ecology and evolution in the ocean, but studies of their distribution and impact in the deep sea are lacking. Our results indicate that viral diversity is high in venting fluids from hydrothermal systems, and the viruses we analyzed have restricted biogeographic distributions and host ranges. This implies that viruses do not spread widely between vent sites and that the viral role in mediating horizontal gene transfer across taxa and between vent sites is relatively restricted. However, the high abundance of integrated proviruses and the large number of virus-host CRISPR connections implies rapid viral mutation and ongoing viral infection, indicating active and ongoing interactions between viruses and their microbial hosts in venting fluids from both Axial and Mid-Cayman Rise hydrothermal vents.

Vents host diverse and active assemblages of viruses with restricted host range

Our viral infection networks show diverse microbial lineages across several samples infected by many different viruses at the time of sampling. Although individual microbial lineages could be targeted by multiple viral strains, the viruses we identified in these

hydrothermal habitats had fairly restricted host ranges, infecting specific individual microbial strains. The only clear example of viral infection across microbial taxa emerged at Axial Seamount, where viral sequences associated with Ignavibacteria were also linked to Aquificales (Figure 5). All other examples of viral infection across microbial taxa may either be true examples of viral-host cross-infection or may instead result from shared direct repeats. The limited host range of viruses from hydrothermal systems is consistent with previous work indicating that viruses generally tend to be host-specific [84], and we found very little evidence for generalist viruses, despite the fact that this has been reported previously [84, 85]. We identified many viral connections to Epsilonbacteraeota, which are abundant and active at both Axial Seamount and the Mid-Cayman Rise [4, 33, 39, 40, 86], showing that these highly successful microbial groups in hydrothermal vent fluids are also susceptible to viral infection.

Matches between CRISPR loci and viruses provide an indication of which viruses are being targeted by the CRISPR immune response. We found that while some CRISPR spacers target relatively abundant viruses, most CRISPRs target relatively rare viruses in these systems. At the Mid-Cayman Rise, we found only one viral cluster with high relative abundance to be targeted by a host CRISPR system (Figure 5), and we did not find any instances of viral clusters with high relative abundance targeted by CRISPRs at Axial Seamount (Figure 4). This is consistent with previous observations in an archaea-dominated hypersaline lake [87], where the vast majority of CRISPRs were found to target viruses with populations too small to allow for the assembly of contigs. The relative scarcity of viruses targeted by CRISPRs may result from an evolutionary arms race: CRISPRs limit the abundance of the viral populations they target, while concurrently, viruses undergo mutations, limiting the ability of CRISPR spacers to target them. Alternatively, these observations could result from an abundance of inactive spacers inherited over multiple generations. However, we do not believe this to be the case because CRISPR spacer clusters were infrequently present across multiple samples, suggesting that spacers were integrated on subgenerational timescales.

Viral assemblages are spatially restricted across vent sites

Previous studies of viruses in marine systems have observed viral populations to be commonly found across multiple samples [88], whereas others have found that most viruses are biogeographically restricted, with only a few cosmopolitan groups [84, 89]. In Mid-Cayman Rise and Axial Seamount hydrothermal vents, we found high viral diversity with a limited distribution, potentially indicating rapid diversification at vent sites. Most viral clusters we observed were found only at individual vent sites, with very few cosmopolitan viruses, and the clustering of viral assemblages by location further indicated restricted dispersal. Thus, viral biogeographic patterns correspond to those of their microbial hosts: previous work examining microbial distribution in hydrothermal systems through fine-scale 16S rRNA gene analyses indicates that while some microbial lineages are endemic to individual vent sites, others are widespread across vent fields [28, 40]. Often, sample sites in close proximity are often more similar to each one another in terms of microbial community structure than to geographically distant sites [4, 27, 40, 90]. This may result from subseafloor plumbing that restricts fluid flux between sites, creating “islands” of microbial diversity that are distinct from one vent site to the next [26]. Here, we show barriers to dispersal apply to viruses as well, and our results suggested that viruses were even more spatially restricted than their microbial hosts. Microbial lineages that spread between vent sites may face infection from novel viral strains not found in other vent

sites. These endemic viral populations thus further shape distinct microbial community structure at individual vent sites.

Although the viral assemblages at hydrothermal vents had high diversity and restricted dispersal between vent sites and vent fields, our results show that the viral assemblages in vents persist over time, particularly at Axial Seamount. Viral assemblages from samples from the same vent sites at Axial across three years clustered together in the hierarchical dendrograms (Figure 2a), and our viral infection networks revealed that a number of viruses at Axial Seamount were linked to specific microbial hosts at the same vent over multiple years (Figure 4). These patterns match those observed in the microbial communities (Figure 2b) and are consistent with previous work showing spatially restricted but temporally stable microbial communities over time at Axial Seamount [4, 27]. Our work extends this to the viral world at Axial Seamount, indicating that viral assemblages follow the same temporal patterns as their microbial hosts, and that virus-host relationships persist over time.

However, temporal stability in the viral and microbial community was generally not preserved at the Mid-Cayman Rise for sites sampled in both 2012 and 2013. Viral and CRISPR spacer clusters were also more widespread (i.e., more often found at multiple locations) in the Mid-Cayman Rise than Axial Seamount, potentially implying more rapid dispersal or slower diversification of both viruses and their microbial hosts at the Mid-Cayman Rise compared to Axial. It is unclear whether this is due to sampling bias (fewer time points), to abiotic factors, such as higher degrees of fluid flux facilitating exchange at the Mid-Cayman Rise, or biotic factors, such as a set of conditions that encourage slower viral replication and diversification at the Mid-Cayman Rise compared to Axial.

Patterns recorded in microbial CRISPR loci do not reflect the contemporary viral assemblage

In contrast to the viral clusters, the CRISPR spacers did not demonstrate any clear biogeographic patterns. We expected CRISPR spacers to be more widespread than viruses since viral composition represents the virus community at the time of sampling, while spacers represent a history of viral infection. In contrast to our predictions, while both had limited distributions, we found viruses to be more widespread than CRISPR spacers at all scales examined (Table 1). These results are in contrast to previous studies of CRISPR spacer biogeography in terrestrial hot springs, where both viral and CRISPR spacers showed clear biogeographic structure [91] This suggests that there is selective pressure for CRISPR spacer composition to evolve more rapidly than viral sequences via the loss or mutation of CRISPR spacers. However, it is also possible that we found viruses to be more widespread because of undersampling of CRISPR spacers, or our use of different clustering algorithms for viruses and spacers: spacers may have been clustered at a finer resolution, resulting in a more narrow distribution for each spacer.

