Diverse viruses in deep-sea hydrothermal vent fluids have restricted dispersal across ocean basins

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 microbial metagenomes generated from venting fluids venting fluids from several vent sites within three different geochemically and geographically distinct hydrothermal systems: Piccard and Von Damm vent fields at the Mid-Cayman Rise in the Caribbean Sea, and at several vent sites within Axial Seamount in the Pacific Ocean. Analysis of viral sequences and CRISPR spacers revealed highly diverse viral assemblages and evidence of active infection. 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, we show that hydrothermal vent fluids are home to novel, diverse viral assemblages that are highly localized to specific regions and taxa.

Importance

Viruses play important roles in manipulating microbial ecology and evolution in the ocean, yet not much is known about viruses in deep-sea hydrothermal vents. However, viral ecology and evolution are of particular interest in hydrothermal vent habitats because of their unique nature: previous studies have indicated that most viruses in hydrothermal vents are lysogenic rather than lytic, and it has been established that rates of horizontal gene transfer (HGT) are particularly high among thermophilic vent microbes, and viruses are common vectors for HGT. If viruses have broad host range or are widespread across vent sites, they have increased potential to act as gene-sharing "highways" between vent sites. By examining viral diversity, distribution, and infection networks across disparate vent sites, this study provides the opportunity to better characterize and constrain the viral impact on hydrothermal vent microbial communities. We show that viruses in vent sites are diverse and apparently active, but have restricted host range and are not widely distributed among vent sites. Thus, the impacts of viral infection are likely to be highly localized and constrained to specific taxa in these habitats.

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 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.

83 One large gap in our understanding of eco-evolutionary dynamics in the deep sea is how
84 broadly viruses are distributed within and between vent systems, and how viral interactions shift
85 across hydrothermal vent types. Previous work has shown that microbial communities exhibit high
86 endemism locally (4, 7, 26, 27), but some cosmopolitan species are distributed globally (28–30).
87 However, less is known about the geographic distribution of viral populations. If viruses exhibit a
88 restricted distribution, this would limit their role as vectors of gene flow across and between
89 hydrothermal systems. Microbial populations and their viruses are also limited by environmental
90 conditions: hydrothermal systems hosted in basalt rocks are characterized by metal-enriched, low pH
91 fluids up to 400°C (31). In contrast, hydrothermal systems hosted in peridotite are influenced by
92 serpentinization and feature organic carbon-enriched, high pH fluids with slightly lower temperature
93 (32). Microbial populations in basalt-hosted and peridotite-hosted vents exhibit distinct patterns of
94 genomic variation (33), suggesting that microbes are subject to different selection pressures
95 depending on the vent type, but we know little about the role viruses play in driving those
96 differences, nor how viral diversity varies across hydrothermal systems.
97
98 To gain further insight into viral diversity, distribution, and host range across hydrothermal
99 systems, we compared viral sequences recovered from microbial metagenomes collected from two
100 hydrothermal regions in two distinct ocean basins: Axial Seamount, a submarine volcano located on
101 the Juan de Fuca Ridge in the Pacific Ocean at ~1520 meters depth, and the Mid-Cayman Rise, an
102 ultraslow spreading ridge in the Caribbean Sea. Axial Seamount is a basalt-hosted, magma-driven
103 system with fluids that are low in pH and high in carbon dioxide and hydrogen sulfide (34).
104 Microbial communities at Axial are spatially restricted but temporally stable at individual vent sites
105 over time (4, 27). In contrast, the Mid-Cayman Rise hosts two geologically and geochemically
106 distinct hydrothermal vent fields in close proximity to each other: Piccard hydrothermal field,
107 located in basalt rocks along the ridge axis, is the deepest hydrothermal field discovered to date
108 (~4950 meters depth) and is characterized by fluids that are acidic with high hydrogen and hydrogen
109 sulfide content, whereas Von Damm vent field, located approximately 20 km away on a nearby
110 massif at ~2350 meters depth, is influenced by serpentinization and is characterized by fluids that
111 are high in hydrogen, methane, and small carbon compounds (35–38). Previous work found distinct
112 community composition but similar functional potential in the microbial communities at Piccard and
113 Von Damm (7, 27, 33, 39, 40) . Here, we compare viral sequences in venting fluids from Von
114 Damm vent field, Piccard vent field, and Axial Seamount in a comparative survey of viral diversity

115 and gene content across geographically and geochemically distinct hydrothermal vent sites. By
116 examining viruses from the microbial fraction (0.22 μm), we are closely examining the relationship
117 between prophage, actively infecting viruses, and their hosts. We examined viral sequences identified
118 in the metagenomes as well as CRISPR sequences, which provide context for historical viral
119 infections within each of the communities, and provide a means to match viruses with putative
120 hosts. We show that viral populations have high diversity but restricted distribution and host range
121 across hydrothermal systems, limiting the capacity of viruses to act as vectors of gene flow between
122 disparate hosts and vent locations.
123
124 **Methods**
125
126 *Sample collection and DNA preparation and sequencing*
127 Low-temperature diffuse flow fluid samples were collected from the vent fields Piccard and
128 Von Damm at the Mid-Cayman Rise in January 2012 and June 2013 during research cruises on the
129 R/V Atlantis and R/V Falkor, respectively. We analyzed a total of 11 metagenomes from 8 different
130 vents at Von Damm vent field and 4 metagenomes from 4 different vents at Piccard vent field.
131 Sample locations, depth, and metagenomic data are described in Supplementary Table 1. The ROV
132 Jason II and Mat Sampler were used to collect the 2012 Mid-Cayman Rise samples, as previously
133 described in (41). The 2013 Mid-Cayman Rise samples were collected using the SUPR version 2
134 sampler and HROV Nereus (42). For microbial DNA collection, approximately 3 to 6 L of diffuse
135 flow fluid were pumped through 0.22 μm Sterivex filters (Millipore). Shipboard, the filters were
136 flooded with RNALater (Ambion), sealed with Luer Caps, stored in sterile Falcon tubes, and frozen
137 at −80 °C. Sample collection and preservation are further described in Anderson et al. (33) and
138 Reveillaud et al. (40). Genomic DNA was extracted and metagenomic libraries constructed as
139 described in Anderson et al. (33). Sequencing was done on an Illumina Hi Seq 1000 at the W.M.
140 Keck Facility in the Josephine Bay Paul Center at the Marine Biological Laboratory.
141
142 Diffuse flow fluid samples were collected from Axial Seamount in September 2013, August
143 2014, and August 2015 (approximately five months after the eruption of Axial Seamount) during
144 research cruises on the R/V Falkor and R/V Thompson in 2013, R/V Brown in 2014, and R/V
145 Thompson in 2015 (Supplementary Table 1). Diffuse flow samples were collected from four
146 different vent fields within Axial Seamount (ASHES, International District, North Rift Zone, and

147 Dependable) using the ROVs ROPOS and Jason. We analyzed a total of 16 metagenomes from 10
148 different vents, 1 plume, and 2 deep seawater samples at Axial Seamount. For collection of microbial
149 DNA, 3 L of diffuse fluid was pumped through 0.22 μm, 47 mm GWSP filters (Millipore) and the
150 filters were flooded with RNALater (Ambion) on the seafloor, as described in Fortunato et al. (4).
151 Fluids from a hydrothermal plume above Anemone and background seawater were collected using a
152 Seabird SBE911 CTD and 10 L Niskin bottles and 3 L of the plume and seawater fluid was filtered
153 through 0.22 μm Sterivex filters (Millipore). DNA was extracted from the filters using a phenol-
154 chloroform method adapted from Crump et al. (43) and Zhou et al. (44). The Ovation Ultralow
155 Library DR multiplex system (Nugen) was used to prepare metagenomic libraries following the
156 manufacturer instructions. DNA extraction and metagenomic library construction is further
157 described in Fortunato and Huber (45). The 2013 and 2014 samples were sequenced on an Illumina
158 HiSeq 1000 and the 2015 samples on a NextSeq 500. Sequencing was done at the W.M. Keck
159 sequencing facility at the Marine Biological Laboratory.
160
161 For all metagenomes, paired-end partially overlapping reads were merged and quality filtered
162 using the illumina-utils package (46) using the iu-quality-filter-minoche flag, then assembled using
163 idba-ud (47) v1.1.2 with default settings. All data from both sites is available in the European
164 Nucleotide Archives archives under study accession number PRJEB15541 for the Mid-Cayman Rise
165 and under study accession numbers PRJEB7866, PRJEB12000, and PRJEB19456 for Axial samples
166 in the years 2013, 2014 and 2015 respectively (Supplementary Table 1).
167
168 *Identification of CRISPR loci, viral clusters, and spacer clusters*
169 Crass (48) v1.0.1 was used to identify CRISPR loci (i.e., unique direct repeat types) and
170 CRISPR spacers in the metagenomic reads. Viral contigs in the metagenomic assemblies were
171 identified with VirFinder (48, 49) v1.1 using a p-value threshold of 0.05. Given that viruses from
172 hydrothermal systems are not well-represented in sequence databases, we elected to use VirFinder
173 for identification of viral contigs for diversity and abundance analysis because VirFinder is a k-mer
174 frequency-based method that avoids gene-based similarity searches, and thus has higher potential to
175 identify novel viruses. To cluster the CRISPR spacers, we performed an all-v-all blast using
176 BLASTn v2.5.0 (E-value threshold of 10^{-08}) and then clustered using Markov Cluster Algorithm
177 (MCL) (50) v14-137 (inflation 1.2 and scheme 7) based on bitscore. We used ClusterGenomes

178 (https://bitbucket.org/MAVERICLab/stampede-clustergenomes) v1.1.3 (95% identity, 80%
179 coverage) to cluster viral contigs.
180
181 *Recovery of MAGs*
182 All metagenome assembled genomes (MAGs) were recovered from metagenomic assemblies
183 using anvi'o (51). Supervised clustering was used to recover bins from Piccard and Von Damm
184 contigs using anvi'o v2.1.0 (33). MAGs from Axial Seamount assemblies were recovered using
185 unsupervised binning with CONCOCT (50, 52, 53) within the anvi'o v4.0 pipeline, followed by
186 manual refinement within anvi'o (39). For all analyses, only bins with >=70% completeness and
187 <=10% redundancy were retained as MAGs for this analysis. The MAGs were assigned taxonomies
188 using PhyloSift (50, 52) v1.0.1 as described in Anderson et al. (33). Only MAGs for which
189 taxonomies could be identified were used in analyses. MAG coverage was normalized by the
190 number of merged reads in the metagenome. All of the Mid-Cayman Rise MAGs were previously
191 described in Anderson et al. (33).
192
193 *Taxonomic identification of microbial and viral reads*
194 The reads from all metagenomes were mapped to the Silva SSU and LSU Parc databases (54,
195 55) (release 132) with bowtie2 (56) v2.2.9 using default settings and local alignment. Mapped reads
196 were assigned taxonomies using the classify.seqs function in mothur (57, 58) v1.38.1 with the Silva
197 16S rRNA database (release 119) and a cutoff of 50. Reads mapping to 16S rRNA gene sequences
198 that were classified as eukaryotes were excluded from analyses.
199 In order to identify the viral taxa at Von Damm, Piccard, and Axial Seamount, we conducted
200 a blast of the representative sequence of each viral cluster against all of the viral genomic sequences
201 in RefSeq (downloaded on September 16th, 2020; E-value <=10-5). The best match for each viral
202 cluster representative sequence was selected based on E-value and percent identity. The taxonomy
203 of the RefSeq viral sequence with the best alignment was assigned to the given viral cluster. To
204 quantify the relative abundance of viral families in each sample, the reads that mapped to viral
205 clusters of a viral family were divided by the total number of reads in the sample.
206
207 *Viral, spacer, and host diversity*
208 Rarefaction curves for viral clusters, CRISPR spacer clusters, and reads matching 16S rRNA
209 genes categorized at the class level were created using the Vegan R package (59) v2.4-5. The number

