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 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 Clustered Regularly InterSpaced Palindromic Repeats (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 communities and their 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 temperate 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 hydrothermal vents are diverse and apparently active, but most 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 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). Temperate viruses, which often 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 temperate 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

84 One large gap in our understanding of eco-evolutionary dynamics in deep-sea hydrothermal
85 vents is how broadly viruses are distributed within and between vent systems, and how viral
86 interactions shift across hydrothermal vent types. Previous work has shown that microbial
87 communities exhibit high endemism locally (4, 7, 26, 27), but some cosmopolitan species are
88 distributed globally (28–30). However, less is known about the geographic distribution of viral
89 populations. If viruses exhibit a restricted distribution, this would limit their role as vectors of gene
90 flow across and between hydrothermal systems. Microbial populations and their viruses are also
91 limited by environmental conditions: hydrothermal systems hosted in basalt rocks are characterized
92 by metal-enriched, low pH fluids up to 400°C (31). In contrast, hydrothermal systems hosted in
93 peridotite are influenced by serpentinization and feature organic carbon-enriched, high pH fluids
94 with slightly lower temperature (32). Microbial populations in basalt-hosted and peridotite-hosted
95 vents exhibit distinct patterns of genomic variation (33), suggesting that microbes are subject to
96 different selection pressures depending on the vent type, but we know little about the role viruses
97 play in driving those differences, nor how viral diversity varies across hydrothermal systems.
98

99 To gain further insight into viral diversity, distribution, and host range across hydrothermal
100 systems, we compared viral sequences recovered from microbial metagenomes collected from two
101 hydrothermal regions in two distinct ocean basins: Axial Seamount, a submarine volcano located on
102 the Juan de Fuca Ridge in the Pacific Ocean at ~1520 meters depth, and the Mid-Cayman Rise, an
103 ultraslow spreading ridge in the Caribbean Sea. Axial Seamount is a basalt-hosted, magma-driven
104 system with fluids that are low in pH and high in carbon dioxide and hydrogen sulfide (34).
105 Microbial communities at Axial are spatially restricted but temporally stable at individual vent sites
106 over time (4, 27). In contrast, the Mid-Cayman Rise hosts two geologically and geochemically
107 distinct hydrothermal vent fields in close proximity to each other: Piccard hydrothermal field,
108 located in basalt rocks along the ridge axis, is the deepest hydrothermal vent field discovered to date
109 (~4950 meters depth) and is characterized by fluids that are acidic with high hydrogen and hydrogen
110 sulfide content, whereas the Von Damm vent field, located approximately 20 km away on a nearby
111 massif at ~2350 meters depth, is influenced by serpentinization and is characterized by fluids that
112 are high in hydrogen, methane, and small carbon compounds (35–38). Previous work found distinct
113 community composition but similar functional potential in the microbial communities at Piccard and
114 Von Damm (7, 27, 33, 39, 40) . Here, we compare viral sequences in venting fluids from Von

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

147 different vent fields within Axial Seamount (ASHES, International District, North Rift Zone, and
148 Dependable) using the ROVs ROPOS and Jason. We analyzed a total of 16 metagenomes from 10
149 different vents, 1 plume, and 2 deep seawater samples at Axial Seamount. For collection of microbial
150 DNA, 3 L of diffuse fluid was pumped through 0.22 μm, 47 mm GWSP filters (Millipore) and the
151 filters were flooded with RNALater (Ambion) on the seafloor, as described in Fortunato et al. (4).
152 Fluids from a hydrothermal plume above Anemone and background seawater were collected using a
153 Seabird SBE911 CTD and 10 L Niskin bottles and 3 L of the plume and seawater fluid was filtered
154 through 0.22 μm Sterivex filters (Millipore). DNA was extracted from the filters using a phenol-
155 chloroform method adapted from Crump et al. (43) and Zhou et al. (44). The Ovation Ultralow
156 Library DR multiplex system (Nugen) was used to prepare metagenomic libraries following the
157 manufacturer instructions. DNA extraction and metagenomic library construction is further
158 described in Fortunato and Huber (45). The 2013 and 2014 samples were sequenced on an Illumina
159 HiSeq 1000 and the 2015 samples on a NextSeq 500. Sequencing was done at the W.M. Keck
160 sequencing facility at the Marine Biological Laboratory.
161
162 For all metagenomes, paired-end partially overlapping reads were merged and quality filtered
163 using the illumina-utils package (46) using the iu-quality-filter-minoche flag, then assembled using
164 idba-ud (47) v1.1.2 with default settings. All data from both sites is available in the European
165 Nucleotide Archives archives under study accession number PRJEB15541 for the Mid-Cayman Rise
166 and under study accession numbers PRJEB7866, PRJEB12000, and PRJEB19456 for Axial samples
167 in the years 2013, 2014 and 2015 respectively (Supplementary Table 1).
168
169 *Identification of CRISPR loci, viral populations, and spacer assemblages*
170 Crass (48) v1.0.1 was used to identify CRISPR loci (i.e., unique direct repeat types) and
171 CRISPR spacers in the metagenomic reads. Viral contigs in the metagenomic assemblies were
172 identified with VirFinder (48, 49) v1.1 using a p-value threshold of 0.05. Given that viruses from
173 hydrothermal systems are not well-represented in sequence databases, we elected to use VirFinder
174 for identification of viral contigs for diversity and abundance analysis because VirFinder is a k-mer
175 frequency-based method that avoids gene-based similarity searches, and thus has higher potential to
176 identify novel viruses. We used ClusterGenomes (https://bitbucket.org/MAVERICLab/stampede-
177 clustergenomes) v1.1.3 (95% identity, 80% coverage) to identify viral clusters (or viral operational
178 taxonomic units, vOTUs) of viral contigs. To cluster the CRISPR spacers, we performed an all-v-all