Given that CRISPR spacers provide a history of infection, comparing the record of past viral infections via CRISPR arrays with viral sequences in the metagenomes can provide insights into whether CRISPR arrays provide an accurate representation of the viral assemblage at the time of sampling, as well as the rate at which CRISPR spacers are accumulated. We did not find that the most common or cosmopolitan viruses and CRISPR spacers matched each other. This may arise from a temporal mismatch: it takes time for spacers to be incorporated into and lost

from CRISPR loci as virus abundances shift and viruses evolve. Additionally, just one SNP (which we allowed for when aligning spacers to viruses) can prevent a CRISPR spacer from providing resistance against a virus [66]. This could explain the discrepancy between viruses and spacers: once resistance of a spacer to a particular virus is suppressed, the population of the virus is freed to shift independently from the spacer in the host population. The discrepancy we observed between viruses and spacers is important to note when attempting to use CRISPR spacers to study viral populations or vice versa.

Auxiliary metabolic genes and the role of virus-driven horizontal gene transfer in these systems

Previous work has suggested that viruses in hydrothermal systems have the genomic capacity to alter their hosts' microbial metabolism through harboring auxiliary metabolic genes (AMGs) [17–19]. We found a higher percentage of microbial genomes with putative viral sequences (40%) than has been previously reported for single-cell genomes (SAGs, 10%) in diffuse flow hydrothermal fluids [15]. The high abundance of prophage identified here is consistent with previous work documenting a high incidence of prophage in diffuse flow systems compared to deep seawater [14].

The ORFs on the viral contigs identified in these hydrothermal vent metagenomes encoded a wide range of functions, some of which may function as auxiliary metabolic genes. Many viral contigs contained genes related to cell wall or membrane proteins. The function of these genes is unknown; it is possible that these genes are involved in the synthesis of membranes surrounding viral capsids. Genes related to outer membrane proteins and protein glycosylation are commonly observed as variable genes within microbial pangenomes [92–98], possibly as a means to vary membrane proteins to evade viral infection. Viruses often act as a source of genes for horizontal gene transfer via transduction, and it is possible that the introduction of these novel genes may enable some microbial strains to avoid infection by other viruses. We also observed several genes with functions related to energy metabolism, inorganic ion transport, and signal transduction. Previous work has found that genes related to inorganic ion transport are differentially distributed in the *Sulfurovum* pangenome according to nutrient availability [39] and further work on viral sequences across vent sites is needed to determine whether these genes are carried by viruses to benefit their hosts. Moreover, work on viruses in hydrothermal plumes [19] and in diffuse flow vent fluid [18] has indicated that viruses in these systems encode energy-metabolizing AMGs, potentially to supplement the host's ability to generate sufficient energy for cellular processes during the course of infection. The abundance and diversity of metabolic genes observed on these contigs provide further evidence that viral-encoded AMGs are widespread and diverse within hydrothermal vent ecosystems. These genes also have the potential to be horizontally transferred via viral transduction. However, our analyses of viral biogeography in hydrothermal vents has revealed virus-host interactions to have limited distributions because viruses are spatially restricted and highly host-specific. Thus, the potential for viruses to act as mechanisms for horizontal gene transfer between spatially and phylogenetically distant hosts is likely limited.

Taken together, our results show that hydrothermal vent viruses are active, abundant, diverse, are commonly found as integrated prophage in microbial genomes, and carry auxiliary metabolic genes with the potential to alter the phenotype and evolutionary trajectory of their microbial hosts. These viruses are restricted in their host range and biogeographic extent, but

their interactions with hosts persist over time. Thus, while viruses in venting fluids from deep-sea hydrothermal systems have the capacity to play an important role in driving the evolution and ecology of microbial communities, their influence appears to be highly localized to specific regions and taxa. Future work examining viral diversity and distribution across higher resolution transects over space and time should reveal further insights into the extent of the viral influence in deep-sea hydrothermal vents.

Acknowledgements

We thank Julie Reveillaud, Caroline Fortunato, and Emily Reddington for support in the collection and generation of metagenomic data, and Chip Breier, David Butterfield, Bill Chadwick, Chris German, Jim Holden, Jill McDermott, and Jeff Seewald for sample collection support at sea. Funding for ET was provided by Carleton College. REA was supported by a NASA Postdoctoral Fellowship with the NASA Astrobiology Institute. This work was supported by a NASA Exobiology grant number 80NSSC18K1076 to REA and JAH; a NASA Astrobiology Science and Technology for Exploring Planets (ASTEP) grant NNX-327 09AB75G and a grant from Deep Carbon Observatory's Deep Life Initiative to JAH; the NSF Science and Technology Center for Dark Energy Biosphere Investigations (C-DEBI) to JAH; and the Gordon and Betty Moore Foundation Grant GBMF3297 to JAH. Samples were collected from the Mid-Cayman Rise with the assistance of the captains and crew of the R/V Atlantis and R/V Falkor as well as ROVs Jason and Nereus. For Mid-Cayman Rise, ship and vehicle time in 2012 were supported by the NSF-OCE great OCE-1061863 to Chris German and Jeff Seewald and in 2013 by the Schmidt Ocean Institute during cruise FX008-2013 aboard the R/V Falkor. Samples collected from Axial Seamount were collected with the assistance of the captains and crew of the R/V Falkor, R/V Thompson, and R/V Brown as well as the ROV ROPOS and Jason groups, and in 2013 the Schmidt Ocean Institute during cruise FK010-2013 aboard the R/V Falkor. This is C-DEBI Contribution number [number to be determined].

References

1. Jannasch HW, Mottl MJ. Geomicrobiology of Deep-Sea Hydrothermal Vents. *Science* . 1985. , **229**: 717–725
2. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, et al. Microbial population structures in the deep marine biosphere. *Science* 2007; **318**: 97–100.
3. Meier DV, Pjevac P, Bach W, Hourdez S, Girguis PR, Vidoudez C, et al. Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J* 2017; **11**: 1545–1558.
4. Fortunato CS, Larson B, Butterfield DA, Huber JA. Spatially distinct, temporally stable microbial populations mediate biogeochemical cycling at and below the seafloor in hydrothermal vent fluids. *Environ Microbiol* 2018; **20**: 769–784.
5. McNichol J, Stryhanyuk H, Sylva SP, Thomas F, Musat N, Seewald JS, et al. Primary productivity below the seafloor at deep-sea hot springs. *Proc Natl Acad Sci U S A* 2018; **115**: 6756–6761.
6. Trembath-Reichert E, Butterfield DA, Huber JA. Active subseafloor microbial communities from Mariana back-arc venting fluids share metabolic strategies across different thermal niches and taxa. *The ISME Journal* . 2019. , **13**: 2264–2279
7. Galambos D, Anderson RE, Reveillaud J, Huber JA. Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents. *Environ Microbiol* 2019; **21**: 4395–4410.
8. Suttle CA. Viruses in the sea. *Nature* 2005; **437**: 356–361.
9. Suttle CA. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* 2007; **5**: 801–812.
10. Chow C-ET, Suttle CA. Biogeography of Viruses in the Sea. *Annu Rev Virol* 2015; **2**: 41–66.
11. Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? *Environ Microbiol* 2004; **6**: 1–11.
12. Ortmann AC, Suttle CA. High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. *Deep Sea Research Part I: Oceanographic Research Papers* . 2005. , **52**: 1515–1527
13. Anderson RE, Brazelton WJ, Baross JA. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol Ecol* 2011; **77**: 120–133.
14. Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, Winget D, et al. Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *ISME J* 2008; **2**: 1112–1121.
15. Labonté JM, Pachiadaki M, Fergusson E, McNichol J, Grosche A, Gulmann LK, et al. Single Cell Genomics-Based Analysis of Gene Content and Expression of Prophages in a Diffuse-Flow Deep-Sea Hydrothermal System. *Frontiers in Microbiology* . 2019. , **10**
16. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. *Nat Microbiol* 2018; **3**: 754–766.
17. He T, Li H, Zhang X. Deep-Sea Hydrothermal Vent Viruses Compensate for Microbial Metabolism in Virus-Host Interactions. *MBio* 2017; **8**.
18. Anderson RE, Sogin ML, Baross JA. Evolutionary strategies of viruses, bacteria and archaea in