210 of viral clusters per contig and spacer clusters per read (paired reads) were used as proxies for viral
211 diversity, and the number of different taxa at the class level matching 16S rRNA genes reflected
212 microbial diversity, calculated on a per read basis (merged reads).
213
214 *Viral, spacer, and CRISPR relative abundance*
215 To calculate relative viral abundance, the reads in each of the metagenomes were mapped
216 against all of the viral contigs from the corresponding geographic region using bowtie2 (56) v2.2.9.
217 The reads from each sample were mapped against all of the viral contigs from the corresponding
218 geographic region rather than solely the viral contigs in the sample because there were viral reads in
219 samples that did not assemble into viral contigs. This method therefore allowed the identification of
220 more viral sequences. The number of reads that mapped to viral contigs was normalized by the
221 number of merged reads as a measure of relative viral abundance. This measure of relative viral
222 abundance reflects only the proportion of viruses that were retained on the filter as viral capsids or
223 prophages. We used the number of spacers per read and CRISPR direct repeat types per read as
224 measures of spacer and CRISPR relative abundance, respectively. Paired rather than merged reads
225 were used for these analyses. Relative abundance and diversity of viruses, microbes, CRISPR spacers
226 and CRISPR loci were visualized using the Seaborn library within Python (60).
227
228 *Relative compositions and abundances of viral clusters, spacer clusters, and hosts*
229 The relative compositions of the microbial community, the viral assemblage, and CRISPR
230 spacers were compared between vent sites within Von Damm, Piccard, and Axial Seamount. To
231 calculate the relative abundance of each viral cluster in each sample, the number of reads in the
232 sample that mapped to the viral contigs in the viral cluster was determined using bowtie2 (56) v2.2.9.
233 The number of reads in each metagenome that mapped to the viral cluster was normalized by the
234 total length of the viral contigs in the viral cluster and the number of merged reads in the
235 metagenome. We defined the most common viral clusters as the six clusters with the highest relative
236 abundance in each sample. The relative abundance of each CRISPR spacer cluster in each sample
237 was calculated as the percent of spacers in the sample that were part of the spacer cluster. For spacer
238 clusters, we defined the most common as the three clusters with the highest relative abundance in
239 each sample. To compare the compositions of viral contigs and CRISPR spacers, all of the spacers
240 were blasted against all of the viral contigs using blastn (E-value $<=10^{-05}$, <=1 mismatch, as per
241 Emerson et al. (61). The relative abundance of each microbial host was measured as the number of

242 reads that mapped to 16S rRNA gene sequences of the given taxon, normalized by the number of
243 16S rRNA gene reads in the sample. Analyses of microbial taxa were done at either the class level or
244 the lowest taxonomic level available, depending on the analysis.
245
246 Microbiome datasets are compositional in nature because sequencing instruments impose an
247 arbitrary total (62). Therefore, to conduct hierarchical clustering of samples based on viral, spacer,
248 and host composition, we used the protocol outlined by Gloor et al. (62) and Gloor and Reid (63)
249 for computing distances between samples containing compositional data. For hierarchical clustering,
250 we did not normalize; we performed analyses on the number of reads that mapped to each viral
251 cluster, the number of CRISPR spacers in each spacer cluster, and the number of reads that mapped
252 to 16S rRNA gene sequences of each microbial host in each sample; for microbial hosts, we did not
253 include reads that mapped to unclassified sequences or sequences classified as eukaryotes. We
254 replaced zero counts with estimates using the count zero multiplicative method for viral clusters and
255 hosts and the Bayes-Laplace Bayesian multiplicative method for spacer clusters via the
256 zCompositions R package (64) v1.2.0. Using the CoDaSeq R package
257 (https://github.com/ggloor/CoDaSeq) (62) v0.99.3, we applied a centered log-ratio (clr)
258 transformation to the count data (that lacked zero counts), thereby capturing the ratios between
259 parts. To calculate distances between samples for hierarchical clustering, we used the ward.D2
260 method on the transformed counts.
261
262 *Networks of viral infection*
263 Infection networks of viral-host interactions were created using CRISPR sequences to
264 connect microbial hosts with clusters of viral contigs, adapted from the methods used by Daly et al.
265 (65) and Emerson et al. (61). First, MAGs were connected to CRISPR direct repeat types within
266 each sample using BLASTn v2.5.0 (E-value $<=10^{-10}$, 100% nucleotide identity, as per Emerson et al.
267 (61). Then, each CRISPR direct repeat type was matched to a set of CRISPR spacers as identified by
268 Crass (Skennerton et al. 2013) v1.0.1. Finally, the CRISPR spacers in each sample were matched to
269 viral contigs in the corresponding region using BLASTn v2.5.0 with an E-value cutoff of 10^{-05} and a
270 maximum of one mismatch, as per Emerson et al. (61). Only one mismatch was allowed because
271 resistance has been found to be lost by single nucleotide differences between bacterial spacers and
272 target phage sequences (66). In contrast, resistance by archaeal CRISPR systems can still be provided
273 when there are up to three mismatches between spacers and target phage sequences (67, 68). We did

274 not allow for more mismatches between archaeal spacers and phage sequences because resistance is
275 weakened by more mismatches (68). Each direct repeat type provided by Crass v1.0.1 does not
276 necessarily represent an individual CRISPR locus (48).
277
278 *Data analysis software*
279 The majority of analyses were conducted in RStudio (69) v1.0.136. R packages used were
280 readr (70) v1.1.1, readxl (71) v1.0.0, tidyr (72) v0.7.2, stringr (73) v1.3.0, dplyr (74) v0.7.4, ggnetwork
281 (75) v0.5.1, statnet (76) v2016.9, ggpubr (77) v0.1.7, ggplot2 (78) v2.2.1, and svglite (79) v1.2.1.
282
283 **Results**
284
285 *Taxonomy of viruses and hosts*
286 All analyses were carried out from diffuse flow fluids sampled directly from the vent orifice, as
287 well as vent plume waters, which were sampled up to 100m above the vent orifice and had much
288 lower temperatures (Supplementary Table 1). It is possible to examine viruses in 0.22 μm-filtered
289 fluids because this fraction includes integrated prophage, lytic infections in progress, and free viral
290 particles captured on filters. However, our analysis misses free viral particles that were not retained
291 on the filters, and we are only capturing the viral diversity and variation across sites based on the
292 viral sequences identified in the metagenomes.
293 The viral sequences identified at Axial Seamount, Von Damm, and Piccard vents represent 28
294 families of viruses (Figure 1). 13 of the viral families were present at all three vent fields. 14 viral
295 families were present at both of the Mid-Cayman Rise vent fields, Von Damm and Piccard, while
296 Axial Seamount shared 13 viral families with each of the Mid-Cayman Rise vent fields. We observed
297 a distinct difference in viral taxonomic groups between the vent fields: whereas the *Myoviridae* viral
298 family had the highest relative abundance in 12 of the 16 Axial Seamount samples, the *Myoviridae*
299 were in much lower abundance at Von Damm and Piccard, which had higher abundances of reads
300 matching the *Podoviridae* family of viruses and the the *Guttaviridae*.
301
302 *Relative abundance of viral and microbial reads*
303 In order to examine the distribution and abundance of viruses across individual vent sites,
304 vent fields, and regions, we quantified the relative abundance of viral sequences, CRISPRs, and
305 spacers in each of the metagenomes (Figure 2, samples from background seawater and plumes are

306 excluded; Supplementary Table 2). We also quantified the diversity and abundance of loci and
307 spacers of Clustered Regularly InterSpaced Palindromic Repeats (CRISPR) found in each of the
308 metagenomes. CRISPR loci are a microbial immune system found in archaeal and bacterial genomes,
309 consisting of direct repeats interspersed by "spacers," which match foreign DNA (predominantly
310 viruses, but also including plasmids and other forms of foreign DNA) that the cell has been exposed
311 to previously. The relative abundance of CRISPR loci serves as an indication of how many microbial
312 lineages use CRISPR as a mechanism for viral immunity. It is also important to note that CRISPR
313 loci vary in number across microbial genomes, and we did not distinguish between CRISPR loci
314 with the same direct repeat type. Therefore, our measure of CRISPR relative abundance is not a
315 direct proxy for the abundance of viruses. Instead, it gives an indication of how commonly CRISPR
316 loci are used as an antiviral mechanism within the community. Moreover, CRISPRs serve as a record
317 of past infections, and thus while viral diversity reflects the diversity of viral particles sequenced at
318 the time of sampling, CRISPR spacer diversity reflects the diversity of past viral infections.
319 Our results indicate that the relative abundance of viral sequences (Figure 2A) was similar
320 both across and within each of the hydrothermal vent regions we studied ($p = 0.416$, t-test). There
321 was a higher relative abundance of CRISPR loci in samples collected at Axial Seamount compared to
322 both Piccard and Von Damm vent fields (Figure 2B; $p = 1.49e-05$, t-test). However, there was no
323 difference in the relative abundance of spacers within CRISPR loci per read (Figure 2C; $p = 0.101$, t-
324 test). Within the Mid-Cayman Rise, we compared samples from mafic-hosted (Piccard) versus
325 ultramafic-hosted (Von Damm) hydrothermal systems. We did not observe significant differences in
326 the relative abundance of viral sequences, CRISPR loci, or CRISPR spacers between Piccard and
327 Von Damm (viral sequences: $p = 0.66$, t-test, CRISPR loci: 0.40, t-test, spacers: 0.37, t-test) (Figure
328 2A-C). Similar results emerged from comparisons among vent fields within Axial Seamount: we did
329 not observe significant differences in the relative abundance of viral sequences, CRISPR loci, or
330 CRISPR spacers between vent fields at Axial Seamount (viral sequences: $p = 0.17$, t-test, CRISPR
331 loci: 0.41, t-test, spacers: 0.11, t-test). Finally, within Axial Seamount, we also compared the relative
332 abundance of viruses in samples taken from plume and diffuse flow hydrothermal fluid at Anemone
333 vent. The Anemone diffuse flow samples had a higher relative abundance of CRISPR spacers and
334 CRISPR loci compared to the Anemone plume sample, which was sampled above the vent.
335 However, the relative abundance of viral sequences did not differ between the Anemone plume and
336 diffuse flow samples (Supplementary Table 2).
337

338 *Diversity of viral assemblage and microbial community*

339 Overall, viral diversity analyses revealed that the viral assemblages within these vents had high
340 richness and were not dominated by specific viral strains. With the exception of a few samples, the
341 rarefaction curves for the viruses and microbes did not reach saturation (Supplementary Fig. 1),
342 indicating that the viral clusters recovered in this study did not capture the total diversity in the
343 samples. In each sample, there were no dominant viral or spacer clusters, and each of the viral and
344 spacer clusters present was relatively rare (Supplementary Figs. 2-3). Moreover, the viral and spacer
345 clusters with the highest coverage did not correspond across samples: in both Piccard and Axial
346 Seamount, one of the most common spacer clusters matched with one of the most common viral
347 clusters by BLAST, while none of the most common spacer and viral clusters aligned in Von
348 Damm. Thus, there were no dominant viral sequences that were consistently found across all
349 samples.
350
351 We observed a more diverse viral assemblage at Axial Seamount compared to the Mid-
352 Cayman Rise. We observed higher richness of both viruses (Figure 2D; $p = 0.0148$, t-test) and their
353 microbial hosts (Figure 2E; $p = 0.00563$, t-test) in samples from vents at Axial Seamount compared
354 to samples from Von Damm and Piccard vent fields. However, we did not observe meaningful
355 correlations between viral cluster diversity, spacer cluster diversity, or host diversity. Moreover, we
356 did not observe significant differences in either viral or host diversity between the Piccard and Von
357 Damm vent fields at the Mid-Cayman Rise (viral diversity: $p = 0.05306071$, t-test, host diversity: $p =$
358 0.6648, t-test). We also examined the relative diversity of CRISPR spacers, which provides an
359 indication of past viral infections rather than the current virus pool. We did not observe a significant
360 difference in the diversity of CRISPR spacers between samples from Piccard, Von Damm, and Axial
361 Seamount (Figure 2F, $p = 0.0927$, t-test).
362
363 We observed significant differences in viral diversity between vent sites within Axial Seamount
364 ($p = 0.02165$, t-test). However, no significant differences emerged in terms of the diversity of
365 microbial hosts and CRISPR spacers (host diversity: $p = 0.7444$, t-test, spacer diversity: $p = 0.1081$,
366 t-test). At Anemone vent, the diffuse flow samples had higher CRISPR spacer and microbial
367 diversity than the plume sample, but viral diversity did not differ between these samples
368 (Supplementary Table 2).
369

370 *Viral distribution across vent sites*

371 In order to characterize viral distribution and community similarity across hydrothermal vent
372 fluids, we evaluated the extent to which viral sequences and CRISPR spacers were distributed across
373 samples, then compared these results to the host microbial community. As before, we clustered
374 sequences based on similarity to compare across all diffuse flow samples, excluding background
375 seawater and plume samples. On the whole, viral sequences and CRISPR spacers had fairly limited
376 distributions. Only a few viral clusters (0.02%) were present at all vent sites (Von Damm, Piccard,
377 and Axial Seamount), and no CRISPR spacer clusters were present in all three regions (Table 1). The
378 most cosmopolitan viruses and CRISPR spacers did not match each other: at Von Damm, two of
379 the 100 most widely distributed viral clusters matched with three of the 100 most widely distributed
380 CRISPR spacer clusters according to BLAST. At Piccard, one of the most widely distributed viral
381 clusters aligned to one of the most widely distributed spacer clusters. At Axial Seamount, one of the
382 100 most widely distributed viral clusters matched with two of the 100 most widely distributed
383 CRISPR spacer clusters. In contrast, microbial taxa were much more cosmopolitan, with ~17% of
384 taxa shared between Von Damm, Piccard, and Axial Seamount according to 16S rRNA gene
385 classification at the lowest taxonomic level available (Table 1). Viral and CRISPR spacer clusters
386 were shared more widely among vent sites within Von Damm and Piccard compared to Axial
387 Seamount, but microbial lineages were shared more widely among vent sites at Axial Seamount
388 (Table 1).