179 blast using BLASTn v2.5.0 (E-value threshold of 10^{-08}) and then clustered using Markov Cluster
180 Algorithm (MCL) (50) v14-137 (inflation 1.2 and scheme 7) based on bitscore.
181
182 *Recovery of MAGs*
183 All metagenome assembled genomes (MAGs) were recovered from metagenomic assemblies
184 using anvi'o (51). Supervised clustering was used to recover bins from Piccard and Von Damm
185 contigs using anvi'o v2.1.0 (33). MAGs from Axial Seamount assemblies were recovered using
186 unsupervised binning with CONCOCT (50, 52, 53) within the anvi'o v4.0 pipeline, followed by
187 manual refinement within anvi'o (39). For all analyses, only bins with >=70% completeness and
188 <=10% redundancy were retained as MAGs for this analysis. The MAGs were assigned taxonomies
189 using PhyloSift (50, 52) v1.0.1 as described in Anderson et al. (33). Only MAGs for which
190 taxonomies could be identified were used in analyses. MAG coverage was normalized by the
191 number of merged reads in the metagenome. All of the Mid-Cayman Rise MAGs were previously
192 described in Anderson et al. (33).
193
194 *Taxonomic identification of microbial and viral reads*
195 The reads from all metagenomes were mapped to the Silva SSU and LSU Parc databases (54,
196 55) (release 132) with bowtie2 (56) v2.2.9 using default settings and local alignment. Mapped reads
197 were assigned taxonomies using the classify.seqs function in mothur (57, 58) v1.38.1 with the Silva
198 16S rRNA database (release 119) and a cutoff of 50. Reads mapping to 16S rRNA gene sequences
199 that were classified as eukaryotes were excluded from analyses.
200 In order to identify the viral taxa at Von Damm, Piccard, and Axial Seamount, we conducted
201 a blast of the representative sequence of each vOTU against all of the viral genomic sequences in
202 RefSeq (downloaded on September 16th, 2020; E-value <=10-5). The best match for each vOTU
203 representative sequence was selected based on E-value and percent identity. The taxonomy of the
204 RefSeq viral sequence with the best alignment was assigned to the given vOTU. To quantify the
205 relative abundance of viral families in each sample, the reads that mapped to vOTUs of a viral family
206 were divided by the total number of reads in the sample.
207
208 *Viral, spacer, and host diversity*
209 Rarefaction curves for vOTUs, CRISPR spacer clusters, and reads matching 16S rRNA
210 genes categorized at the class level were created using the Vegan R package (59) v2.4-5. The number

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

243 rRNA gene sequences of the given taxon, normalized by the number of 16S rRNA gene reads in the
244 sample. Analyses of microbial taxa were done at either the class level or the lowest taxonomic level
245 available, depending on the analysis.
246
247 Microbiome datasets are compositional in nature because sequencing instruments impose an
248 arbitrary total (62). Therefore, to conduct hierarchical clustering of samples based on viral, spacer,
249 and host composition, we used the protocol outlined by Gloor et al. (62) and Gloor and Reid (63)
250 for computing distances between samples containing compositional data. For hierarchical clustering,
251 we did not normalize; we performed analyses on the number of reads that mapped to each vOTU,
252 the number of CRISPR spacers in each spacer cluster, and the number of reads that mapped to 16S
253 rRNA gene sequences of each microbial host in each sample; for microbial hosts, we did not include
254 reads that mapped to unclassified sequences or sequences classified as eukaryotes. We replaced zero
255 counts with estimates using the count zero multiplicative method for vOTUs and hosts and the
256 Bayes-Laplace Bayesian multiplicative method for spacer clusters via the zCompositions R package
257 (64) v1.2.0. Using the CoDaSeq R package (https://github.com/ggloor/CoDaSeq) (62) v0.99.3, we
258 applied a centered log-ratio (clr) transformation to the count data (that lacked zero counts), thereby
259 capturing the ratios between parts. To calculate distances between samples for hierarchical
260 clustering, we used the ward.D2 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 *Identification of putative viral contigs*
286 We used VirFinder to identify putative viral contigs because it is a k-mer frequency-based
287 method that has higher potential to identify novel viruses (Ren *et al.*, 2017). However, this method
288 allows for the possibility of false positives. VirFinder assigns scores between 0-1 to indicate the
289 likelihood that a contig is viral, with higher values reflecting a higher likelihood that the sequence is
290 viral. Almost all VirFinder sequences from Axial Seamount (99.96%), Piccard (98.54%), and Von
291 Damm (98.96%) were assigned scores greater than 0.7. 47.35% of Axial Seamount, 44.21% of
292 Piccard, and 37.92% of Von Damm VirFinder sequences had scores greater than 0.9. The mean
293 score of VirFinder sequences was 0.89, 0.88 and 0.87 for Axial Seamount, Piccard, and Von Damm,
294 respectively. Supplementary Figure 1a shows the distribution of scores for the contigs identified with
295 VirFinder at the Axial Seamount, Piccard, and Von Damm vent fields. Similarly, the majority of
296 contigs identified as putatively viral were short, with shorter average contigs at Axial Seamount (625
297 bp) than those at Piccard (2172 bp) or Von Damm (2037 bp). The VirFinder score did not increase
298 with sequence length (Supplementary Figure 1b).
299
300 *Taxonomy of viruses and hosts*
301 All analyses were carried out from diffuse flow fluids sampled directly from the vent orifice, as
302 well as vent plume waters, which were sampled up to 100m above the vent orifice and had much
303 lower temperatures (Supplementary Table 1). It is possible to examine viruses in 0.22 μm-filtered
304 fluids because this fraction includes integrated prophage, lytic infections in progress, and free viral
305 particles captured on filters. However, our analysis misses free viral particles that were not retained