hydrothermal vent ecosystems revealed through metagenomics. *PLoS One* 2014; **9**: e109696.
19. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. Sulfur oxidation genes in diverse deep-sea viruses. *Science* 2014; **344**: 757–760.
20. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 2001; **55**: 709–742.
21. Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A* 2005; **102**: 14332–14337.
22. Fuchsman CA, Collins RE, Rocap G, Brazelton WJ. Effect of the environment on horizontal gene transfer between bacteria and archaea. *PeerJ* 2017; **5**: e3865.
23. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000; **405**: 299–304.
24. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 2005; **3**: 722–732.
25. Brazelton WJ, Baross JA. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *ISME J* 2009; **3**: 1420–1424.
26. Stewart LC, Algar CK, Fortunato CS, Larson BI, Vallino JJ, Huber JA, et al. Fluid geochemistry, local hydrology, and metabolic activity define methanogen community size and composition in deep-sea hydrothermal vents. *ISME J* 2019; **13**: 1711–1721.
27. Opatkiewicz AD, Butterfield DA, Baross JA. Individual hydrothermal vents at Axial Seamount harbor distinct subseafloor microbial communities. *FEMS Microbiol Ecol* 2009; **70**: 413–424.
28. Anderson RE, Sogin ML, Baross JA. Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents. *FEMS Microbiol Ecol* 2015; **91**: 1–11.
29. Han Y, Perner M. The globally widespread genus Sulfurimonas: versatile energy metabolisms and adaptations to redox clines. *Front Microbiol* 2015; **6**: 989.
30. Dick GJ. The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally. *Nat Rev Microbiol* 2019; **17**: 271–283.
31. Damm KV, Von Damm K. Seafloor Hydrothermal Activity: Black Smoker Chemistry And Chimneys. *Annual Review of Earth and Planetary Sciences* . 1990. , **18**: 173–204
32. Kelley DS. A Serpentinite-Hosted Ecosystem: The Lost City Hydrothermal Field. *Science* . 2005. , **307**: 1428–1434
33. Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, et al. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents. *Nat Commun* 2017; **8**: 1114.
34. Butterfield DA, Roe KK, Lilley MD, Huber JA, Baross JA, Embley RW, et al. Mixing, reaction and microbial activity in the sub-seafloor revealed by temporal and spatial variation in diffuse flow vents at Axial volcano. *The Subseafloor Biosphere at Mid-Ocean Ridges* . 2004. , 269–289
35. McDermott JM, Seewald JS, German CR, Sylva SP. Pathways for abiotic organic synthesis at submarine hydrothermal fields. *Proc Natl Acad Sci U S A* 2015; **112**: 7668–7672.
36. German CR, Bowen A, Coleman ML, Honig DL, Huber JA, Jakuba MV, et al. Diverse styles of submarine venting on the ultraslow spreading Mid-Cayman Rise. *Proc Natl Acad Sci U S A* 2010; **107**: 14020–14025.
37. McDermott JM. Geochemistry of deep-sea hydrothermal vent fluids from the Mid-Cayman Rise, Caribbean Sea. 2015.
38. McDermott JM, Sylva SP, Ono S, German CR, Seewald JS. Geochemistry of fluids from Earth's

deepest ridge-crest hot-springs: Piccard hydrothermal field, Mid-Cayman Rise. *Geochimica et Cosmochimica Acta* . 2018. , **228**: 95–118
39. Moulana A, Anderson RE, Fortunato CS, Huber JA. Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus in Geographically Distinct Deep-Sea Hydrothermal Vents. *mSystems* 2020; **5**.
40. Reveillaud J, Reddington E, McDermott J, Algar C, Meyer JL, Sylva S, et al. Subseafloor microbial communities in hydrogen-rich vent fluids from hydrothermal systems along the Mid-Cayman Rise. *Environ Microbiol* 2016; **18**: 1970–1987.
41. Breier JA, Gomez-Ibanez D, Reddington E, Huber JA, Emerson D. A precision multi-sampler for deep-sea hydrothermal microbial mat studies. *Deep Sea Research Part I: Oceanographic Research Papers* . 2012. , **70**: 83–90
42. Breier JA, Rauch CG, McCartney K, Toner BM, Fakra SC, White SN, et al. A suspended-particle rosette multi-sampler for discrete biogeochemical sampling in low-particle-density waters. *Deep Sea Research Part I: Oceanographic Research Papers* . 2009. , **56**: 1579–1589
43. Crump BC, Hopkinson CS, Sogin ML, Hobbie JE. Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. *Appl Environ Microbiol* 2004; **70**: 1494–1505.
44. Zhou J, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. *Appl Environ Microbiol* 1996; **62**: 316–322.
45. Fortunato CS, Huber JA. Coupled RNA-SIP and metatranscriptomics of active chemolithoautotrophic communities at a deep-sea hydrothermal vent. *ISME J* 2016; **10**: 1925–1938.
46. Eren AM, Vineis JH, Morrison HG, Sogin ML. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One* 2013; **8**: e66643.
47. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012; **28**: 1420–1428.
48. Skennerton CT, Imelfort M, Tyson GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res* 2013; **41**: e105.
49. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017; **5**: 69.
50. van Dongen SM. Graph Clustering by Flow Simulation. 2000.
51. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015; **3**: e1319.
52. Darling AE, Jospin G, Lowe E, Matsen FA 4th, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2014; **2**: e243.
53. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; **11**: 1144–1146.
54. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007; **35**: 7188–7196.
55. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013; **41**: D590–6.
56. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–359.

57. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006; **72**: 5069–5072.
58. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537–7541.
59. Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* . 2003. , **14**: 927
60. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. mwaskom/seaborn: v0.8.1. *Zenodo* 2017.
61. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* 2018; **3**: 870–880.
62. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* 2017; **8**: 2224.
63. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can J Microbiol* 2016; **62**: 692–703.
64. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* . 2015. , **143**: 85–96
65. Daly RA, Roux S, Borton MA, Morgan DM, Johnston MD, Booker AE, et al. Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat Microbiol* 2019; **4**: 352–361.
66. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007; **315**: 1709–1712.
67. Garrett RA, Vestergaard G, Shah SA. Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol* 2011; **19**: 549–556.
68. Manica A, Zebec Z, Teichmann D, Schleper C. In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. *Mol Microbiol* 2011; **80**: 481–491.
69. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015; **3**: e985.
70. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30**: 2068–2069.
71. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 2015; **43**: D261–9.
72. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003; **4**: 41.
73. Team R, Others. RStudio: integrated development for R. *RStudio, Inc , Boston, MA URL http://www rstudio com* 2015; **42**: 14.
74. Wickham H, Hester J, Francois R. Readr: Read rectangular text data. *R package version* 2017; **1**.
75. Wickham H, Bryan J. readxl: Read Excel files. *R package version* 2017; **1**.
76. Wickham H, Henry L, Others. tidyr: Easily Tidy Data with 'spread ()'and 'gather ()'Functions. *R package version 0 8* 2018; **2**.
77. Wickham H. stringr: Simple, consistent wrappers for common string operations. *R package version* 2017; **1**.
78. Wickham H, Francois R, Henry L, Müller K, Others. dplyr: A grammar of data manipulation. *R*

package version 0 4 2015; **3**.
79. Briatte F. ggnetwork: Geometries to Plot Networks with 'ggplot2'. *R package version 0 5* 2016; **1**.
80. Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *J Stat Softw* 2008; **24**: 1548.
81. Kassambara A. ggpubr:'ggplot2' based publication ready plots. *R package version 0 1* 2017; **6**.
82. Wickham H. ggplot2. *WIREs Comp Stat* 2011; **3**: 180–185.
83. Luciani TJ, Decorde M, Lise V. RSvgDevice: An R SVG graphics device. *R package version 0 6* 2009; **4**.
84. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. *Nature* 2016; **536**: 425–430.
85. Peters DL, Lynch KH, Stothard P, Dennis JJ. The isolation and characterization of two Stenotrophomonas maltophilia bacteriophages capable of cross-taxonomic order infectivity. *BMC Genomics* 2015; **16**: 664.
86. Yoshida-Takashima Y, Takaki Y, Shimamura S, Nunoura T, Takai K. Genome sequence of a novel deep-sea vent epsilonproteobacterial phage provides new insight into the co-evolution of Epsilonproteobacteria and their phages. *Extremophiles* 2013; **17**: 405–419.
87. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, et al. Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* 2013; **2013**: 370871.
88. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* 2015; **348**: 1261498.
89. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 2016; **537**: 689–693.
90. Flores GE, Campbell JH, Kirshtein JD, Meneghin J, Podar M, Steinberg JI, et al. Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ Microbiol* 2011; **13**: 2158–2171.
91. Held NL, Whitaker RJ. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. *Environ Microbiol* 2009; **11**: 457–466.
92. Anderson RE, Kouris A, Seward CH, Campbell KM, Whitaker RJ. Structured Populations of Sulfolobus acidocaldarius with Susceptibility to Mobile Genetic Elements. *Genome Biol Evol* 2017; **9**: 1699–1710.
93. Meyer JL, Huber JA. Strain-level genomic variation in natural populations of Lebetimonas from an erupting deep-sea volcano. *ISME J* 2014; **8**: 867–880.
94. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, et al. Genomic islands and the ecology and evolution of Prochlorococcus. *Science* 2006; **311**: 1768–1770.
95. Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D'Auria G, Zhaxybayeva O, Papke RT, et al. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J* 2007; **1**: 235–245.
96. Peña A, Teeling H, Huerta-Cepas J, Santos F, Yarza P, Brito-Echeverría J, et al. Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains. *ISME J* 2010; **4**: 882–895.
97. Grote J, Cameron Thrash J, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, et al. Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio* . 2012. , **3**
98. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell

genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 2014; **344**: 416–420.

Tables

Table 1. Distribution of viral clusters, CRISPR spacer clusters, and microbial hosts (classified at the lowest taxonomic level available based on 16S rRNA genes) across vent sites and vent fields in the Mid-Cayman Rise and Axial Seamount. The Background and CTD1200 samples were excluded. Reads that mapped to 16S rRNA sequences that were classified as eukaryotes were excluded.

Region	Type	Percent in more than one vent site	Percent in all vent sites in region	Percent in more than one vent field	Percent in all vent fields in region
Mid-Cayman Rise	Viral clusters	14.19	0.00	3.28	3.28
Mid-Cayman Rise	Spacer clusters	6.34	0.00	0.05	0.05
Mid-Cayman Rise	16S rRNA host	49.41	1.19	32.81	32.81
Axial Seamount	Viral clusters	8.88	0.05	7.06	0.22
Axial Seamount	Spacer clusters	1.01	0.00	0.34	0.00
Axial Seamount	16S rRNA host	54.55	2.36	49.83	9.43

Type	Percent in both the Mid-Cayman Rise and Axial Seamount
Viral clusters	0.17
Spacer clusters	1.71E-03
16S rRNA host	40.66

Figure captions

Figure 1. Abundance and diversity of viral sequences, CRISPR spacers, CRISPR loci, and microbes in diffuse flow samples from the Mid-Cayman Rise (blue) and Axial Seamount (orange). Values for individual samples are indicated with black dots. Violins represent the kernel density estimation of the underlying data distribution. Variables with significant differences between the Mid-Cayman Rise and Axial Seamount are indicated with asterisks. For these comparisons, samples from background seawater and plumes were excluded.

Figure 2. Hierarchical clustering of the Mid-Cayman Rise (black; Von Damm is bold; Piccard is non-bold) and Axial Seamount (red) diffuse flow samples based on A) viral assemblage and B) microbial host composition. For viral composition, the number of reads in each sample that mapped to each viral cluster was calculated; the zero counts were replaced with estimates and a centered log-ratio (clr) transformation was applied. For the host composition, the number of reads that mapped to 16S rRNA sequences of each host (classified at the lowest taxonomic level available) in each sample was calculated; the zero counts were replaced with estimates and a centered log-ratio (clr) transformation was applied. Reads that mapped to 16S rRNA sequences that were either unclassified or classified as eukaryotes were excluded. The y-axes indicate distance between samples as calculated by the ward.D2 method based on the transformed counts.

Figure 3. Hierarchical clustering of CRISPR spacer composition of the Mid-Cayman Rise (black; Von Damm is bold; Piccard is non-bold) and Axial Seamount (red) diffuse flow samples. For all samples, the number of CRISPR spacers in each sample in each spacer cluster was calculated; the zero counts were replaced with estimates and a centered log-ratio (clr) transformation was applied. The y-axis indicates distance between samples as calculated by the ward.D2 method based on the transformed counts.