389 We created hierarchical dendrograms to assess the similarity of samples based on their viral
390 content. Viral assemblages in samples from the Mid-Cayman Rise and Axial Seamount grouped
391 separately (Figure 3A). Within Axial Seamount, samples taken in successive years from the same
392 vent tended to have similar viral assemblage compositions (Figure 3A). In contrast, at the Mid-
393 Cayman Rise, samples taken from the same site in two different years did not cluster together, and
394 we observed weak clustering of samples by location (Figure 3A). Samples from Piccard vent field
395 and Von Damm vent field did not cluster separately. Grouping of microbial communities based on
396 classification of 16S rRNA reads in the metagenomes showed similar patterns to the viral
397 assemblages. Samples collected from vents at Axial and the Mid-Cayman Rise grouped separately
398 from each other, with only some clustering of samples at Piccard and Von Damm. Similarly, we
399 observed stronger temporal clustering among samples at Axial Seamount than at Von Damm
400 (Figure 3B).

401 In contrast to the viral and microbial assemblages, hierarchical dendrograms based on
402 CRISPR spacer compositions showed little clustering by location (Figure 3C). Based on spacer
403 assemblages, samples from either Axial Seamount or the Mid-Cayman Rise did not cluster together,
404 and samples taken from the same vent sites in different years did not cluster together. Very few
405 CRISPR spacer clusters were found at multiple vent sites (Table 1).
406
407 *Networks of viral infection*
408 Networks of viral infection generated from viral sequences, CRISPRs, and MAGs were used
409 to examine the distribution and host specificity of vent viruses. These virus networks show
410 connections between CRISPR spacers, which represent a record of previous viral infection in the
411 host, and viral sequences recovered from the metagenome, which represent viral sequences present
412 in the community at the time of sampling. Virus-host networks made for Axial Seamount (Figure
413 4A), Von Damm (Figure 4B), and Piccard (Figure 4C) indicate that vent viruses are restricted in
414 terms of both host range and geographic distribution. We did not observe any virus-host
415 connections shared between Piccard, Von Damm, or Axial Seamount. Within Axial Seamount, there
416 were 89 viral clusters linked to 18 MAGs, and 105 connections between distinct pairs of viral
417 clusters and MAGs (Figure 4A). We did not observe any connections between MAGs and viral
418 clusters in any of the non-diffuse flow samples (Background, CTD1200 and the Anemone plume
419 sample). The number of viral connections appeared to be related to, but was not significantly
420 correlated with, the microbial hosts' relative abundance (Supplementary Table 3): for example, a
421 Clostridia MAG in the 2015 North Rift Zone sample (Clostridia_35) had high normalized coverage
422 and was linked to nine viral clusters. We observed a high number of viruses shared among *Aquificae*
423 MAGs sampled from the Anemone vent in 2013 and 2014 (Figure 4a). Finally, we observed only a
424 single viral cluster that was linked to MAGs of different taxonomic classes from different vent sites,
425 with a viral cluster connected to an *Ignavibacteria* MAG (Ignavibacteria_15) in the 2014 Marker 33
426 sample and an *Aquificae* MAG (Aquificae_43) in the 2013 Anemone sample (Figure 4A).
427 The virus infection networks reflected the geographic separation observed in the cluster
428 dendrograms, and also revealed fairly narrow virus-host ranges for the viral sequences observed in
429 our datasets. Within the Von Damm vent field in the Mid-Cayman Rise, there were 62 viral clusters
430 linked to nine MAGs, with 66 connections between distinct pairs of viral clusters and MAGs (Figure
431 4B). The size of the viral infection network was similar in Piccard, where 48 viral clusters were
432 linked to seven MAGs and there were 64 connections between distinct pairs of viral clusters and

433 MAGs (Figure 4C). All of the viral clusters in the Von Damm and Piccard networks represented
434 relatively rare viruses with the exception of one. This viral cluster was present in seven of the Von
435 Damm vents and all of the Piccard vents, was among the top six most abundant viruses in both the
436 X-19 and Shrimp Gully #2 vents within Piccard, and was linked to a Campylobacterales and a
437 Desulfobacterales MAG, both from Shrimp Gully #2. This was the only viral cluster with high
438 relative abundance in at least one sample that was present in the Axial Seamount, Von Damm, and
439 Piccard viral infection networks (Figure 4). Some MAGs had more viral connections than others: for
440 example, a *Methanomicrobia* MAG (Methanomicrobia_41) from the 2012 sample from Shrimp Hole
441 within Von Damm was linked to 35 viral clusters (Figure 4B). However, this is not a direct
442 indication of the number of different viruses infecting a specific strain because some MAGs had
443 more CRISPR spacers than others, increasing the possibility of finding a viral connection. As with
444 Axial Seamount, there were instances of highly abundant MAGs with many viral connections
445 (Supplementary Table 3): a *Sulfurovum* MAG (Sulfurovum_99) in the 2012 sample from X-19 within
446 Piccard had the highest normalized coverage across Von Damm and Piccard and was linked to 18
447 viral clusters, making it the MAG with the third highest number of viral connections within either
448 network. Another *Sulfurovum* MAG (Sulfurovum_37) had a high relative abundance in the 2012
449 sample from Shrimp Gulley #2 within Piccard, and had two viral connections. Within Von Damm,
450 Piccard, and Axial Seamount, we observed several cases in which viruses were connected to multiple
451 microbial hosts within the network (Figure 4). However, in some cases the MAGs were linked to
452 these shared viral clusters through the same CRISPR direct repeat type and were in the same sample,
453 suggesting that these specific connections may have been due to matching CRISPR direct repeat
454 types rather than true cross-infection. These have been indicated in Figure 4 (red lines).
455
456 Identification of integrated prophage and AMGs in MAGs was conducted using VirSorter and are
457 described in the Supplementary Materials.
458
459 **Discussion**
460
461 Viruses are important drivers of microbial mortality, ecology and evolution in the ocean, but
462 studies of their distribution and impact in the deep sea are lacking. Our results indicate that viral
463 diversity is high in venting fluids from hydrothermal systems, and the viruses we analyzed have
464 restricted geographic distributions and host ranges. This implies that viruses do not spread widely

465 between vent sites and that the viral role in mediating horizontal gene transfer across taxa and
466 between vent sites is relatively restricted. However, the high abundance and diversity of viral
467 sequences as well the large number of virus-host CRISPR connections implies rapid viral mutation
468 and ongoing viral infection, indicating active and ongoing interactions between viruses and their
469 microbial hosts in venting fluids from the hydrothermal vents examined here.
470
471 *Viral assemblages are taxonomically distinct and spatially restricted across vent sites*
472 Previous studies of viruses in marine systems have observed viral populations to be
473 commonly found across multiple samples (80), whereas others have found that most viruses are
474 biogeographically restricted, with only a few cosmopolitan groups (81, 82). In hydrothermal vents at
475 Von Damm, Piccard and Axial Seamount, we found high viral diversity with a limited distribution,
476 potentially indicating rapid diversification in vents. Most viral clusters we observed were found only
477 at individual vent sites, with very few cosmopolitan viruses. Previous work examining microbial
478 distribution in hydrothermal systems through fine-scale 16S rRNA gene analyses indicates that while
479 some microbial lineages are endemic to individual vent sites, others are widespread across vent fields
480 (28, 40). Our work revealed that viral biogeographic patterns roughly corresponded to those of their
481 microbial hosts. Taxonomic analysis revealed that there are distinct differences in viral taxonomy
482 between vent sites-- most notably, we observed a high abundance of *Myoviridae* at Axial Seamount
483 that was not found at either Piccard or Von Damm. These observations indicate either restricted
484 gene flow or strong environmental selection for specific groups within each site. Our analysis of
485 viral sequence similarity using hierarchical clustering revealed similar results: there were some
486 similarities between the viral assemblages in vents at Piccard and Von Damm, which are located
487 approximately 20km apart in the Caribbean Sea, but the viral assemblages from Axial Seamount in
488 the Pacific were very different from those at Piccard and Von Damm, despite the fact that vents at
489 Piccard are geochemically more similar to those at Axial than at Von Damm. Previous work on
490 microbial community similarity and distribution in hydrothermal systems has shown that vents in
491 close proximity are often more similar to one another in terms of microbial community structure
492 than to geographically distant vents (4, 27, 40, 83). This may result from subseafloor plumbing that
493 restricts fluid flux between sites, creating "islands" of microbial diversity that are distinct from one
494 vent site to the next (26). Here, we show that these barriers to dispersal apply to viruses as well, and
495 that viruses may be even more spatially restricted than their microbial hosts. Microbial lineages that
496 spread between vent sites may face infection from novel viral strains not found in other vent sites.

497 These endemic viral populations thus further shape distinct microbial community structure at
498 individual vent sites.
499 Although the viral assemblages at hydrothermal vents had high diversity and restricted
500 dispersal between vent sites and vent fields, our results show that the viral assemblages in vents
501 persist over time, particularly at Axial Seamount where the same vents were sampled over a 3-year
502 period. Viral assemblages from samples from the same vent sites at Axial across three years clustered
503 together in the hierarchical dendrograms (Figure 3A), and our viral infection networks revealed that
504 a number of viruses at Axial Seamount were linked to specific microbial hosts at the same vent over
505 multiple years (Figure 4A). These patterns match those observed in the microbial communities
506 (Figure 3B) and are consistent with previous work showing spatially restricted but temporally stable
507 microbial communities over time at Axial Seamount (4, 27). Our work extends this to the viral world
508 at Axial Seamount, indicating that viral assemblages follow the same temporal patterns as their
509 microbial hosts, and that virus-host relationships persist over time. However, temporal stability in
510 the viral and microbial community was generally not preserved at the Von Damm vent field for sites
511 sampled in both 2012 and 2013. However, there were fewer sites and time points sampled at Von
512 Damm compared to Axial Seamount, so it is unclear whether this results from a true biological
513 signal or insufficient data.
514
515 *Vents host diverse and active assemblages of viruses with restricted host range*
516 The viral infection networks show diverse microbial lineages across several samples infected
517 by many different viruses at the time of sampling. Although individual microbial lineages could be
518 targeted by multiple viral strains, the viruses we identified in these hydrothermal habitats had fairly
519 restricted host ranges, infecting specific individual microbial strains. The only clear example of viral
520 infection across microbial taxa emerged at Axial Seamount, where viral sequences associated with
521 Ignavibacteria were also linked to Aquificales (Figure 4A). All other examples of viral infection
522 across microbial taxa may either be true examples of viral-host cross-infection or may instead result
523 from shared direct repeats. The limited host range of viruses from hydrothermal systems is
524 consistent with previous work indicating that viruses generally tend to be host-specific (82), and we
525 found very little evidence for generalist viruses, despite the fact that this has been reported
526 previously (82, 84). Furthermore, we identified many viral connections to Epsilonbacteraeota,
527 confirming that these highly successful microbial groups in hydrothermal vent diffuse flow fluids are
528 susceptible to viral infection (3, 39, 85–87).