306 on the filters, and we are only capturing the viral diversity and variation across sites based on the
307 viral sequences identified in the metagenomes.
308 The viral sequences identified at Axial Seamount, Von Damm, and Piccard vents represent 28
309 families of viruses (Figure 1). 13 of the viral families were present at all three vent fields. 14 viral
310 families were present at both of the Mid-Cayman Rise vent fields, Von Damm and Piccard, while
311 Axial Seamount shared 13 viral families with each of the Mid-Cayman Rise vent fields. We observed
312 a distinct difference in viral taxonomic groups between the vent fields: whereas the *Myoviridae* viral
313 family had the highest relative abundance in 12 of the 16 Axial Seamount samples, the *Myoviridae*
314 were in much lower abundance at Von Damm and Piccard, which had higher abundances of reads
315 matching the *Podoviridae* family of viruses and the *Guttaviridae*.
316
317 *Relative abundance of viral and microbial reads*
318 In order to examine the distribution and abundance of viruses across individual vent sites,
319 vent fields, and regions, we quantified the relative abundance of viral sequences, CRISPRs, and
320 spacers in each of the metagenomes (Figure 2, samples from background seawater and plumes are
321 excluded; Supplementary Table 2). We also quantified the diversity and abundance of loci and
322 spacers within CRISPRs each of the metagenomes. CRISPR loci are a microbial immune system
323 found in archaeal and bacterial genomes, consisting of direct repeats interspersed by "spacers,"
324 which match foreign DNA (predominantly viruses, but also including plasmids and other forms of
325 foreign DNA) that the cell has been exposed to previously. The relative abundance of CRISPR loci
326 serves as an indication of how many microbial lineages use CRISPR as a mechanism for viral
327 immunity. It is also important to note that CRISPR loci vary in number across microbial genomes,
328 and we did not distinguish between CRISPR loci with the same direct repeat type. Therefore, our
329 measure of CRISPR relative abundance is not a direct proxy for the abundance of viruses. Instead, it
330 gives an indication of how commonly CRISPR loci are used as an antiviral mechanism within the
331 community. Moreover, CRISPRs serve as a record of past infections, and thus while viral diversity
332 reflects the diversity of viral particles sequenced at the time of sampling, CRISPR spacer diversity
333 reflects the diversity of past viral infections.
334 Our results indicate that the relative abundance of viral sequences (Figure 2A) was similar
335 both across and within each of the hydrothermal vent regions we studied ($p = 0.42$, t-test). There
336 was a higher relative abundance of CRISPR loci in samples collected at Axial Seamount compared to
337 both Piccard and Von Damm vent fields (Figure 2B; $p = 1.49e-05$, t-test). However, there was no

338 difference in the relative abundance of spacers within CRISPR loci per read (Figure 2C; p = 0.10, t-
339 test). Within the Mid-Cayman Rise, we compared samples from mafic-hosted (Piccard) versus
340 ultramafic-hosted (Von Damm) hydrothermal systems. We did not observe significant differences in
341 the relative abundance of viral sequences, CRISPR loci, or CRISPR spacers between Piccard and
342 Von Damm (viral sequences: p = 0.66, t-test, CRISPR loci: 0.40, t-test, spacers: 0.37, t-test) (Figure
343 2A-C). Similar results emerged from comparisons among vent fields within Axial Seamount: we did
344 not observe significant differences in the relative abundance of viral sequences, CRISPR loci, or
345 CRISPR spacers between vent fields at Axial Seamount (viral sequences: p = 0.17, t-test, CRISPR
346 loci: 0.41, t-test, spacers: 0.11, t-test). Finally, within Axial Seamount, we also compared the relative
347 abundance of viruses in samples taken from plume and diffuse flow hydrothermal fluid at Anemone
348 vent. The Anemone diffuse flow samples had a higher relative abundance of CRISPR spacers and
349 CRISPR loci compared to the Anemone plume sample, which was sampled above the vent.
350 However, the relative abundance of viral sequences did not differ between the Anemone plume and
351 diffuse flow samples (Supplementary Table 2).
352
353 *Diversity of viral assemblage and microbial community*
354 Overall, viral diversity analyses revealed that the viral assemblages within these vents had high
355 richness and were not dominated by specific viral strains. With the exception of a few samples, the
356 rarefaction curves for the viruses and microbes did not reach saturation (Supplementary Fig. 2),
357 indicating that the vOTUs recovered in this study did not capture the total diversity in the samples.
358 In each sample, there were no dominant viral or spacer clusters, and each of the viral and spacer
359 clusters present was relatively rare (Supplementary Figs. 3-4). Moreover, the viral and spacer clusters
360 with the highest coverage did not correspond across samples: in both Piccard and Axial Seamount,
361 one of the most common spacer clusters matched with one of the most common vOTUs by
362 BLAST, while none of the most common spacer clusters and vOTUs aligned in Von Damm. Thus,
363 there were no dominant viral sequences that were consistently found across all samples.
364
365 We observed a more diverse viral assemblage at Axial Seamount compared to the Mid-
366 Cayman Rise. We observed higher richness of both viruses (Figure 2D; p = 0.015, t-test) and their
367 microbial hosts (Figure 2E; p = 0.0056, t-test) in samples from vents at Axial Seamount compared
368 to samples from Von Damm and Piccard vent fields. However, we did not observe meaningful
369 correlations between vOTU, spacer cluster, or host diversity. Moreover, we did not observe

370 significant differences in either viral or host diversity between the Piccard and Von Damm vent
371 fields at the Mid-Cayman Rise (viral diversity: $p = 0.053$, t-test, host diversity: $p = 0.66$, t-test). We
372 also examined the relative diversity of CRISPR spacers, which provides an indication of past viral
373 infections rather than the current virus pool. We did not observe a significant difference in the
374 diversity of CRISPR spacers between samples from Piccard, Von Damm, and Axial Seamount
375 (Figure 2F, $p = 0.093$, t-test).
376
377 We observed significant differences in viral diversity between vent sites within Axial Seamount
378 ($p = 0.022$, t-test). However, no significant differences emerged in terms of the diversity of
379 microbial hosts and CRISPR spacers (host diversity: $p = 0.74$, t-test, spacer diversity: $p = 0.11$, t-
380 test). At Anemone vent, the diffuse flow samples had higher CRISPR spacer and microbial diversity
381 than the plume sample, but viral diversity did not differ between these samples (Supplementary
382 Table 2).
383
384 *Viral distribution across vent sites*
385 In order to characterize viral distribution and community similarity across hydrothermal vent
386 fluids, we evaluated the extent to which viral sequences and CRISPR spacers were distributed across
387 samples, then compared these results to the host microbial community. As before, we clustered
388 sequences based on similarity to compare across all diffuse flow samples, excluding background
389 seawater and plume samples. On the whole, viral sequences and CRISPR spacers had fairly limited
390 distributions. Only a few vOTUs (0.02%) were present at all vent sites (Von Damm, Piccard, and
391 Axial Seamount), and no CRISPR spacer clusters were present in all three regions (Table 1). The
392 most cosmopolitan viruses and CRISPR spacers did not match each other: at Von Damm, two of
393 the 100 most widely distributed vOTUs matched with three of the 100 most widely distributed
394 CRISPR spacer clusters according to BLAST. At Piccard, one of the most widely distributed vOTUs
395 aligned to one of the most widely distributed spacer clusters. At Axial Seamount, one of the 100
396 most widely distributed vOTUs matched with two of the 100 most widely distributed CRISPR
397 spacer clusters. In contrast, microbial taxa were much more cosmopolitan, with ~17% of taxa
398 shared between Von Damm, Piccard, and Axial Seamount according to 16S rRNA gene
399 classification at the lowest taxonomic level available (Table 1). Viral and CRISPR spacer clusters
400 were shared more widely among vent sites within Von Damm and Piccard compared to Axial