Figure 4. Infection network showing the links between viral clusters and MAGs at Axial Seamount. Viral clusters were linked to MAGs via the spacers and direct repeats in CRISPR loci. There are no red edges because there are no viral clusters linked to multiple MAGs in the same sample through the same direct repeat type. No viral clusters are outlined in black because none of the viral clusters with high relative abundance (top six most abundant) in at least one sample are present in the network.

Figure 5. Infection network showing the links between viral clusters and MAGs at the Mid-Cayman Rise. Viral clusters were linked to MAGs via the spacers and direct repeats in CRISPR loci. Edges are colored red when a viral cluster is linked to multiple MAGs in the same sample through the same direct repeat type. Due to the Crass algorithm, when the same direct repeat type is found in multiple MAGs in the same sample, it cannot be determined which MAG the spacers associated with direct repeat type came from. Therefore, the red edges are links between viral clusters and MAGs that could not be definitely proven. Viral clusters with high relative abundance (top six most abundant) in at least one sample are outlined in black.

Supplementary Figure captions

Supplementary Figure 1. Rarefaction curves for viral clusters, spacer clusters, and 16S rRNA host classes from samples collected at A) the Mid-Cayman Rise and B) Axial Seamount. There is no viral rarefaction curve for Old Man Tree, 2012 because there were no viral clusters in this sample.

Supplementary Figure 2. Rank abundance curves of viral clusters at A) the Mid-Cayman Rise and B) Axial Seamount. The y-axis represents the number of reads that mapped to each viral cluster divided by the total length of the viral contigs in the cluster and the number of reads in the sample. The x-axis represents the viral clusters ordered according to relative abundance, with the most abundant viral clusters at the left. The Axial viral clusters in all cases were too rare to be seen on the graphs.

Supplementary Figure 3. Rank abundance curves of CRISPR spacer clusters in A) the Mid-Cayman Rise and B) Axial Seamount. The y-axis represents the percent of all CRISPR spacers in a sample falling into that cluster. The x-axis represents the CRISPR spacer clusters ordered according to relative abundance, with the most abundant clusters at the left.

Supplementary Figure 4. COG categories assigned to ORFs on contigs identified as being derived from viruses or prophage by VirSorter. A) Samples from the Mid-Cayman Rise; B) Samples from Axial Seamount.