529 Matches between CRISPR loci and viruses provide an indication of which viruses are being
530 targeted by the CRISPR immune response. We found that while some CRISPR spacers target
531 relatively abundant viruses, most CRISPRs target relatively rare viruses in these systems. This is
532 consistent with previous observations in an archaea-dominated hypersaline lake (88), where the vast
533 majority of CRISPRs were found to target viruses with populations too small to allow for the
534 assembly of contigs. The relative scarcity of viruses targeted by CRISPRs may result from an
535 evolutionary arms race: CRISPRs limit the abundance of the viral populations they target, while
536 concurrently, viruses undergo mutations, limiting the ability of CRISPR spacers to target them.
537 Alternatively, these observations could result from an abundance of inactive spacers inherited over
538 multiple generations. However, we do not believe this to be the case because CRISPR spacer
539 clusters were infrequently present across multiple samples, suggesting that spacers were integrated
540 on subgenerational timescales.
541
542 *Patterns recorded in microbial CRISPR loci do not reflect the contemporary viral assemblage*
543 In contrast to the viral clusters, the CRISPR spacers did not demonstrate any clear
544 biogeographic patterns. We expected CRISPR spacers to be more widespread than viruses since viral
545 composition represents the virus community at the time of sampling, while spacers represent a
546 history of viral infection. In contrast to our predictions, while both had limited distributions, we
547 found viruses to be more widespread than CRISPR spacers at all scales examined (Table 1). These
548 results are in contrast to previous studies of CRISPR spacer biogeography in terrestrial hot springs,
549 where both viral and CRISPR spacers showed clear biogeographic structure (89). This suggests that
550 there is selective pressure for CRISPR spacer composition to evolve more rapidly than viral
551 sequences via the loss or mutation of CRISPR spacers. However, it is also possible that we found
552 viruses to be more widespread because of undersampling of CRISPR spacers, or our use of different
553 clustering algorithms for viruses and spacers: spacers may have been clustered at a finer resolution,
554 resulting in a more narrow distribution for each spacer.
555 Given that CRISPR spacers provide a history of infection, comparing the record of past viral
556 infections via CRISPR arrays with viral sequences in the metagenomes can provide insights into
557 whether CRISPR arrays provide an accurate representation of the viral assemblage at the time of
558 sampling, as well as the rate at which CRISPR spacers are accumulated. We did not find that the
559 most common or cosmopolitan viruses and CRISPR spacers matched each other. This may arise
560 from a temporal mismatch: it takes time for spacers to be incorporated into and lost from CRISPR

561 loci as virus abundances shift and viruses evolve. Additionally, just one SNP (which we allowed for
562 when aligning spacers to viruses) can prevent a CRISPR spacer from providing resistance against a
563 virus (66). This could explain the discrepancy between viruses and spacers: once resistance of a
564 spacer to a particular virus is suppressed, the population of the virus is freed to shift independently
565 from the spacer in the host population. The discrepancy we observed between viruses and spacers is
566 important to note when attempting to use CRISPR spacers to study viral populations or vice versa.
567 Examination of the prophage abundance in MAGs revealed a relatively high abundance of
568 prophage encoded in each MAG, confirming previous results indicating that lysogeny is a common
569 lifestyle in hydrothermal systems (14, 15). However, while examination of the gene content of viral
570 contigs revealed several ORFs encoding a range of functions including cell membrane function and
571 energy production, it was difficult to prove conclusively that these genes were viral-encoded AMGs
572 rather than potential cellular contamination, and thus no concrete conclusions were provided here
573 (see Supplementary Materials). Previous research has suggested that vent viruses encode AMGs (18,
574 19) and thus it is likely that many of these genes represent viral-encoded AMGs, but further research
575 is necessary to determine AMG diversity and prevalence across vent systems.
576 Taken together, our results show that hydrothermal vent viruses are active, abundant, and
577 diverse. These viruses are restricted in their host range and biogeographic extent, but their
578 interactions with hosts persist over time. Thus, while viruses in venting fluids from deep-sea
579 hydrothermal systems have the capacity to play an important role in driving the evolution and
580 ecology of microbial communities, their influence appears to be highly localized to specific regions
581 and taxa. Future work examining viral diversity and distribution across higher resolution transects
582 over space and time should reveal further insights into the extent of the viral influence in deep-sea
583 hydrothermal vents.
584
585 **Acknowledgements**
586 We thank Julie Reveillaud, Caroline Fortunato, and Emily Reddington for support in the
587 collection and generation of metagenomic data, and Chip Breier, David Butterfield, Bill Chadwick,
588 Chris German, Jim Holden, Jill McDermott, and Jeff Seewald for sample collection support at sea.
589 Funding for ET was provided by Carleton College. REA was supported by a NASA Postdoctoral
590 Fellowship with the NASA Astrobiology Institute. This work was supported by a NASA Exobiology
591 grant number 80NSSC18K1076 to REA and JAH; a NASA Astrobiology Science and Technology
592 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. 1985. Geomicrobiology of Deep-Sea Hydrothermal Vents. Science.

2. 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.

3. Meier DV, Pjevac P, Bach W, Hourdez S, Girguis PR, Vidoudez C, Amann R, Meyerdierks A. 2017. Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. ISME J 11:1545–1558.

4. Fortunato CS, Larson B, Butterfield DA, Huber JA. 2018. Spatially distinct, temporally stable microbial populations mediate biogeochemical cycling at and below the seafloor in hydrothermal vent fluids. Environ Microbiol 20:769–784.

5. McNichol J, Stryhanyuk H, Sylva SP, Thomas F, Musat N, Seewald JS, Sievert SM. 2018. Primary productivity below the seafloor at deep-sea hot springs. Proc Natl Acad Sci U S A 115:6756–6761.

6. Trembath-Reichert E, Butterfield DA, Huber JA. 2019. Active subseafloor microbial communities from Mariana back-arc venting fluids share metabolic strategies across different thermal niches and taxa. The ISME Journal.

7. Galambos D, Anderson RE, Reveillaud J, Huber JA. 2019. Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at

622 deep-sea hydrothermal vents. Environ Microbiol 21:4395–4410.

623 8. Suttle CA. 2005. Viruses in the sea. Nature 437:356–361.

624 9. Suttle CA. 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol 5:801–812.

625 10. Chow C-ET, Suttle CA. 2015. Biogeography of Viruses in the Sea. Annu Rev Virol 2:41–66.

626 11. Weinbauer MG, Rassoulzadegan F. 2004. Are viruses driving microbial diversification and diversity?
627 Environ Microbiol 6:1–11.

628 12. Ortmann AC, Suttle CA. 2005. High abundances of viruses in a deep-sea hydrothermal vent system
629 indicates viral mediated microbial mortality. Deep Sea Research Part I: Oceanographic Research Papers.

630 13. Anderson RE, Brazelton WJ, Baross JA. 2011. Using CRISPRs as a metagenomic tool to identify
631 microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol 77:120–133.

632 14. Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, Winget D, Wommack KE. 2008.
633 Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J
634 2:1112–1121.

635 15. Labonté JM, Pachiadaki M, Fergusson E, McNichol J, Grosche A, Gulmann LK, Vetriani C, Sievert SM,
636 Stepanauskas R. 2019. Single Cell Genomics-Based Analysis of Gene Content and Expression of
637 Prophages in a Diffuse-Flow Deep-Sea Hydrothermal System. Frontiers in Microbiology.

638 16. Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the marine microbial
639 realm. Nat Microbiol 3:754–766.

640 17. He T, Li H, Zhang X. 2017. Deep-Sea Hydrothermal Vent Viruses Compensate for Microbial
641 Metabolism in Virus-Host Interactions. MBio 8.

642 18. Anderson RE, Sogin ML, Baross JA. 2014. Evolutionary strategies of viruses, bacteria and archaea in
643 hydrothermal vent ecosystems revealed through metagenomics. PLoS One 9:e109696.

644 19. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. 2014. Sulfur oxidation
645 genes in diverse deep-sea viruses. Science 344:757–760.

646 20. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and
647 classification. Annu Rev Microbiol 55:709–742.

648 21. Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U
649 S A 102:14332–14337.

650 22. Fuchsman CA, Collins RE, Rocap G, Brazelton WJ. 2017. Effect of the environment on horizontal gene
651 transfer between bacteria and archaea. PeerJ 5:e3865.

652 23. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial
653 innovation. Nature 405:299–304.

654 24. Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open
655 source evolution. Nat Rev Microbiol 3:722–732.

656 25. Brazelton WJ, Baross JA. 2009. Abundant transposases encoded by the metagenome of a hydrothermal
657 chimney biofilm. ISME J 3:1420–1424.

658 26. Stewart LC, Algar CK, Fortunato CS, Larson BI, Vallino JJ, Huber JA, Butterfield DA, Holden JF. 2019.
659 Fluid geochemistry, local hydrology, and metabolic activity define methanogen community size and
660 composition in deep-sea hydrothermal vents. ISME J 13:1711–1721.

661 27. Opatkiewicz AD, Butterfield DA, Baross JA. 2009. Individual hydrothermal vents at Axial Seamount
662 harbor distinct subseafloor microbial communities. FEMS Microbiol Ecol 70:413–424.

663 28. Anderson RE, Sogin ML, Baross JA. 2015. Biogeography and ecology of the rare and abundant
664 microbial lineages in deep-sea hydrothermal vents. FEMS Microbiol Ecol 91:1–11.

665 29. Han Y, Perner M. 2015. The globally widespread genus Sulfurimonas: versatile energy metabolisms and
666 adaptations to redox clines. Front Microbiol 6:989.

667 30. Dick GJ. 2019. The microbiomes of deep-sea hydrothermal vents: distributed globally, shaped locally.
668 Nat Rev Microbiol 17:271–283.

669 31. Damm KV, Von Damm K. 1990. Seafloor Hydrothermal Activity: Black Smoker Chemistry And
670 Chimneys. Annual Review of Earth and Planetary Sciences.

671 32. Kelley DS. 2005. A Serpentinite-Hosted Ecosystem: The Lost City Hydrothermal Field. Science.

672 33. Anderson RE, Reveillaud J, Reddington E, Delmont TO, Eren AM, McDermott JM, Seewald JS, Huber
673 JA. 2017. Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea
674 hydrothermal vents. Nat Commun 8:1114.

34. Butterfield DA, Roe KK, Lilley MD, Huber JA, Baross JA, Embley RW, Massoth GJ. 2004. 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.

35. McDermott JM, Seewald JS, German CR, Sylva SP. 2015. Pathways for abiotic organic synthesis at submarine hydrothermal fields. Proc Natl Acad Sci U S A 112:7668–7672.

36. German CR, Bowen A, Coleman ML, Honig DL, Huber JA, Jakuba MV, Kinsey JC, Kurz MD, Leroy S, McDermott JM, de Lépinay BM, Nakamura K, Seewald JS, Smith JL, Sylva SP, Van Dover CL, Whitcomb LL, Yoerger DR. 2010. Diverse styles of submarine venting on the ultraslow spreading Mid-Cayman Rise. Proc Natl Acad Sci U S A 107:14020–14025.

37. McDermott JM. 2015. Geochemistry of deep-sea hydrothermal vent fluids from the Mid-Cayman Rise, Caribbean Sea.

38. McDermott JM, Sylva SP, Ono S, German CR, Seewald JS. 2018. Geochemistry of fluids from Earth's deepest ridge-crest hot-springs: Piccard hydrothermal field, Mid-Cayman Rise. Geochimica et Cosmochimica Acta.

39. Moulana A, Anderson RE, Fortunato CS, Huber JA. 2020. 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 5.

40. Reveillaud J, Reddington E, McDermott J, Algar C, Meyer JL, Sylva S, Seewald J, German CR, Huber JA. 2016. Subseafloor microbial communities in hydrogen-rich vent fluids from hydrothermal systems along the Mid-Cayman Rise. Environ Microbiol 18:1970–1987.

41. Breier JA, Gomez-Ibanez D, Reddington E, Huber JA, Emerson D. 2012. A precision multi-sampler for deep-sea hydrothermal microbial mat studies. Deep Sea Research Part I: Oceanographic Research Papers.

42. Breier JA, Rauch CG, McCartney K, Toner BM, Fakra SC, White SN, German CR. 2009. A suspended-particle rosette multi-sampler for discrete biogeochemical sampling in low-particle-density waters. Deep Sea Research Part I: Oceanographic Research Papers.

43. Crump BC, Hopkinson CS, Sogin ML, Hobbie JE. 2004. Microbial biogeography along an estuarine salinity gradient: combined influences of bacterial growth and residence time. Appl Environ Microbiol 70:1494–1505.