401 Seamount, but microbial lineages were shared more widely among vent sites at Axial Seamount
402 (Table 1).
403 We created hierarchical dendrograms to assess the similarity of samples based on their viral
404 content. Viral assemblages in samples from the Mid-Cayman Rise and Axial Seamount grouped
405 separately (Figure 3A). Within Axial Seamount, samples taken in successive years from the same
406 vent tended to have similar viral assemblage compositions (Figure 3A). In contrast, at the Mid-
407 Cayman Rise, samples taken from the same site in two different years did not cluster together, and
408 we observed weak clustering of samples by location (Figure 3A). Samples from Piccard vent field
409 and Von Damm vent field did not cluster separately. Grouping of microbial communities based on
410 classification of 16S rRNA reads in the metagenomes showed similar patterns to the viral
411 assemblages. Samples collected from vents at Axial and the Mid-Cayman Rise grouped separately
412 from each other, with only some clustering of samples at Piccard and Von Damm. Similarly, we
413 observed stronger temporal clustering among samples at Axial Seamount than at Von Damm
414 (Figure 3B).
415 In contrast to the viral and microbial assemblages, hierarchical dendrograms based on
416 CRISPR spacer compositions showed little clustering by location (Figure 3C). Based on spacer
417 assemblages, samples from either Axial Seamount or the Mid-Cayman Rise did not cluster together,
418 and samples taken from the same vent sites in different years did not cluster together. Very few
419 CRISPR spacer clusters were found at multiple vent sites (Table 1).
420
421 *Networks of viral infection*
422 Networks of viral infection generated from viral sequences, CRISPRs, and MAGs were used
423 to examine the distribution and host specificity of vent viruses. These virus networks show
424 connections between CRISPR spacers, which represent a record of previous viral infection in the
425 host, and viral sequences recovered from the metagenome, which represent viral sequences present
426 in the community at the time of sampling. Virus-host networks made for Axial Seamount (Figure
427 4A), Von Damm (Figure 4B), and Piccard (Figure 4C) indicate that vent viruses are restricted in
428 terms of both host range and geographic distribution. We did not observe any virus-host
429 connections shared between Piccard, Von Damm, or Axial Seamount. Within Axial Seamount, there
430 were 89 vOTUs linked to 18 MAGs, and 105 connections between distinct pairs of vOTUs and
431 MAGs (Figure 4A). We did not observe any connections between MAGs and vOTUs in any of the
432 non-diffuse flow samples (Background, CTD1200 and the Anemone plume sample). The number of

433 viral connections appeared to be related to, but was not significantly correlated with, the microbial
434 hosts' relative abundance (Supplementary Table 3): for example, a Clostridia MAG in the 2015
435 North Rift Zone sample (Clostridia_35) had high normalized coverage and was linked to nine
436 vOTUs. We observed a high number of viruses shared among *Aquificae* MAGs sampled from the
437 Anemone vent in 2013 and 2014 (Figure 4a). Finally, we observed only a single vOTU that was
438 linked to MAGs of different taxonomic classes from different vent sites, with a vOTU connected to
439 an *Ignavibacteria* MAG (Ignavibacteria_15) in the 2014 Marker 33 sample and an *Aquificae* MAG
440 (Aquificae_43) in the 2013 Anemone sample (Figure 4A).
441 The virus infection networks reflected the geographic separation observed in the cluster
442 dendrograms, and also revealed fairly narrow virus-host ranges for the viral sequences observed in
443 our datasets. Within the Von Damm vent field in the Mid-Cayman Rise, there were 62 vOTUs
444 linked to nine MAGs, with 66 connections between distinct pairs of vOTUs and MAGs (Figure 4B).
445 The size of the viral infection network was similar in Piccard, where 48 vOTUs were linked to seven
446 MAGs and there were 64 connections between distinct pairs of vOTUs and MAGs (Figure 4C). All
447 of the vOTUs in the Von Damm and Piccard networks represented relatively rare viruses with the
448 exception of one. This vOTU was present in seven of the Von Damm vents and all of the Piccard
449 vents, was among the top six most abundant viruses in both the X-19 and Shrimp Gully #2 vents
450 within Piccard, and was linked to a Campylobacterales and a Desulfobacterales MAG, both from
451 Shrimp Gully #2. This was the only vOTU with high relative abundance in at least one sample that
452 was present in the Axial Seamount, Von Damm, and Piccard viral infection networks (Figure 4).
453 Some MAGs had more viral connections than others: for example, a *Methanomicrobia* MAG
454 (Methanomicrobia_41) from the 2012 sample from Shrimp Hole within Von Damm was linked to
455 35 vOTUs (Figure 4B). However, this is not a direct indication of the number of different viruses
456 infecting a specific strain because some MAGs had more CRISPR spacers than others, increasing
457 the possibility of finding a viral connection. As with Axial Seamount, there were instances of highly
458 abundant MAGs with many viral connections (Supplementary Table 3): a *Sulfurovum* MAG
459 (Sulfurovum_99) in the 2012 sample from X-19 within Piccard had the highest normalized coverage
460 across Von Damm and Piccard and was linked to 18 vOTUs, making it the MAG with the third
461 highest number of viral connections within either network. Another *Sulfurovum* MAG
462 (Sulfurovum_37) had a high relative abundance in the 2012 sample from Shrimp Gulley #2 within
463 Piccard, and had two viral connections. Within Von Damm, Piccard, and Axial Seamount, we
464 observed several cases in which viruses were connected to multiple microbial hosts within the