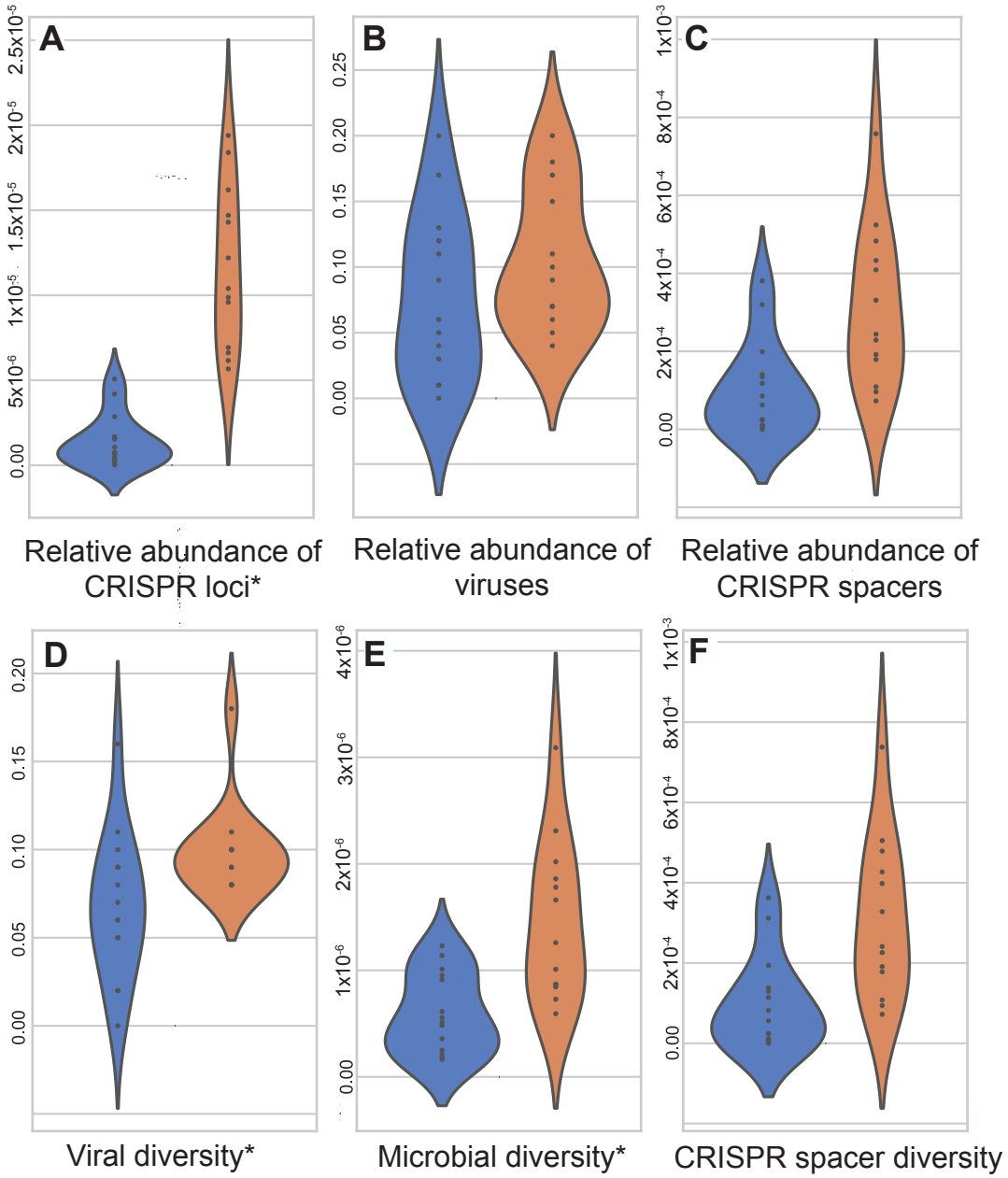
Supplementary Table captions

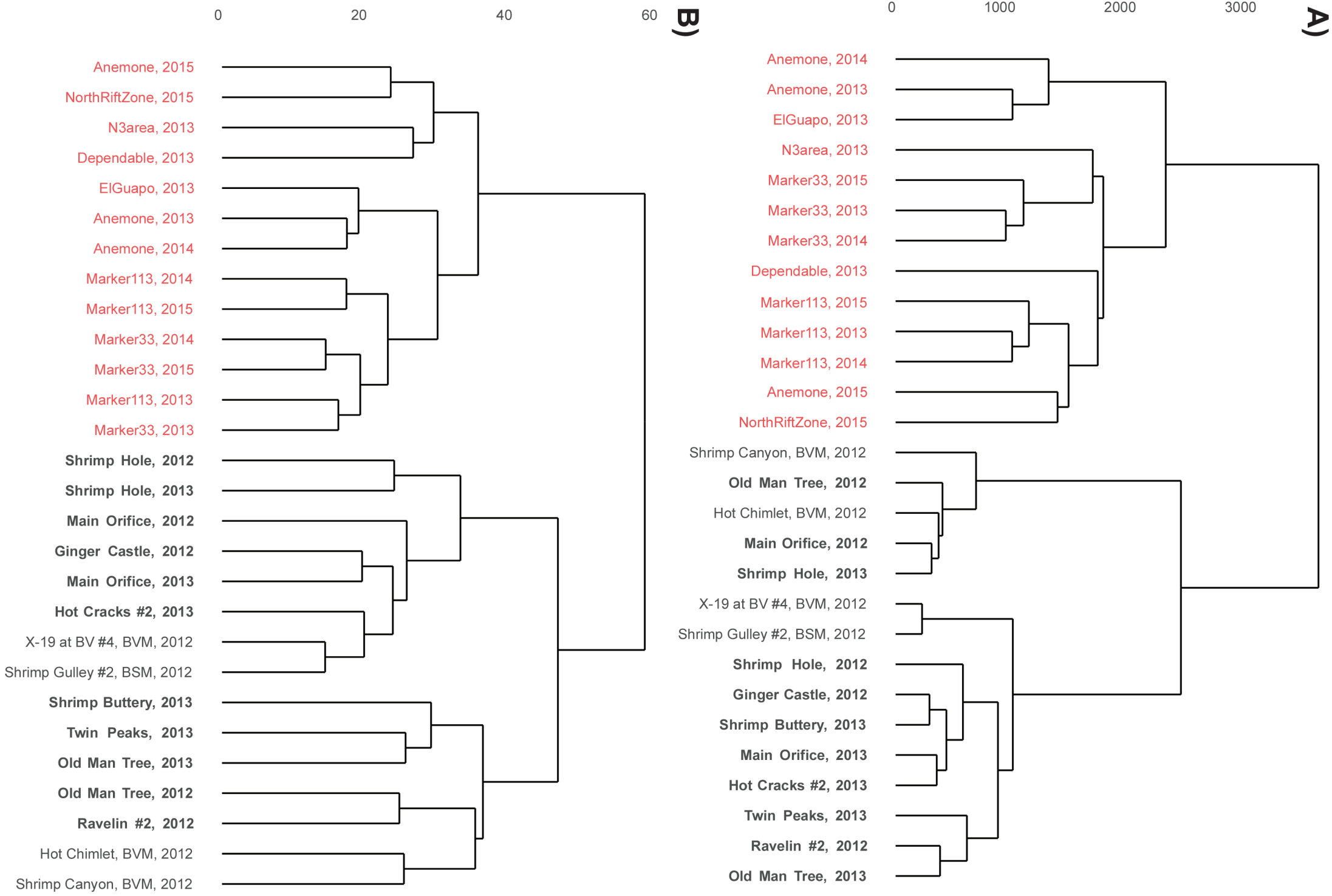
Supplementary Table 1. Information regarding sample location, sample type, metagenomic sequencing, and assembly of metagenomic reads.

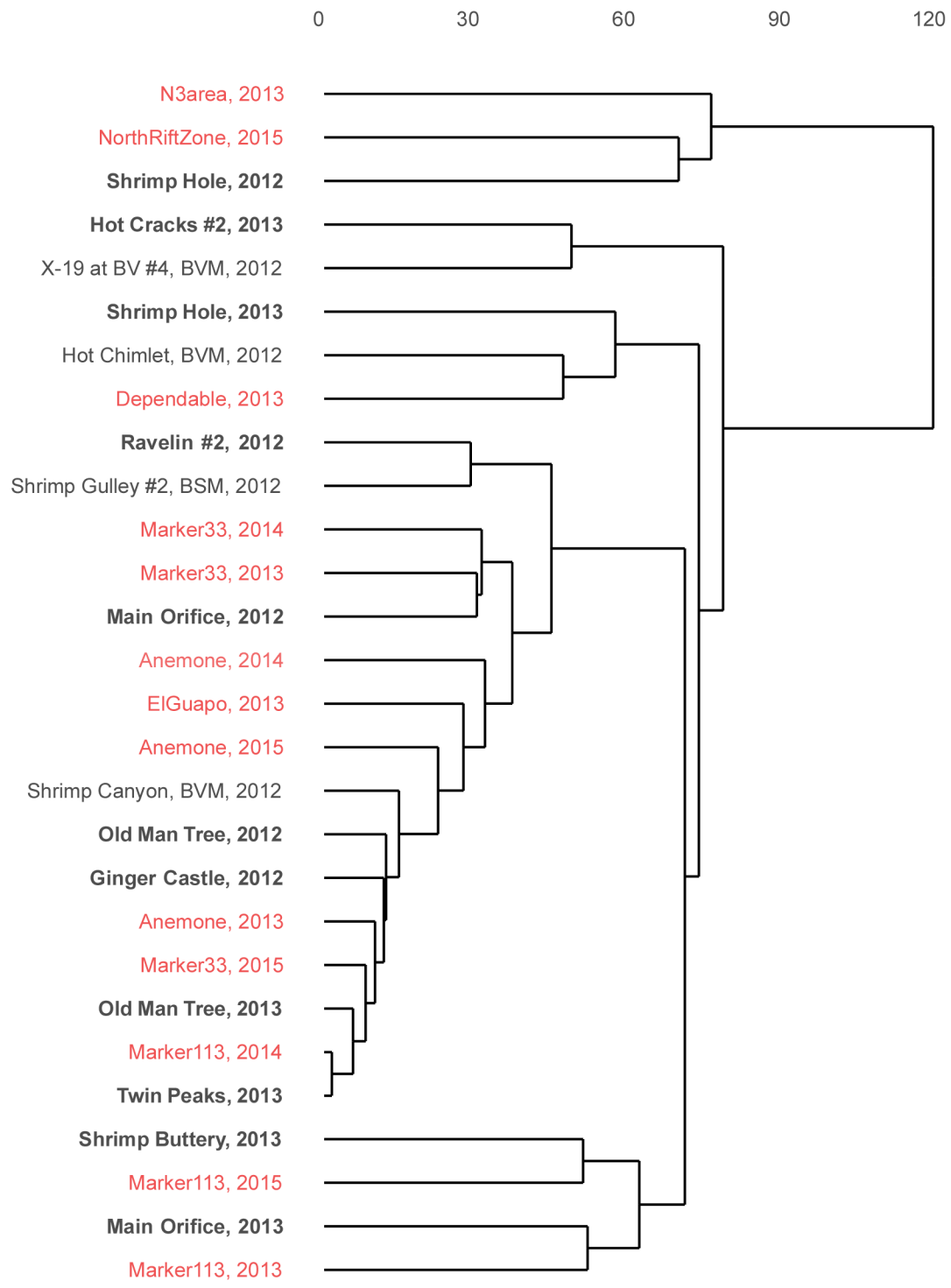
Supplementary Table 2. Relative abundance of viral sequences, spacers, and CRISPRs in metagenomes; in addition, diversity of viral clusters, spacer clusters, and 16S rRNA host classes. The number of reads that mapped to viral contigs per read, spacers per read, and CRISPR direct repeat types per read were used to measure the relative abundance of viral sequences, spacers, and CRISPRs, respectively. We used the number of viral clusters per contig, spacer clusters per read, and different classes per read matching 16S rRNA as measures of viral diversity, spacer diversity, and host class diversity, respectively. Reads mapping to 16S rRNA sequences that were classified as eukaryotes were excluded.