704 44. Zhou J, Bruns MA, Tiedje JM. 1996. DNA recovery from soils of diverse composition. Appl Environ
705 Microbiol 62:316–322.

706 45. Fortunato CS, Huber JA. 2016. Coupled RNA-SIP and metatranscriptomics of active
707 chemolithoautotrophic communities at a deep-sea hydrothermal vent. ISME J 10:1925–1938.

708 46. Eren AM, Vineis JH, Morrison HG, Sogin ML. 2013. A filtering method to generate high quality short
709 reads using illumina paired-end technology. PLoS One 8:e66643.

710 47. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and
711 metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428.

712 48. Skennerton CT, Imelfort M, Tyson GW. 2013. Crass: identification and reconstruction of CRISPR from
713 unassembled metagenomic data. Nucleic Acids Res 41:e105.

714 49. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for
715 identifying viral sequences from assembled metagenomic data. Microbiome 5:69.

716 50. van Dongen SM. 2000. Graph Clustering by Flow Simulation.

717 51. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an
718 advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

719 52. Darling AE, Jospin G, Lowe E, Matsen FA 4th, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic
720 analysis of genomes and metagenomes. PeerJ 2:e243.

721 53. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson
722 AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods
723 11:1144–1146.

724 54. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a
725 comprehensive online resource for quality checked and aligned ribosomal RNA sequence data
726 compatible with ARB. Nucleic Acids Res 35:7188–7196.

727 55. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA
728 ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids
729 Res 41:D590–6.

730 56. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

731 57. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P,
732 Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench
733 compatible with ARB. Appl Environ Microbiol 72:5069–5072.

734 58. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB,
735 Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing
736 mothur: open-source, platform-independent, community-supported software for describing and
737 comparing microbial communities. Appl Environ Microbiol 75:7537–7541.

738 59. Dixon P. 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation
739 Science.

740 60. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, Augspurger T,
741 Halchenko Y, Cole JB, Warmenhoven J, de Ruiter J, Pye C, Hoyer S, Vanderplas J, Villalba S, Kunter G,
742 Quintero E, Bachant P, Martin M, Meyer K, Miles A, Ram Y, Yarkoni T, Williams ML, Evans C,
743 Fitzgerald C, Brian, Fonnesbeck C, Lee A, Qalieh A. 2017. mwaskom/seaborn: v0.8.1. Zenodo
744 https://doi.org/10.5281/zenodo.883859.

745 61. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas
746 AE, Boyd JA, Hodgkins SB, Wilson RM, Trubl G, Li C, Frolking S, Pope PB, Wrighton KC, Crill PM,
747 Chanton JP, Saleska SR, Tyson GW, Rich VI, Sullivan MB. 2018. Host-linked soil viral ecology along a
748 permafrost thaw gradient. Nat Microbiol 3:870–880.

749 62. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome Datasets Are
750 Compositional: And This Is Not Optional. Front Microbiol 8:2224.

751 63. Gloor GB, Reid G. 2016. Compositional analysis: a valid approach to analyze microbiome high-
752 throughput sequencing data. Can J Microbiol 62:692–703.

753 64. Palarea-Albaladejo J, Martín-Fernández JA. 2015. zCompositions — R package for multivariate
754 imputation of left-censored data under a compositional approach. Chemometrics and Intelligent
755 Laboratory Systems.

756 65. Daly RA, Roux S, Borton MA, Morgan DM, Johnston MD, Booker AE, Hoyt DW, Meulia T, Wolfe
757 RA, Hanson AJ, Mouser PJ, Moore JD, Wunch K, Sullivan MB, Wrighton KC, Wilkins MJ. 2019.
758 Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing.
759 Nat Microbiol 4:352–361.

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

762 67. Garrett RA, Vestergaard G, Shah SA. 2011. Archaeal CRISPR-based immune systems: exchangeable
763 functional modules. Trends Microbiol 19:549–556.

764 68. Manica A, Zebec Z, Teichmann D, Schleper C. 2011. In vivo activity of CRISPR-mediated virus defence
765 in a hyperthermophilic archaeon. Mol Microbiol 80:481–491.

766 69. Team R, Others. 2015. RStudio: integrated development for R. RStudio, Inc , Boston, MA URL
767 http://www rstudio com 42:14.

768 70. Wickham H, Hester J, Francois R. 2017. Readr: Read rectangular text data. R package version 1.

769 71. Wickham H, Bryan J. 2017. readxl: Read Excel files. R package version 1.

770 72. Wickham H, Henry L, Others. 2018. tidyr: Easily Tidy Data with "spread ()"and "gather ()"Functions. R
771 package version 0 8 2.

772 73. Wickham H. 2017. stringr: Simple, consistent wrappers for common string operations. R package
773 version 1.

774 74. Wickham H, Francois R, Henry L, Müller K, Others. 2015. dplyr: A grammar of data manipulation. R
775 package version 0 4 3.

776 75. Briatte F. 2016. ggnetwork: Geometries to Plot Networks with'ggplot2'. R package version 0 5 1.

777 76. Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M. 2008. statnet: Software tools for the
778 representation, visualization, analysis and simulation of network data. J Stat Softw 24:1548.

779 77. Kassambara A. 2017. ggpubr:"ggplot2" based publication ready plots. R package version 0 1 6.

780 78. Wickham H. 2011. ggplot2. WIREs Comp Stat 3:180–185.

781 79. Luciani TJ, Decorde M, Lise V. 2009. RSvgDevice: An R SVG graphics device. R package version 0 6 4.

782 80. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de
783 Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant
784 S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans
785 Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Ocean

786 plankton. Patterns and ecological drivers of ocean viral communities. Science 348:1261498.

787 81. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E,
788 Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte
789 CM, Gasol JM, Vaqué D, Tara Oceans Coordinators, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016.
790 Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature
791 537:689–693.

792 82. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin
793 E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. Nature 536:425–430.

794 83. Flores GE, Campbell JH, Kirshtein JD, Meneghin J, Podar M, Steinberg JI, Seewald JS, Tivey MK,
795 Voytek MA, Yang ZK, Reysenbach A-L. 2011. Microbial community structure of hydrothermal deposits
796 from geochemically different vent fields along the Mid-Atlantic Ridge. Environ Microbiol 13:2158–2171.

797 84. Peters DL, Lynch KH, Stothard P, Dennis JJ. 2015. The isolation and characterization of two
798 Stenotrophomonas maltophilia bacteriophages capable of cross-taxonomic order infectivity. BMC
799 Genomics 16:664.

800 85. Akerman NH, Butterfield DA, Huber JA. 2013. Phylogenetic diversity and functional gene patterns of
801 sulfur-oxidizing subseafloor Epsilonproteobacteria in diffuse hydrothermal vent fluids. Front Microbiol
802 4:185.

803 86. Meyer JL, Huber JA. 2014. Strain-level genomic variation in natural populations of Lebetimonas from an
804 erupting deep-sea volcano. ISME J 8:867–880.

805 87. Yoshida-Takashima Y, Takaki Y, Shimamura S, Nunoura T, Takai K. 2013. Genome sequence of a
806 novel deep-sea vent epsilonproteobacterial phage provides new insight into the co-evolution of
807 Epsilonproteobacteria and their phages. Extremophiles 17:405–419.

808 88. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF. 2013. Virus-
809 host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea
810 2013:370871.

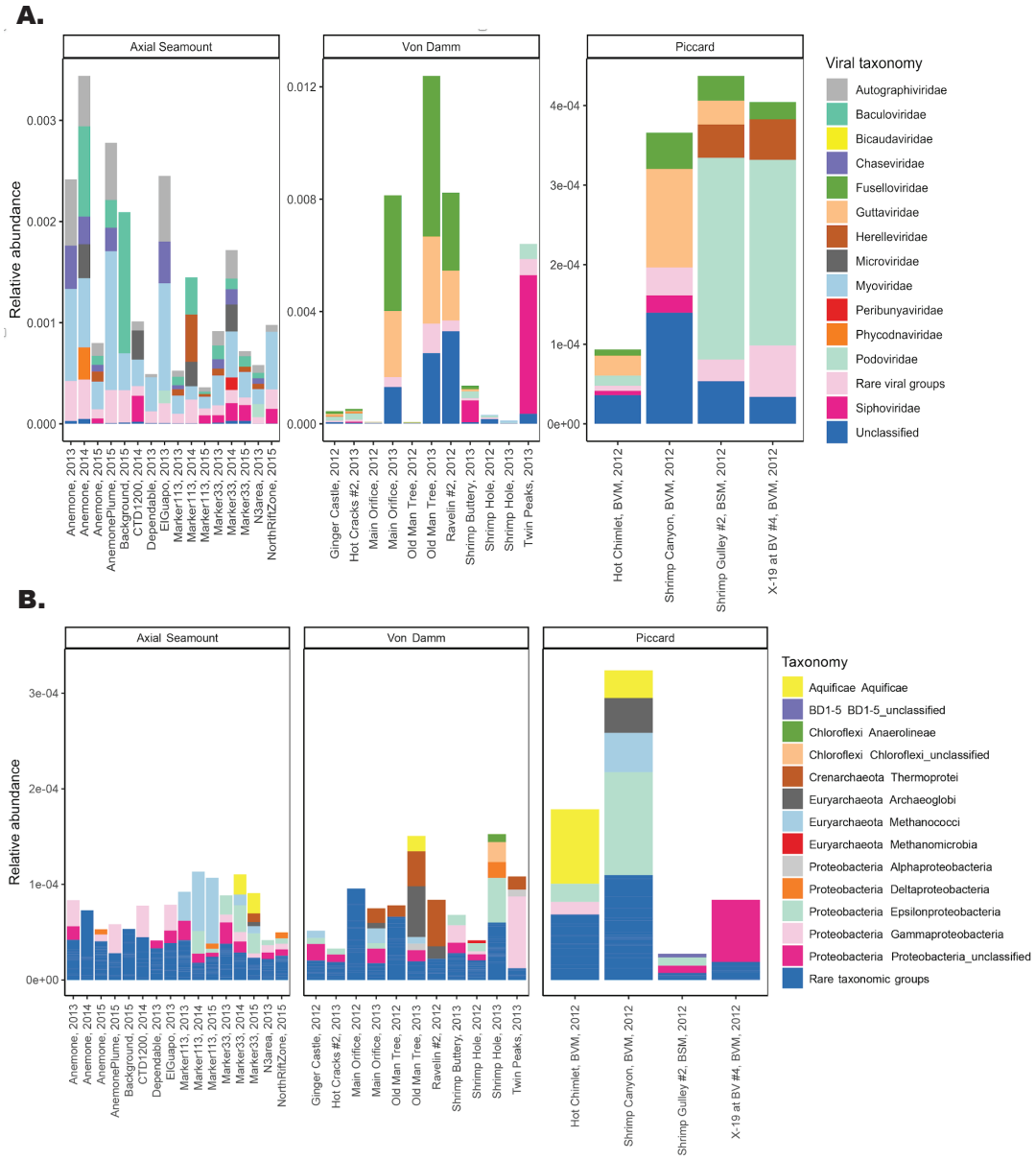
811 89. Held NL, Whitaker RJ. 2009. Viral biogeography revealed by signatures in Sulfolobus islandicus
812 genomes. Environ Microbiol 11:457–466.

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 at Piccard, Von Damm, and Axial Seamount. The Background and CTD1200 samples were excluded. Reads that mapped to 16S rRNA sequences that were classified as eukaryotes were excluded.