465 network (Figure 4). However, in some cases the MAGs were linked to these shared vOTUs through
466 the same CRISPR direct repeat type and were in the same sample, suggesting that these specific
467 connections may have been due to matching CRISPR direct repeat types rather than true cross-
468 infection. These have been indicated in Figure 4 (red lines).
469 Finally, an additional analysis of virus-host interaction was conducted by identifying prophages
470 and auxiliary metabolic genes (AMGs) in MAGs. However, the fragmented nature of the assemblies
471 and the absence of clear viral hallmark genes next to putative AMGs did not allow us to reach strong
472 conclusions regarding the prevalence of lysogeny and AMGs in these hydrothermal vent samples.
473 These analyses are described in the Supplementary Materials.
474
475 **Discussion**
476
477 Viruses are important drivers of microbial mortality, ecology and evolution in the ocean, but
478 studies of their distribution and impact in the deep sea are lacking. Our results indicate that viral
479 diversity is high in venting fluids from hydrothermal systems, and the viruses we analyzed have
480 restricted geographic distributions and host ranges. This implies that viruses do not spread widely
481 between vent sites and that the viral role in mediating horizontal gene transfer across taxa and
482 between vent sites is relatively restricted. However, the high abundance and diversity of viral
483 sequences as well the large number of virus-host CRISPR connections implies rapid viral mutation
484 and ongoing viral infection, indicating active and ongoing interactions between viruses and their
485 microbial hosts in venting fluids from the hydrothermal vents examined here.
486
487 *Viral assemblages are taxonomically distinct and spatially restricted across vent sites*
488 Previous studies of viruses in marine systems have observed viral populations to be
489 commonly found across multiple samples (80), whereas others have found that most viruses are
490 biogeographically restricted, with only a few cosmopolitan groups (81, 82). In hydrothermal vents at
491 Von Damm, Piccard and Axial Seamount, we found high viral diversity with a limited distribution,
492 potentially indicating rapid diversification in vents. Most vOTUs we observed were found only at
493 individual vent sites, with very few cosmopolitan viruses. Previous work examining microbial
494 distribution in hydrothermal systems through fine-scale 16S rRNA gene analyses indicates that while
495 some microbial lineages are endemic to individual vent sites, others are widespread across vent fields
496 (28, 40). Our work revealed that viral biogeographic patterns roughly corresponded to those of their

497 microbial hosts. Taxonomic analysis revealed that there are distinct differences in viral taxonomy
498 between vent sites-- most notably, we observed a high abundance of *Myoviridae* at Axial Seamount
499 that was not found at either Piccard or Von Damm. These observations indicate either restricted
500 gene flow or strong environmental selection for specific groups within each site. Our analysis of
501 viral sequence similarity using hierarchical clustering revealed similar results: there were some
502 similarities between the viral assemblages in vents at Piccard and Von Damm, which are located
503 approximately 20km apart in the Caribbean Sea, but the viral assemblages from Axial Seamount in
504 the Pacific were very different from those at Piccard and Von Damm, despite the fact that vents at
505 Piccard and Axial are both hosted in basalt and have metal-enriched, low pH fluids. Previous work
506 on microbial community similarity and distribution in hydrothermal systems has shown that vents in
507 close proximity are often more similar to one another in terms of microbial community structure
508 than to geographically distant vents (4, 27, 40, 83). This may result from subseafloor plumbing that
509 restricts fluid flux between sites, creating "islands" of microbial diversity that are distinct from one
510 vent site to the next (26). Here, we show that these barriers to dispersal apply to viruses as well, and
511 that viruses may be even more spatially restricted than their microbial hosts. Microbial lineages that
512 spread between vent sites may face infection from novel viral strains not found in other vent sites.
513 These endemic viral populations thus further shape distinct microbial community structure at
514 individual vent sites.
515 Although the viral assemblages at hydrothermal vents had high diversity and restricted
516 dispersal between vent sites and vent fields, our results show that the viral assemblages in vents
517 persist over time, particularly at Axial Seamount where the same vents were sampled over a 3-year
518 period. Viral assemblages from samples from the same vent sites at Axial across three years clustered
519 together in the hierarchical dendrograms (Figure 3A), and our viral infection networks revealed that
520 a number of viruses at Axial Seamount were linked to specific microbial hosts at the same vent over
521 multiple years (Figure 4A). These patterns match those observed in the microbial communities
522 (Figure 3B) and are consistent with previous work showing spatially restricted but temporally stable
523 microbial communities over time at Axial Seamount (4, 27). Our work extends this to the viral world
524 at Axial Seamount, indicating that viral assemblages follow the same temporal patterns as their
525 microbial hosts, and that virus-host relationships persist over time. However, temporal stability in
526 the viral and microbial community was generally not preserved at the Von Damm vent field for sites
527 sampled in both 2012 and 2013. There were fewer sites and time points sampled at Von Damm

528 compared to Axial Seamount, so it is unclear whether this results from a true biological signal or
529 insufficient data.
530
531 *Vents host diverse and active assemblages of viruses with restricted host range*
532 The viral infection networks show diverse microbial lineages across several samples infected
533 by many different viruses at the time of sampling. Although individual microbial lineages could be
534 targeted by multiple viral strains, the viruses we identified in these hydrothermal habitats had fairly
535 restricted host ranges, infecting specific individual microbial strains. The only clear example of viral
536 infection across microbial taxa emerged at Axial Seamount, where viral sequences associated with
537 Ignavibacteria were also linked to Aquificales (Figure 4A). All other examples of viral infection
538 across microbial taxa may either be true examples of viral-host cross-infection or may instead result
539 from shared direct repeats. The limited host range of viruses from hydrothermal systems is
540 consistent with previous work indicating that viruses tend to be host-specific (82), and we found
541 very little evidence for generalist viruses, despite the fact that this has been reported previously (82,
542 84). Furthermore, we identified many viral connections to Epsilonbacteraeota, confirming that these
543 highly successful microbial groups in hydrothermal vent diffuse flow fluids are susceptible to viral
544 infection (3, 39, 85–87).
545 Matches between CRISPR loci and viruses provide an indication of which viruses are being
546 targeted by the CRISPR immune response. We found that while some CRISPR spacers target
547 relatively abundant viruses, most CRISPRs target relatively rare viruses in these systems. This is
548 consistent with previous observations in an archaea-dominated hypersaline lake (88), where the vast
549 majority of CRISPRs were found to target viruses with populations too small to allow for the
550 assembly of contigs. The relative scarcity of viruses targeted by CRISPRs may result from an
551 evolutionary arms race: CRISPRs limit the abundance of the viral populations they target, while
552 concurrently, viruses undergo mutations, limiting the ability of CRISPR spacers to target them.
553 Alternatively, these observations could result from an abundance of inactive spacers inherited over
554 multiple generations. However, we do not believe this to be the case because CRISPR spacer
555 clusters were infrequently present across multiple samples, suggesting that spacers were integrated
556 on sub-generational timescales.
557
558 *Patterns recorded in microbial CRISPR loci do not reflect the contemporary viral assemblage*