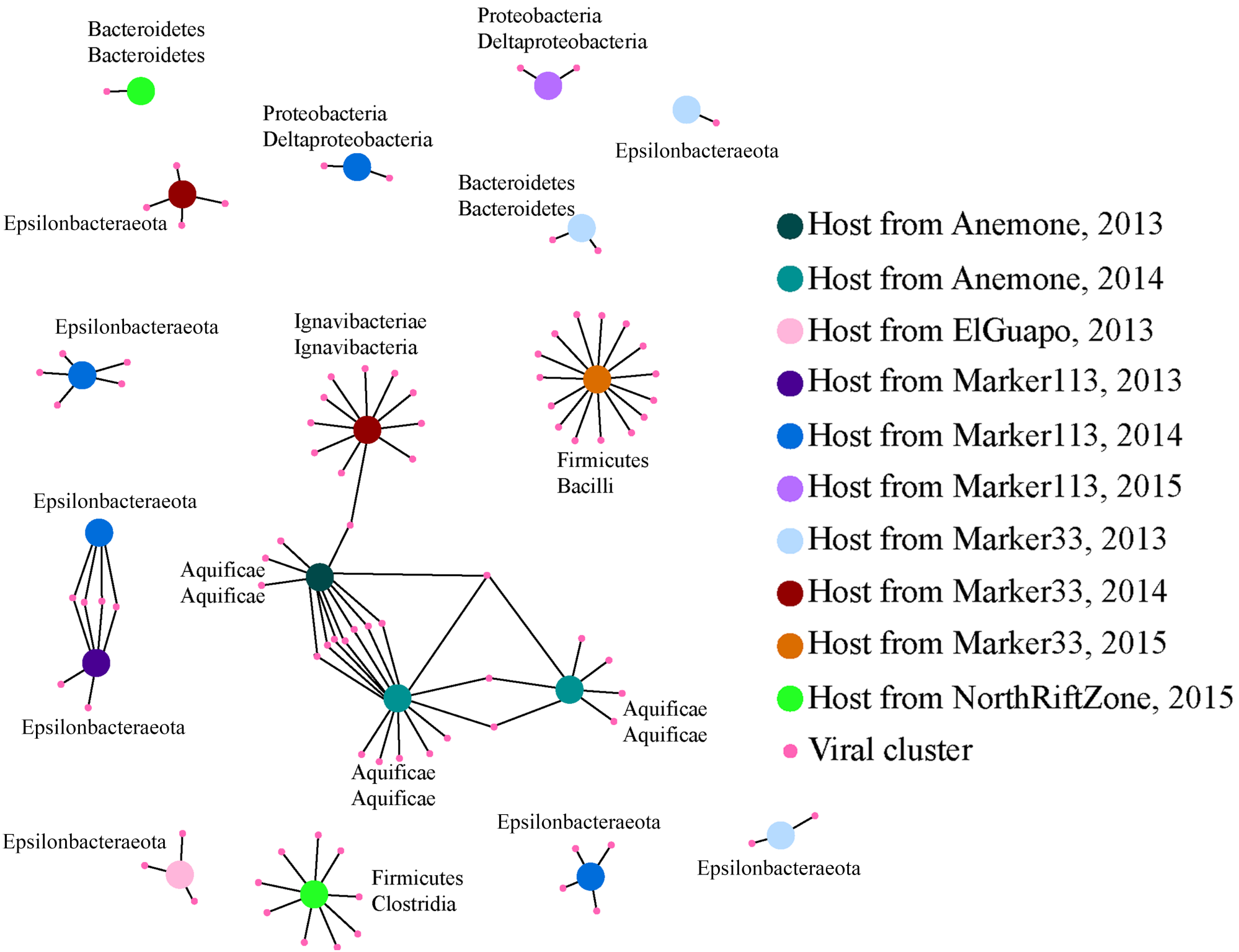
Supplementary Table 3. Percent completeness, percent redundancy, and normalized coverage of each MAG.