Vent field	Type	Percent in more than one vent site	Percent in all vent sites in region
Von Damm	Viral clusters	13.71	0
Von Damm	Spacer clusters	5.1	0
Von Damm	16S rRNA host	46.15	1.44
Piccard	Viral clusters	10.78	0
Piccard	Spacer clusters	9.13	0
Piccard	16S rRNA host	46.09	11.72
Axial Seamount	Viral clusters	8.88	0.05
Axial Seamount	Spacer clusters	1.01	0
Axial Seamount	16S rRNA host	54.55	2.36
Type	**Percent found in all three regions (Axial Seamount, Von Damm, and Piccard)**		
Viral clusters	0.02		
Spacer clusters	0		
16S rRNA host	16.62		

824 **Figures**

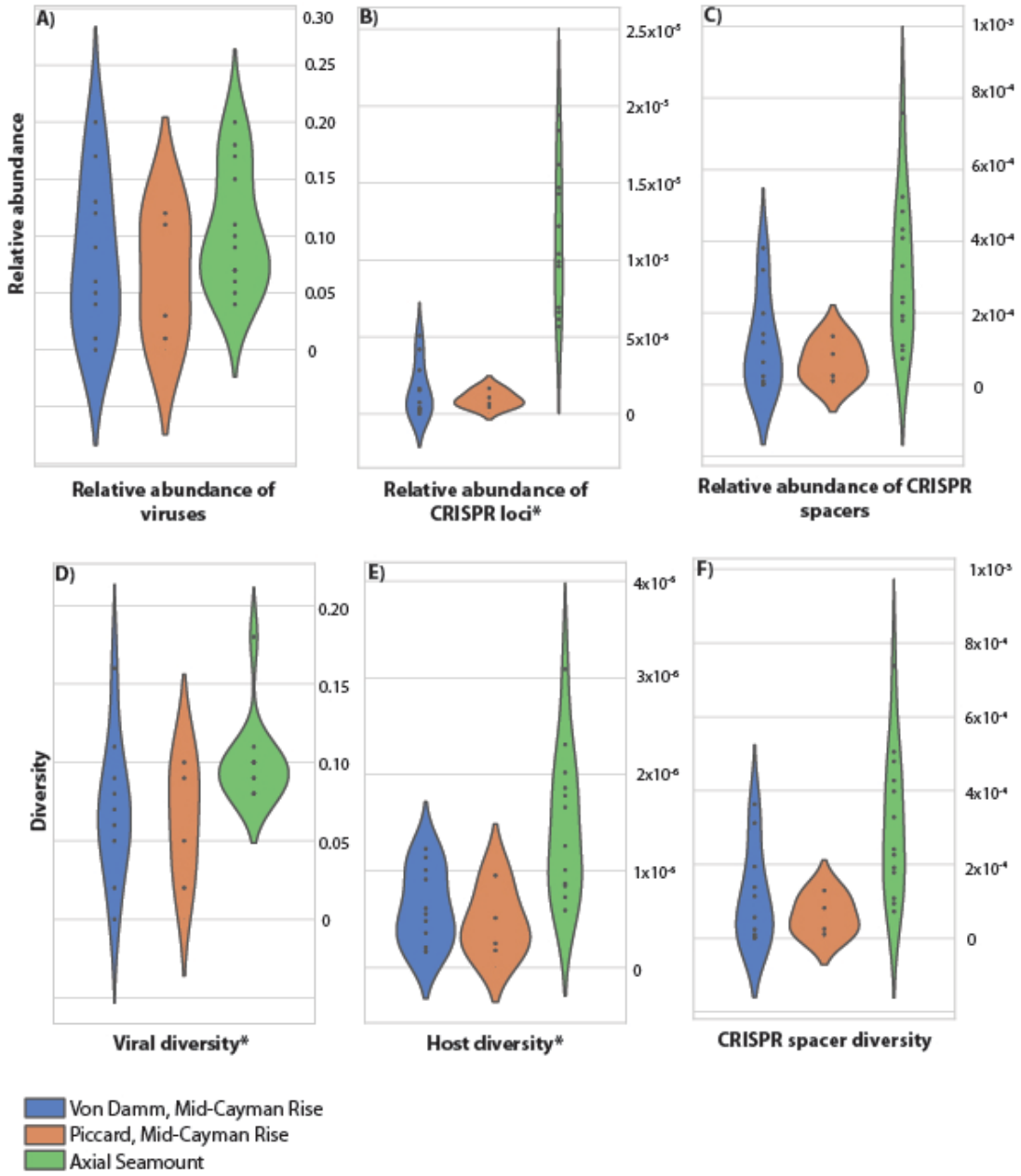


825

826

827 **Figure 1. Taxonomy of A) microbial and B) viral reads** in Von Damm vent field, Piccard vent field, and
828 Axial Seamount. In A) the y-axis represents the number of reads that mapped to 16S rRNA sequences of
829 each host normalized by the number of reads in each sample. Reads that mapped to 16S rRNA sequences
830 that were unclassified or classified as eukaryotes were excluded. Rare taxonomic groups represent taxonomies
831 that comprised less than 1% of reads that mapped to 16S rRNA sequences in a given sample. In B), the y-axis
832 represents the number of reads that mapped to VirFinder contigs of each viral taxonomy normalized by the

833 number of reads in each sample. Rare viral groups represent viral taxonomies that comprised less than 5% of
834 reads that mapped to viral contigs in a given sample.
835
836
837
838



839
840 **Figure 2. Abundance and diversity of viral sequences, CRISPR spacers, CRISPR loci, and microbes**
841 **in diffuse flow samples from the Von Damm vent field (blue), Piccard vent field (orange), and Axial**

842 **Seamount (green).** Relative abundance reflects the relative number of reads mapping to viral contigs within
843 each region, normalized by read abundance (see Methods). Diversity reflects the number of clusters per read,
844 normalized by read abundance (see Methods). Note that the y-axis values are different for each variable.
845 Values for individual samples are indicated with black dots. Violins represent the kernel density estimation of
846 the underlying data distribution. Variables with significant differences between Von Damm, Piccard, and
847 Axial Seamount are indicated with asterisks. For these comparisons, samples from background seawater and
848 plumes were excluded.
849
850
851

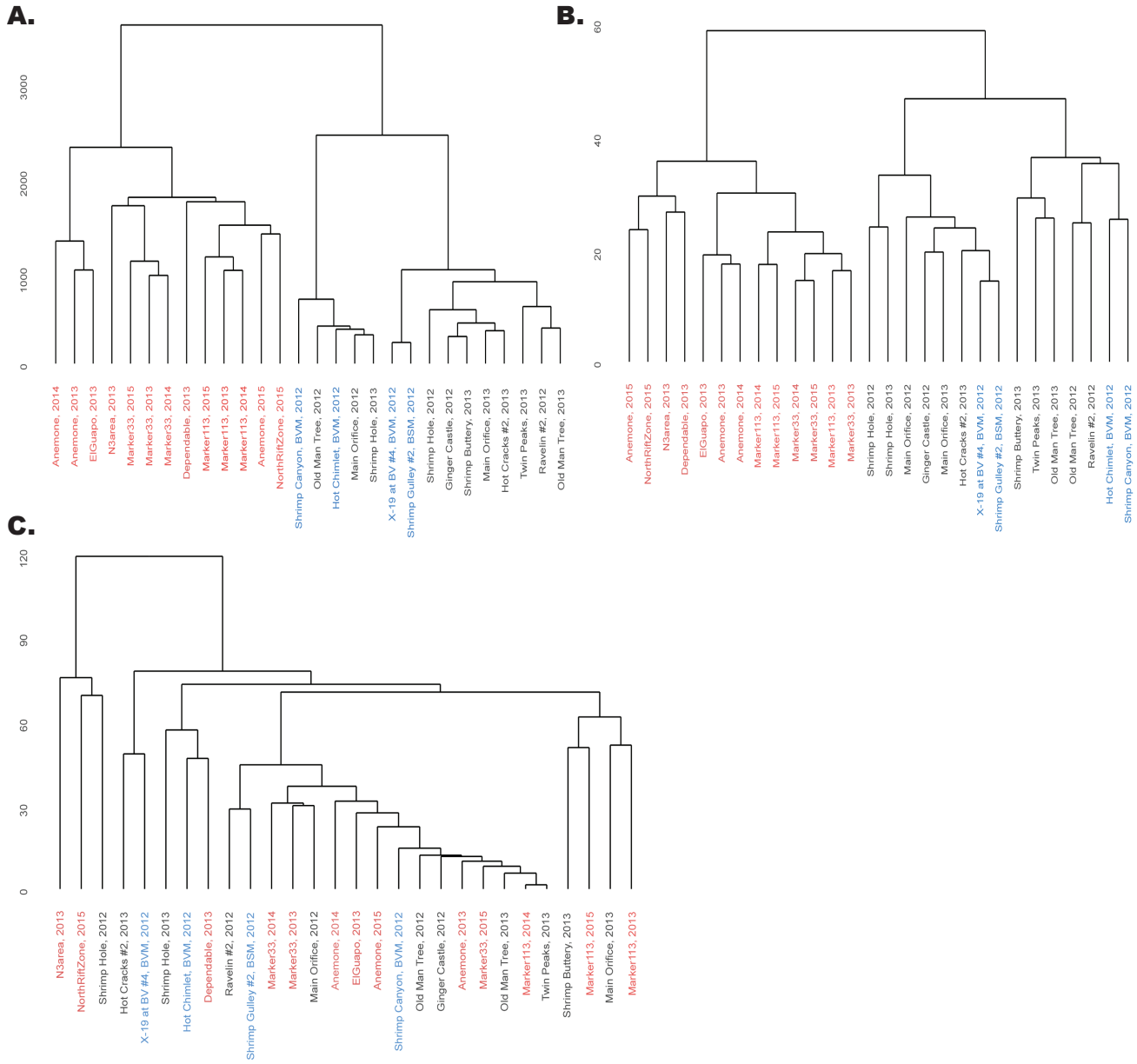


Figure 3. Hierarchical clustering of Von Damm (black) Piccard (blue) and Axial Seamount (red) diffuse flow samples based on A) viral assemblage, B) microbial host composition, and C) CRISPR spacer composition. For viral composition, the number of reads in each sample that mapped to each viral cluster was calculated. 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. Reads that mapped to 16S rRNA sequences that were either unclassified or classified as eukaryotes were excluded. For CRISPR composition, the number of CRISPR spacers in each sample in each spacer cluster was calculated. For all samples, the zero counts were replaced with estimates and a centered log-ratio (clr) transformation was

applied. The y-axes indicate distance between samples as calculated by the ward.D2 method based on the transformed counts.

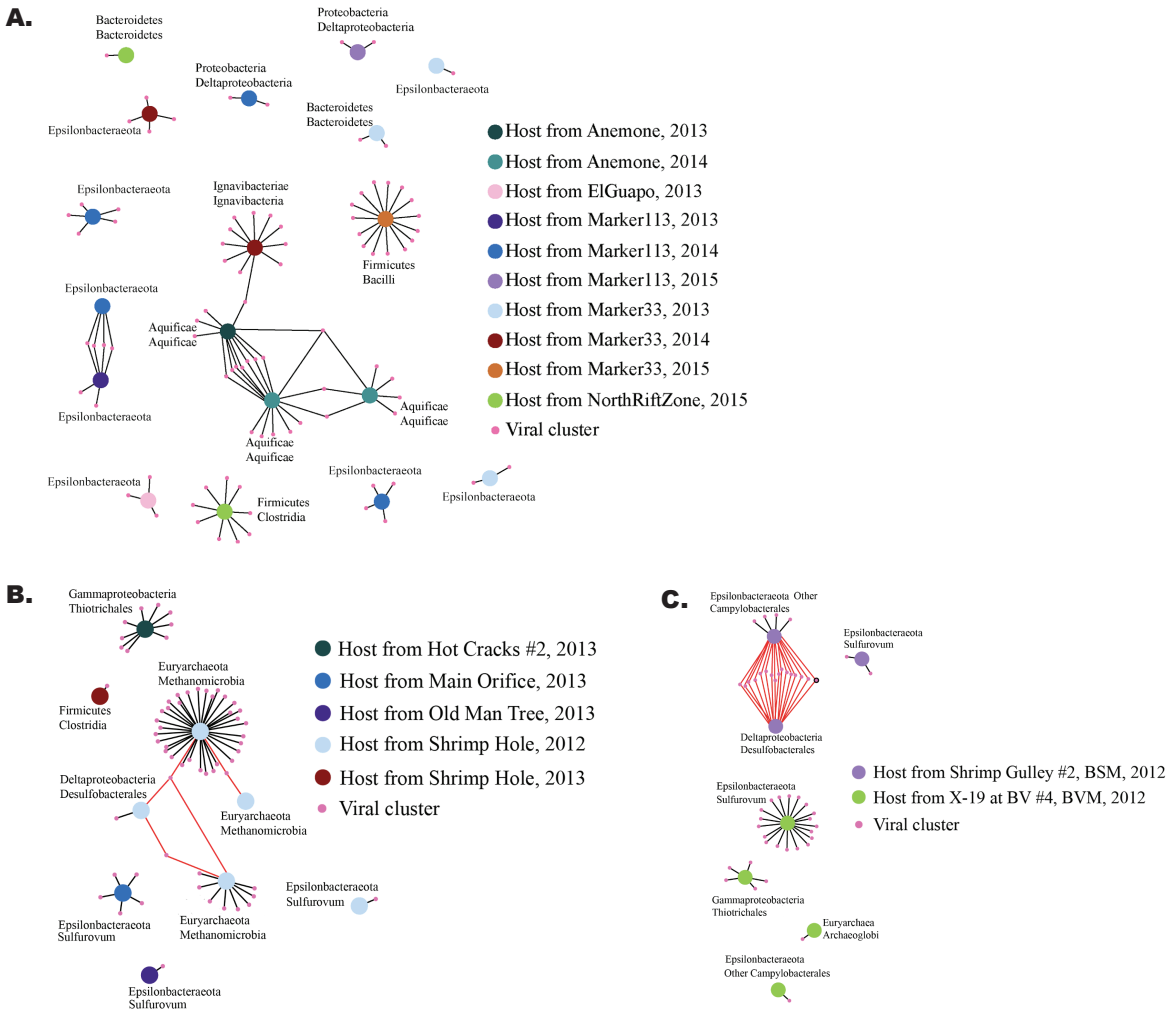


Figure 4. Infection network showing the links between viral clusters and MAGs at A) Axial Seamount, B) Von Damm and C) Piccard. Viral clusters were linked to MAGs via the spacers and direct repeats in CRISPR loci. Edges are colored red when a viral cluster is linked to multiple MAGs in the same sample 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