559 In contrast to the vOTUs, the CRISPR spacers did not demonstrate any clear biogeographic
560 patterns. We expected CRISPR spacers to be more widespread than viruses since viral composition
561 represents the virus community at the time of sampling, while spacers represent a history of viral
562 infection. In contrast to our predictions, while both had limited distributions, we found viruses to be
563 more widespread than CRISPR spacers at all scales examined (Table 1). These results are in contrast
564 to previous studies of CRISPR spacer biogeography in terrestrial hot springs, where both viral and
565 CRISPR spacers showed clear biogeographic structure (89). This suggests that there is selective
566 pressure for CRISPR spacer composition to evolve more rapidly than viral sequences via the loss or
567 mutation of CRISPR spacers. However, it is also possible that we found viruses to be more
568 widespread because of undersampling of CRISPR spacers, or our use of different clustering
569 algorithms for viruses and spacers: spacers may have been clustered at a finer resolution, resulting in
570 a narrower distribution for each spacer.
571 Given that CRISPR spacers provide a history of infection, comparing the record of past viral
572 infections via CRISPR arrays with viral sequences in the metagenomes can provide insights into
573 whether CRISPR arrays provide an accurate representation of the viral assemblage at the time of
574 sampling, as well as the rate at which CRISPR spacers are accumulated. We did not find that the
575 most common or cosmopolitan viruses and CRISPR spacers matched each other. This may arise
576 from a temporal mismatch: it takes time for spacers to be incorporated into and lost from CRISPR
577 loci as virus abundances shift and viruses evolve. Additionally, just one SNP (which we allowed for
578 when aligning spacers to viruses) can prevent a CRISPR spacer from providing resistance against a
579 virus (66). This could explain the discrepancy between viruses and spacers: once resistance of a
580 spacer to a particular virus is suppressed, the population of the virus is freed to shift independently
581 from the spacer in the host population. The discrepancy we observed between viruses and spacers is
582 important to note when attempting to use CRISPR spacers to study viral populations or vice versa.
583 Examination of the prophage abundance in MAGs revealed a relatively high abundance of
584 prophage encoded in each MAG, confirming previous results indicating that lysogeny is a common
585 lifestyle in hydrothermal systems (14, 15). However, while examination of the gene content of viral
586 contigs revealed several ORFs encoding a range of functions including cell membrane function and
587 energy production, it was difficult to prove conclusively that these genes were viral-encoded AMGs
588 rather than potential cellular contamination, and thus no concrete conclusions were provided here
589 (see Supplementary Materials). Previous research has suggested that vent viruses encode AMGs (18,

19), and thus it is likely that many of these genes represent viral-encoded AMGs, but further research is necessary to determine AMG diversity and prevalence across vent systems.

Taken together, our results show that hydrothermal vent viruses are active, abundant, and diverse. 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].

622 **References**

623 1. Jannasch HW, Mottl MJ. 1985. Geomicrobiology of Deep-Sea Hydrothermal Vents. Science.

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

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

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

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

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

636 7. Galambos D, Anderson RE, Reveillaud J, Huber JA. 2019. Genome-resolved metagenomics and
637 metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at
638 deep-sea hydrothermal vents. Environ Microbiol 21:4395–4410.

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

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

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

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

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

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

648 14. Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, Winget D, Wommack KE. 2008.
649 Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J
650 2:1112–1121.

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

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

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

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

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

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

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

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

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

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

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

674 26. Stewart LC, Algar CK, Fortunato CS, Larson BI, Vallino JJ, Huber JA, Butterfield DA, Holden JF. 2019.
675 Fluid geochemistry, local hydrology, and metabolic activity define methanogen community size and

676 composition in deep-sea hydrothermal vents. ISME J 13:1711–1721.

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

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

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

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

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

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

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

691 34. Butterfield DA, Roe KK, Lilley MD, Huber JA, Baross JA, Embley RW, Massoth GJ. 2004. Mixing,
692 reaction and microbial activity in the sub-seafloor revealed by temporal and spatial variation in diffuse
693 flow vents at axial volcano. The Subseafloor Biosphere at Mid-Ocean Ridges.

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

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

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

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

704 Cosmochimica Acta.

705 39. Moulana A, Anderson RE, Fortunato CS, Huber JA. 2020. Selection Is a Significant Driver of Gene
706 Gain and Loss in the Pangenome of the Bacterial Genus in Geographically Distinct Deep-Sea
707 Hydrothermal Vents. mSystems 5.

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

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

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

717 43. Crump BC, Hopkinson CS, Sogin ML, Hobbie JE. 2004. Microbial biogeography along an estuarine
718 salinity gradient: combined influences of bacterial growth and residence time. Appl Environ Microbiol
719 70:1494–1505.

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

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

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

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

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

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

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

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

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

737 53. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson
738 AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods
739 11:1144–1146.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

788 73. Wickham H. 2017. stringr: Simple, consistent wrappers for common string operations. R package

789 version 1.

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

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

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

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

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

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

798 80. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de
799 Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant
800 S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans
801 Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Ocean
802 plankton. Patterns and ecological drivers of ocean viral communities. Science 348:1261498.

803 81. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E,
804 Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte
805 CM, Gasol JM, Vaqué D, Tara Oceans Coordinators, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016.
806 Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature
807 537:689–693.

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

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

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

816 85. Akerman NH, Butterfield DA, Huber JA. 2013. Phylogenetic diversity and functional gene patterns of

817 sulfur-oxidizing subseafloor Epsilonproteobacteria in diffuse hydrothermal vent fluids. Front Microbiol
818 4:185.

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

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

824 88. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF. 2013. Virus-
825 host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea
826 2013:370871.

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

829 **Tables**

830

831 **Table 1.** Distribution of vOTUs, CRISPR spacer clusters, and microbial hosts (classified at the lowest
832 taxonomic level available based on 16S rRNA genes) across vent sites and vent fields at Piccard, Von Damm,
833 and Axial Seamount. The Background and CTD1200 samples were excluded. Reads that mapped to 16S
834 rRNA sequences that were classified as eukaryotes were excluded.