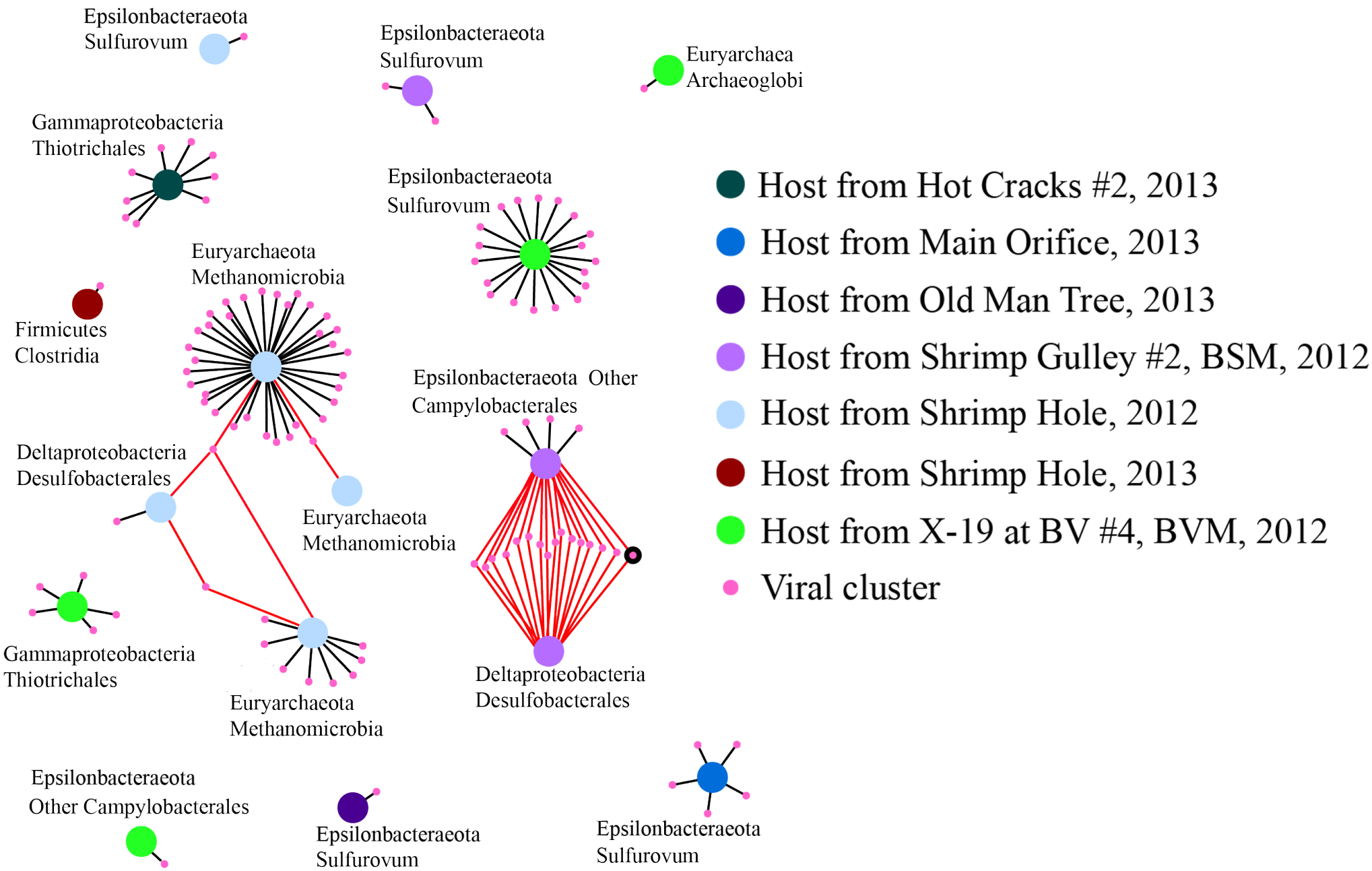
Supplementary Table 4. The number of viral and prophage contigs identified by VirSorter in each MAG in addition to the number of viral clusters linked to each MAG through CRISPR sequences**.**


A
B
C
D
E
F
Relative abundance of CRISPR loci*
Relative abundance of viruses
Relative abundance of CRISPR spacers
Viral diversity*
Microbial diversity*
CRISPR spacer diversity


B)
0
20
40
60
Anemone, 2015
NorthRiftZone, 2015
N3area, 2013
Dependable, 2013
ElGuapo, 2013
Anemone, 2013
Anemone, 2014
Marker113, 2014
Marker113, 2015
Marker33, 2014
Marker33, 2015
Marker113, 2013
Marker33, 2013
Shrimp Hole, 2012
Shrimp Hole, 2013
Main Orifice, 2012
Ginger Castle, 2012
Main Orifice, 2013
Hot Cracks #2, 2013
X-19 at BV #4, BVM, 2012
Shrimp Gulley #2, BSM, 2012
Shrimp Buttery, 2013
Twin Peaks, 2013
Old Man Tree, 2013
Old Man Tree, 2012
Ravelin #2, 2012
Hot Chimlet, BVM, 2012
Shrimp Canyon, BVM, 2012
A)
0
1000
2000
3000
Anemone, 2014
Anemone, 2013
ElGuapo, 2013
N3area, 2013
Marker33, 2015
Marker33, 2013
Marker33, 2014
Dependable, 2013
Marker113, 2015
Marker113, 2013
Marker113, 2014
Anemone, 2015
NorthRiftZone, 2015
Shrimp Canyon, BVM, 2012
Old Man Tree, 2012
Hot Chimlet, BVM, 2012
Main Orifice, 2012
Shrimp Hole, 2013
X-19 at BV #4, BVM, 2012
Shrimp Gulley #2, BSM, 2012
Shrimp Hole, 2012
Ginger Castle, 2012
Shrimp Buttery, 2013
Main Orifice, 2013
Hot Cracks #2, 2013
Twin Peaks, 2013
Ravelin #2, 2012
Old Man Tree, 2013


0
30
60
90
120
N3area, 2013
NorthRiftZone, 2015
Shrimp Hole, 2012
Hot Cracks #2, 2013
X-19 at BV #4, BVM, 2012
Shrimp Hole, 2013
Hot Chimlet, BVM, 2012
Dependable, 2013
Ravelin #2, 2012
Shrimp Gulley #2, BSM, 2012
Marker33, 2014
Marker33, 2013
Main Orifice, 2012
Anemone, 2014
ElGuapo, 2013
Anemone, 2015
Shrimp Canyon, BVM, 2012
Old Man Tree, 2012
Ginger Castle, 2012
Anemone, 2013
Marker33, 2015
Old Man Tree, 2013
Marker113, 2014
Twin Peaks, 2013
Shrimp Buttery, 2013
Marker113, 2015
Main Orifice, 2013
Marker113, 2013


Bacteroidetes
Bacteroidetes
Proteobacteria
Deltaproteobacteria
Proteobacteria
Deltaproteobacteria
Epsilonbacteraeota
Epsilonbacteraeota
Bacteroidetes
Bacteroidetes
Epsilonbacteraeota
Ignavibacteriae
Ignavibacteria
Firmicutes
Bacilli
Epsilonbacteraeota
Aquificae
Aquificae
Epsilonbacteraeota
Aquificae
Aquificae
Aquificae
Aquificae
Epsilonbacteraeota
Epsilonbacteraeota
Firmicutes
Clostridia
Epsilonbacteraeota
Host from Anemone, 2013
Host from Anemone, 2014
Host from ElGuapo, 2013
Host from Marker113, 2013
Host from Marker113, 2014
Host from Marker113, 2015
Host from Marker33, 2013
Host from Marker33, 2014
Host from Marker33, 2015
Host from NorthRiftZone, 2015
Viral cluster


Epsilonbacteraeota
Sulfurovum
Epsilonbacteraeota
Sulfurovum
Euryarchaea
Archaeoglobi
Gammaproteobacteria
Thiotrichales
Epsilonbacteraeota
Sulfurovum
Euryarchaeota
Methanomicrobia
Firmicutes
Clostridia
Epsilonbacteraeota Other
Campylobacterales
Deltaproteobacteria
Desulfobacterales
Euryarchaeota
Methanomicrobia
Gammaproteobacteria
Thiotrichales
Euryarchaeota
Methanomicrobia
Deltaproteobacteria
Desulfobacterales
Epsilonbacteraeota
Other Campylobacterales
Epsilonbacteraeota
Sulfurovum
Epsilonbacteraeota
Sulfurovum
Host from Hot Cracks #2, 2013
Host from Main Orifice, 2013
Host from Old Man Tree, 2013
Host from Shrimp Gulley #2, BSM, 2012
Host from Shrimp Hole, 2012
Host from Shrimp Hole, 2013
Host from X-19 at BV #4, BVM, 2012
Viral cluster