875 with direct repeat type came from. Therefore, the red edges are links between viral clusters and MAGs that
876 could not be definitely proven. Viral clusters with high relative abundance (top six most abundant) in at least
877 one sample are outlined in black.
878

879 **SUPPLEMENTARY MATERIAL**

880 *Viruses and prophages in MAGS and viral and prophage genes*

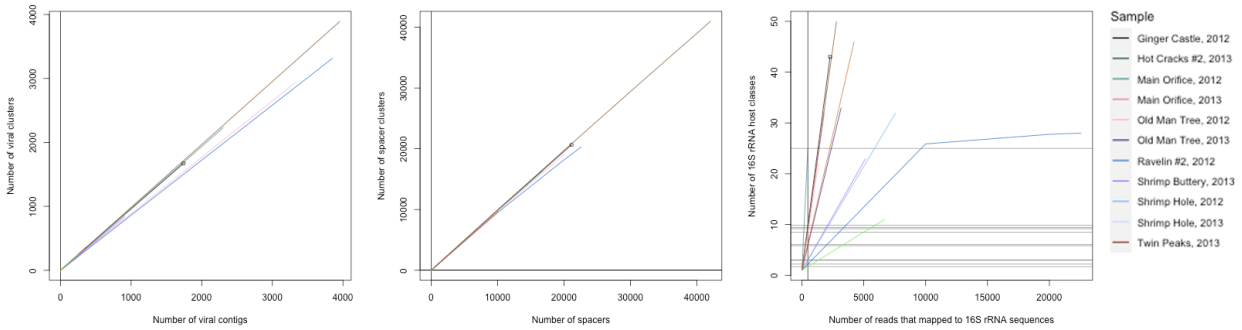
881 In order to identify viral genes and putative auxiliary metabolic genes (AMGs), VirSorter (69)
882 v1.0.3 was used to extract putative viral and prophage sequences from metagenomic assemblies and
883 from individual MAGs. In the case of identifying putative auxiliary metabolic genes, we chose to
884 maximize accuracy rather than recovery. Therefore, we used VirSorter for prophage and AMG
885 identification because VirSorter relies on known viral hallmark genes for viral identification and thus
886 is a more conservative tool for positive viral identification. Category one, category two, category
887 four (prophage), and category five (prophage) VirSorter sequences were used in analyses. Putative
888 viral and prophage sequences from all assembled contigs were annotated via Prokka (69, 70) v1.14
889 with the taxonomic identifier set to "bacteria" as well as to "viruses." We assigned categories to the
890 ORFs that had been identified with the "bacteria" taxonomic identifier using the Clusters of
891 Orthologous Groups of proteins (COG) database (71, 72). We conducted a second round of
892 annotation for ORFs identified by Prokka using the "viruses" taxonomic identifier using VirSorter;
893 we classified these ORFs as having viral function if they were annotated with terms that included
894 "capsid," "tail," "spike," "terminase large subunit," "portal," and "coat."
895
896 Of the 74 Mid-Cayman Rise MAGs, 28 (38%) had at least one putative prophage sequence
897 counting VirSorter categories one, two, four, or five, and 6 (8%) had a putative prophage counting
898 only VirSorter categories four and five (Supplementary Table 4). Of the 98 Axial Seamount MAGs,
899 41 (42%) had at least one putative viral or prophage sequence counting VirSorter categories one,
900 two, four, or five, and 6 (6%) had a putative prophage sequence counting VirSorter categories four
901 and five (Supplementary Table 4). Given the uncertainties associated with binning viral contigs into
902 MAGs, the true number of prophage in hydrothermal vent microbial genomes is likely between
903 these endmember values. VirSorter has not been verified in hydrothermal systems; however, it is an
904 excellent tool that has proved useful results and is commonly used in other non-surface ocean
905 systems (61, 84, 85).
906
907 To determine whether viruses and prophage in these samples carried auxiliary metabolic genes
908 (AMGs), we annotated and characterized viral and prophage gene content in the viral sequences
909 recovered from the metagenomes. The vast majority of genes within the putative viral and prophage
910 sequences were annotated as belonging to the COG categories replication and repair; nucleotide
911 metabolism and transport; and post-translational modification, protein turnover, chaperone
912 functions. Many genes were also annotated as belonging to the cell wall/membrane/envelope
913 biogenesis category. Within the Mid-Cayman Rise, a total of 106 ORFs within putative viral or
914 prophage sequences identified in 10 different metagenomes fell within the broad COG category of
915 "metabolism" (Supplementary Fig. 4). This included genes categorized as energy production and
916 conversion, amino acid transport and metabolism, nucleotide transport and metabolism,
917 carbohydrate transport and metabolism, coenzyme transport and metabolism, lipid transport and
918 metabolism, and inorganic ion transport and metabolism. Samples taken from X-19 and Shrimp
919 Hole (2012) had a high number of metabolism genes (30 and 28, respectively) on putative viral or
920 prophage sequences compared to the other Mid-Cayman Rise samples. A higher proportion of
921 genes on putative viral or prophage sequences were related to metabolism at the Mid-Cayman Rise
922 than Axial Seamount (Supplementary Fig. 4). At Axial Seamount, we identified a total of 185 ORFs
923 on putative viral or prophage sequences from 14 different metagenomes that fell within the broad
924 COG category of "metabolism." However, it is important to note that none of the observed
925 metabolism genes were surrounded by confirmed virus genes (i.e., known viral ORFs both upstream

926 and downstream of the metabolic ORF), and therefore we cannot rule out the possibility that some
927 of these genes were microbial in origin.
928
929 *Auxiliary metabolic genes and the role of virus-driven horizontal gene transfer in these systems*
930 Previous work has suggested that viruses in hydrothermal systems have the genomic capacity
931 to alter their hosts' microbial metabolism through harboring auxiliary metabolic genes (AMGs) (17–
932 19). We found a higher percentage of microbial genomes with putative viral sequences (40%) than
933 has been previously reported for single-cell genomes (SAGs, 10%) in diffuse flow hydrothermal
934 fluids (15). The high abundance of prophage identified here is consistent with previous work
935 documenting a high incidence of prophage in diffuse flow systems compared to deep seawater (14).
936 The ORFs on the viral contigs identified in these hydrothermal vent metagenomes encoded
937 a wide range of functions, some of which may function as auxiliary metabolic genes. Many viral
938 contigs contained genes related to cell wall or membrane proteins. The function of these genes is
939 unknown; it is possible that these genes are involved in the synthesis of membranes surrounding
940 viral capsids. Genes related to outer membrane proteins and protein glycosylation are commonly
941 observed as variable genes within microbial pangenomes (94–100), possibly as a means to vary
942 membrane proteins to evade viral infection. Viruses often act as a source of genes for horizontal
943 gene transfer via transduction, and it is possible that the introduction of these novel genes may
944 enable some microbial strains to avoid infection by other viruses. We also observed several genes
945 with functions related to energy metabolism, inorganic ion transport, and signal transduction.
946 Previous work has found that genes related to inorganic ion transport are differentially distributed in
947 the *Sulfurovum* pangenome according to nutrient availability (39) and further work on viral sequences
948 across vent sites is needed to determine whether these genes are carried by viruses to benefit their
949 hosts. Moreover, work on viruses in hydrothermal plumes (19) and in diffuse flow vent fluid (18)
950 has indicated that viruses in these systems encode energy-metabolizing AMGs, potentially to
951 supplement the host's ability to generate sufficient energy for cellular processes during the course of
952 infection. The abundance and diversity of metabolic genes observed on these contigs provide
953 further evidence that viral-encoded AMGs are widespread and diverse within hydrothermal vent
954 ecosystems. These genes also have the potential to be horizontally transferred via viral transduction.
955 However, our analyses of viral biogeography in hydrothermal vents has revealed virus-host
956 interactions to have limited distributions because viruses are spatially restricted and highly host-
957 specific. Thus, the potential for viruses to act as mechanisms for horizontal gene transfer between
958 spatially and phylogenetically distant hosts is likely limited.
959
960
961
962
963
964
965
966
967
968