835

836

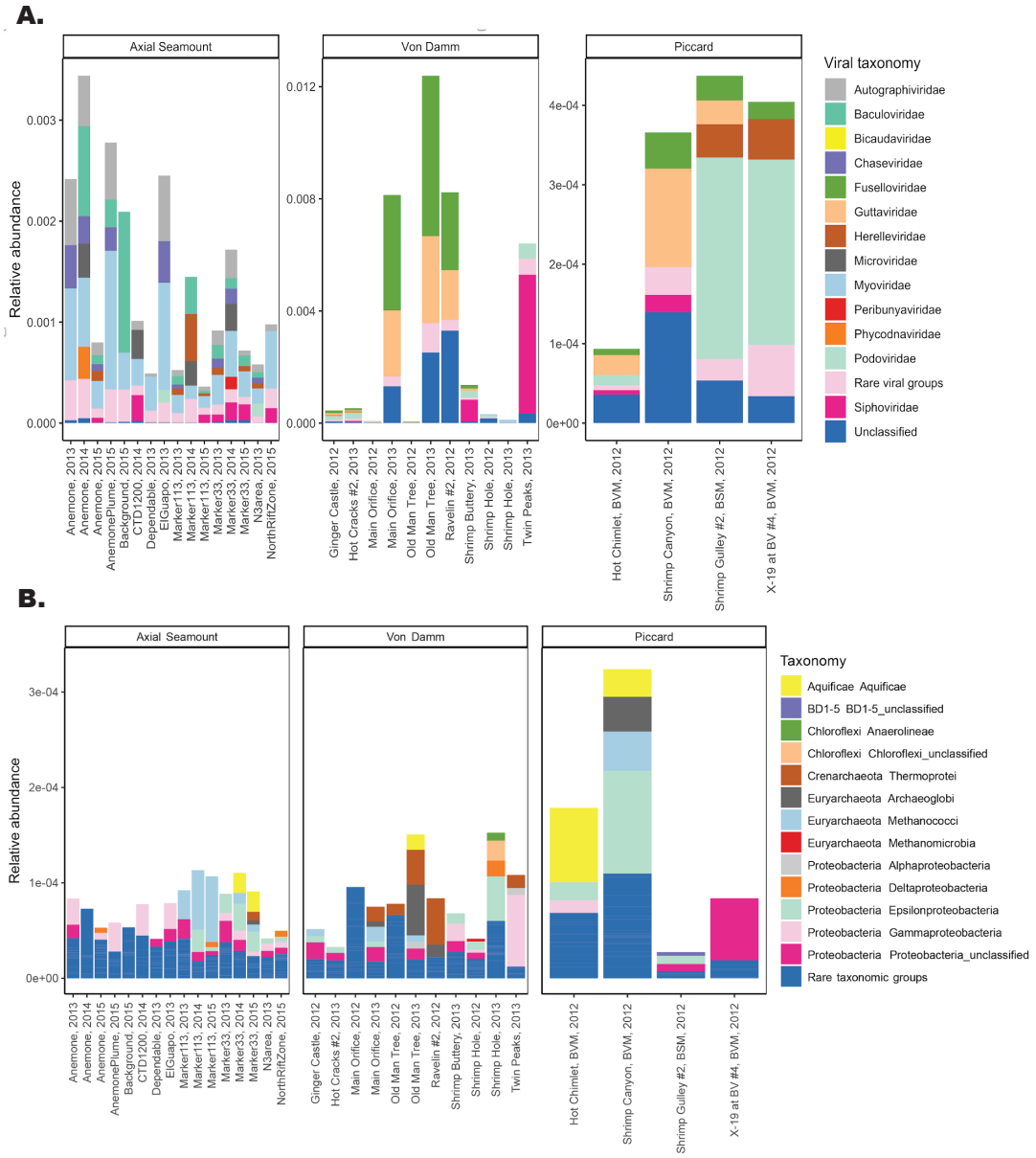
Vent field	Type	Percent in more than one vent site	Percent in all vent sites in region
Von Damm	vOTUs	13.71	0
Von Damm	Spacer clusters	5.1	0
Von Damm	16S rRNA host	46.15	1.44
Piccard	vOTUs	10.78	0
Piccard	Spacer clusters	9.13	0
Piccard	16S rRNA host	46.09	11.72
Axial Seamount	vOTUs	8.88	0.05
Axial Seamount	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 OTUs	0.02		
Spacer clusters	0		
16S rRNA host	16.62		

837

838

839

840 **Figures**



841

842

843 **Figure 1. Taxonomy of A) viral and B) microbial reads** in **the** Von Damm, Piccard, and Axial Seamount
844 vent fields. In A) the y-axis represents the number of reads that mapped to VirFinder contigs of each viral
845 taxonomy normalized by the number of reads in each sample. Rare viral groups represent viral taxonomies
846 that comprised less than 5% of reads that mapped to viral contigs in a given sample. In B) the y-axis
847 represents the number of reads that mapped to 16S rRNA gene sequences of each host normalized by the
848 number of reads in each sample. Reads that mapped to 16S gene rRNA sequences that were unclassified or

classified as eukaryotes were excluded. Rare taxonomic groups represent taxonomies that comprised less than 1% of reads that mapped to 16S rRNA gene sequences in a given sample.