969 **SUPPLEMENTARY FIGURES**

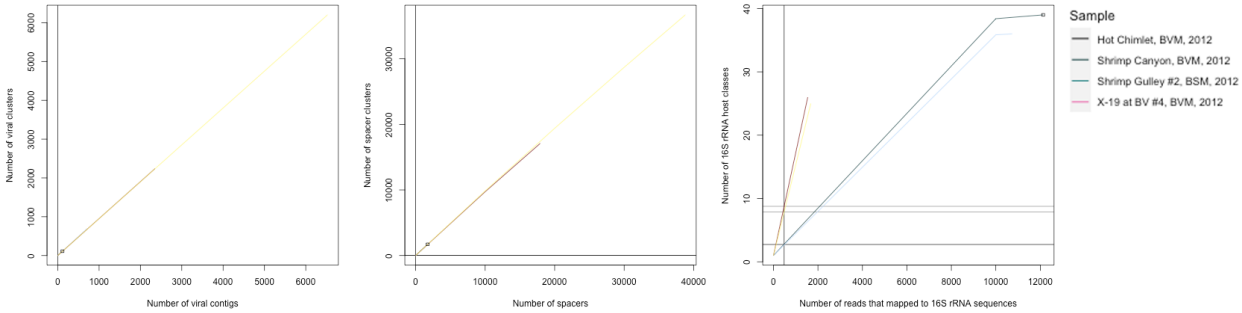
970

971 **A)**



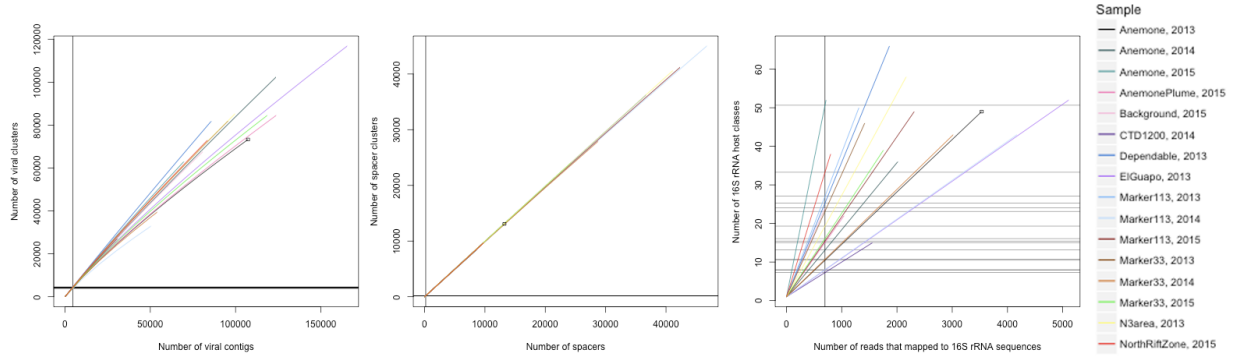
972

973 **B)**



974

975 **C)**



976

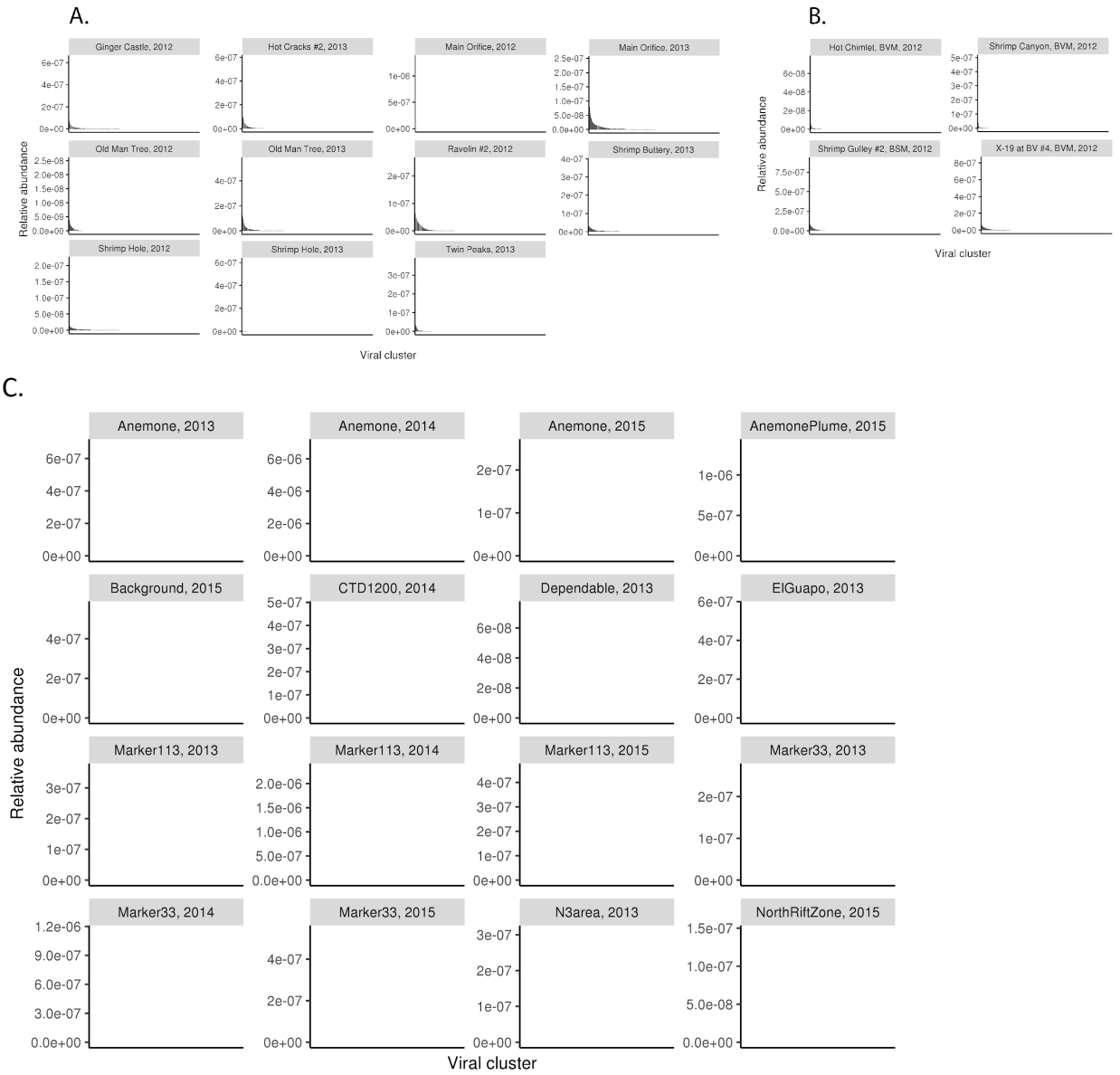
977

978 **Supplementary Figure 1. Rarefaction curves for viral clusters, spacer clusters, and 16S**
979 **rRNA host classes from samples collected at A)Von Damm, B) Piccard and C) Axial**
980 **Seamount.** There is no viral rarefaction curve for Old Man Tree, 2012 because there were no
981 viral clusters in this sample.

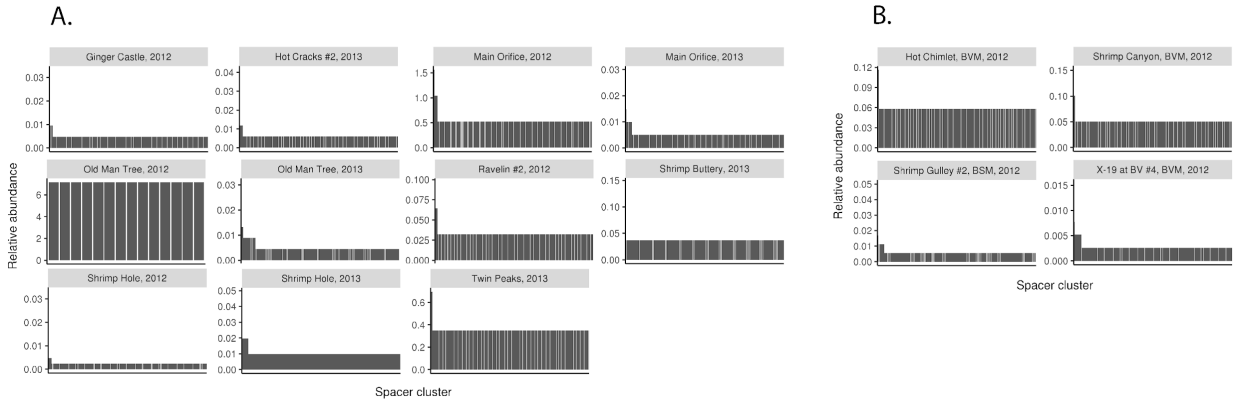
982

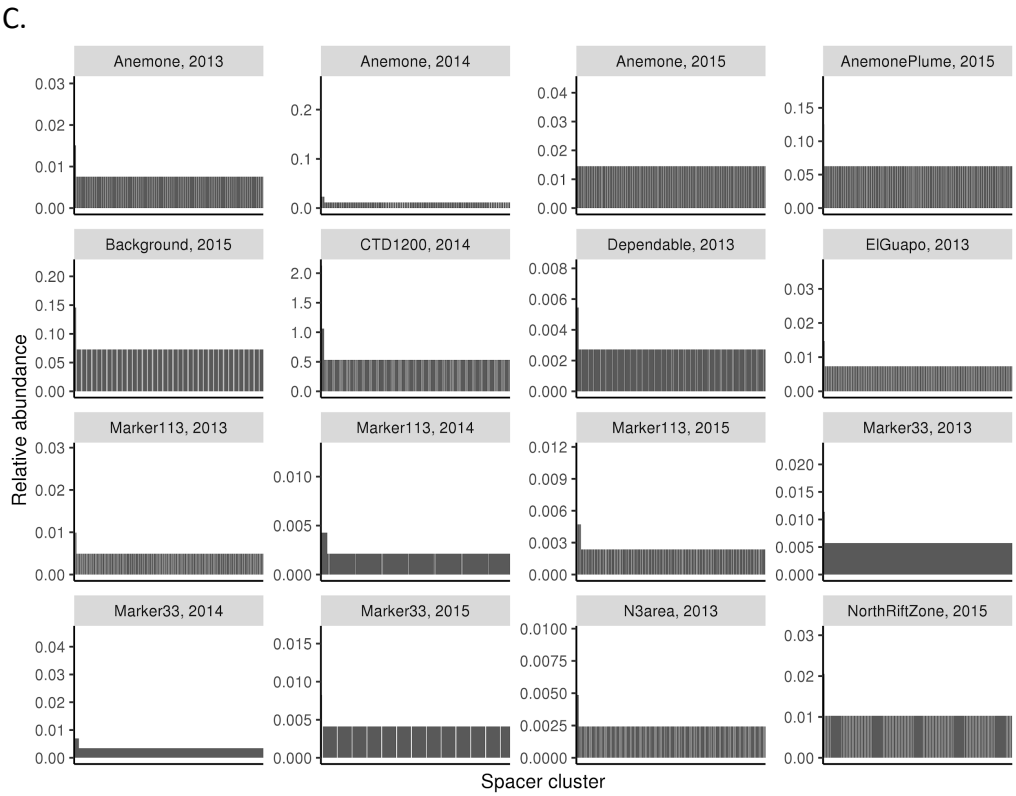
983

984

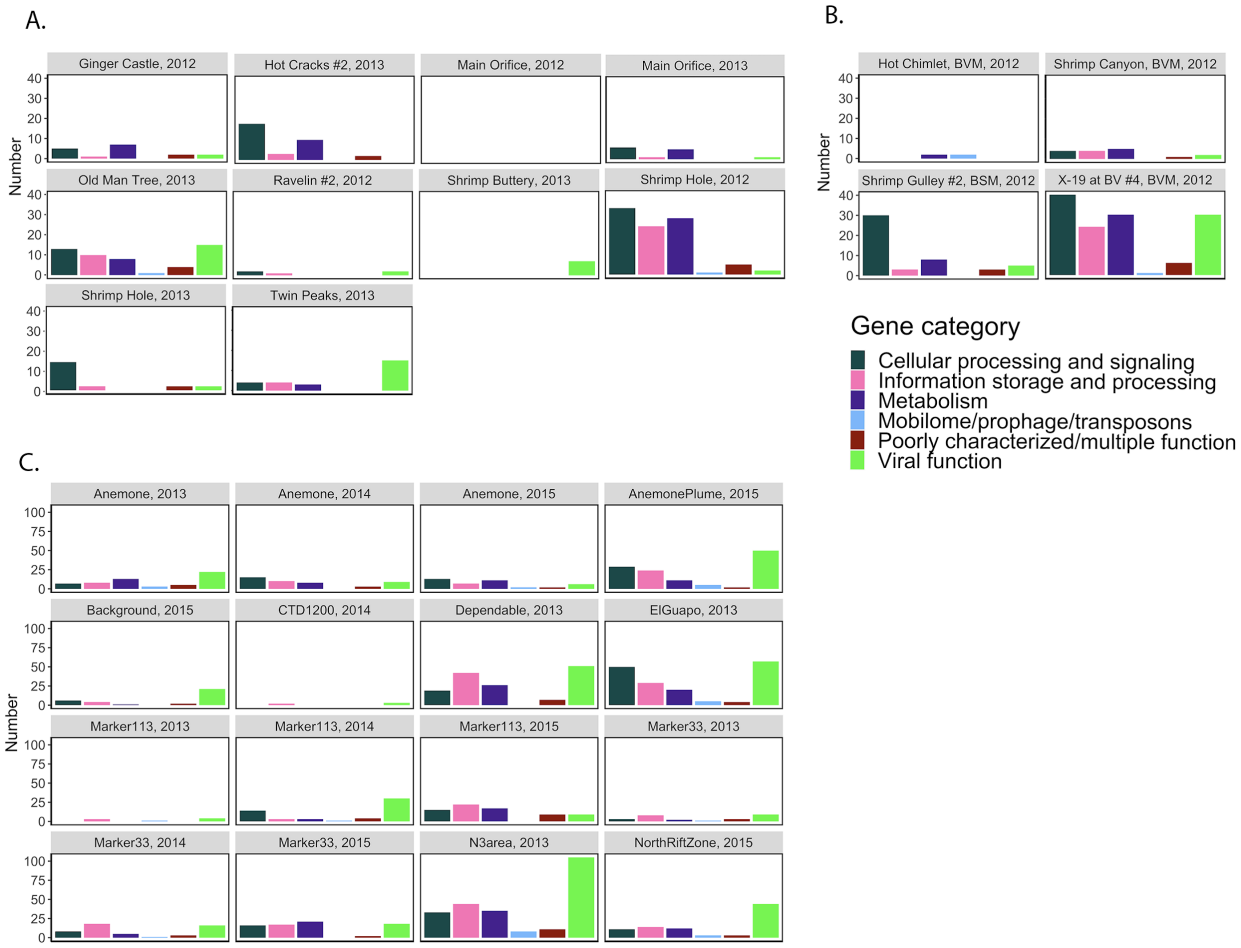


Supplementary Figure 2. Rank abundance curves of viral clusters at A) Von Damm, B) Piccard and C) Axial Seamount. The y-axis represents the number of reads that mapped to each viral cluster divided by the total length of the viral contigs in the cluster and the number of reads in the sample. The x-axis 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) Von Damm, B) Piccard and C) 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 Von Damm; B) Samples from Piccard; C) Samples from Axial Seamount.

Supplementary Tables are provided in Excel format.