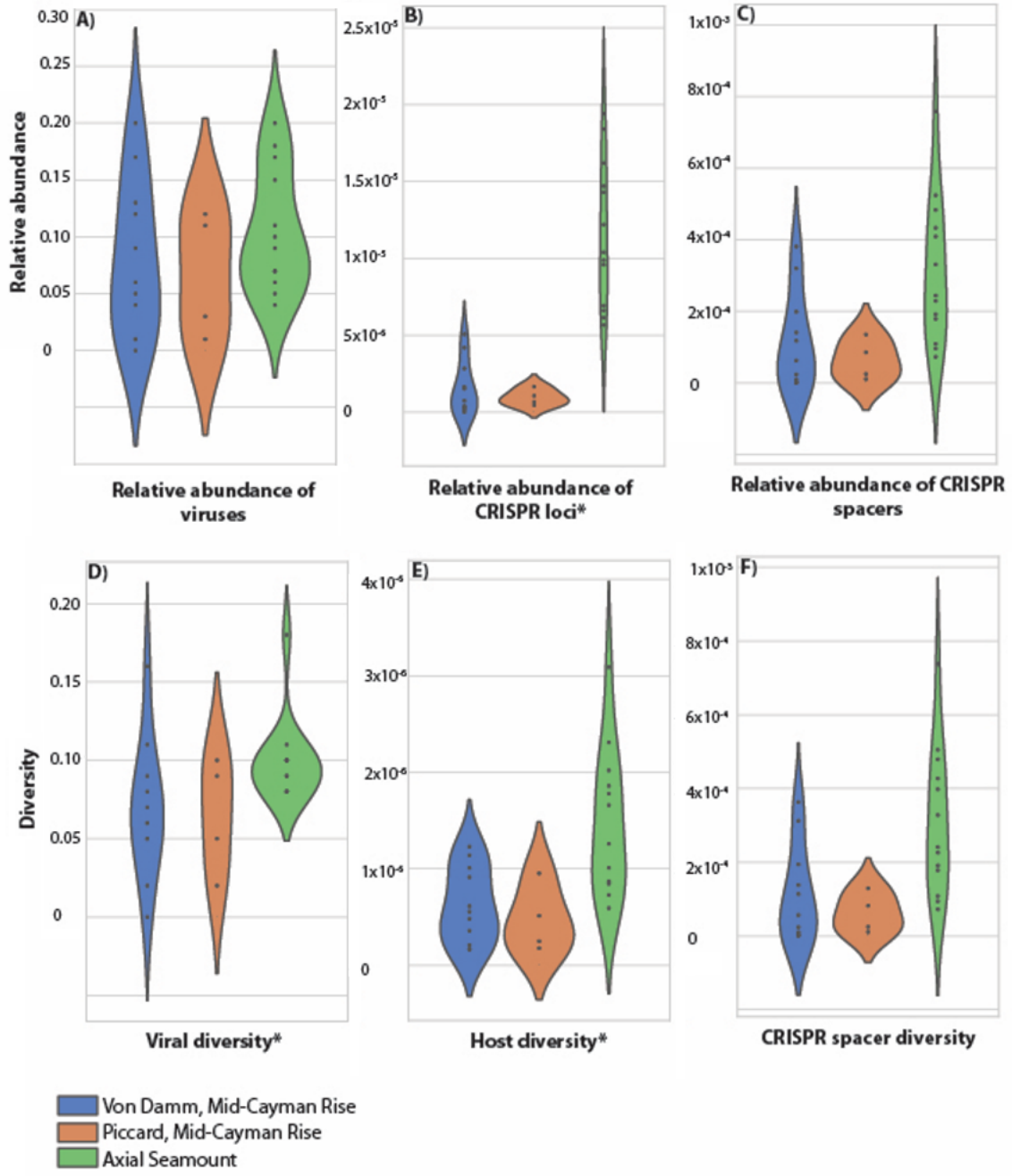
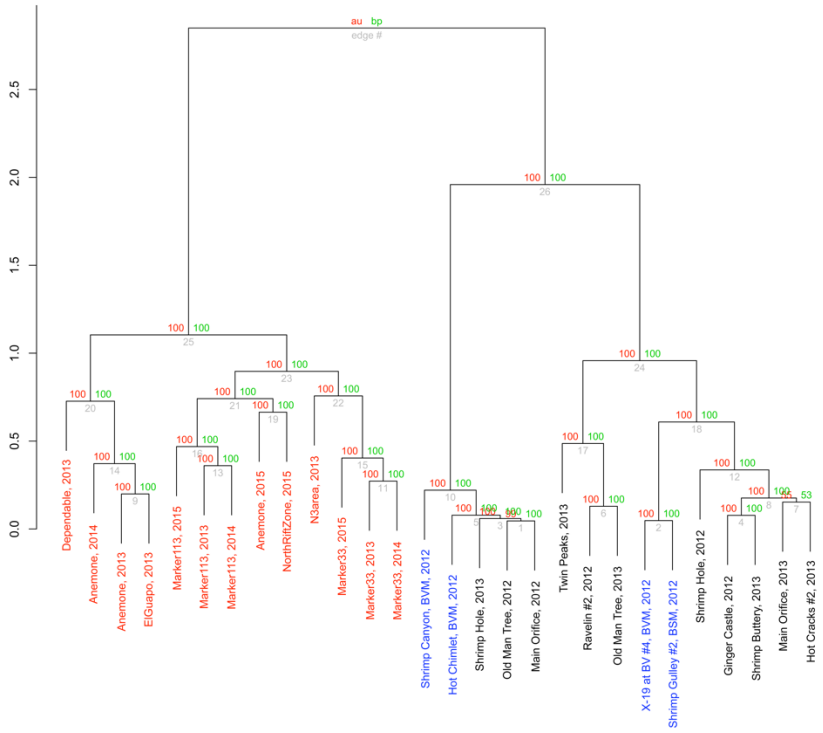
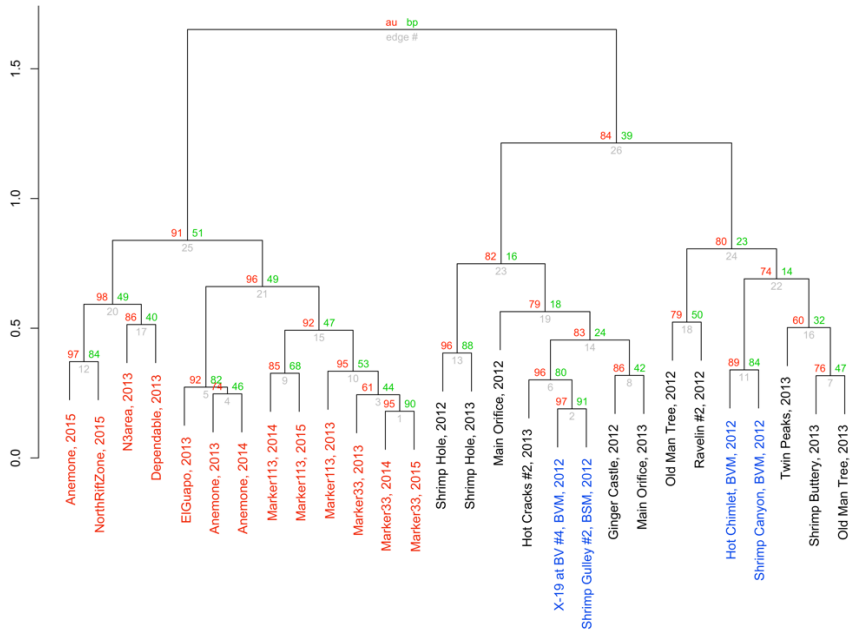


Figure 2. Abundance and diversity of viral sequences, CRISPR spacers, CRISPR loci, and microbes in diffuse flow samples from the Von Damm vent field (blue), Piccard vent field (orange), and Axial Seamount (green). Relative abundance reflects the relative number of reads mapping to viral contigs within

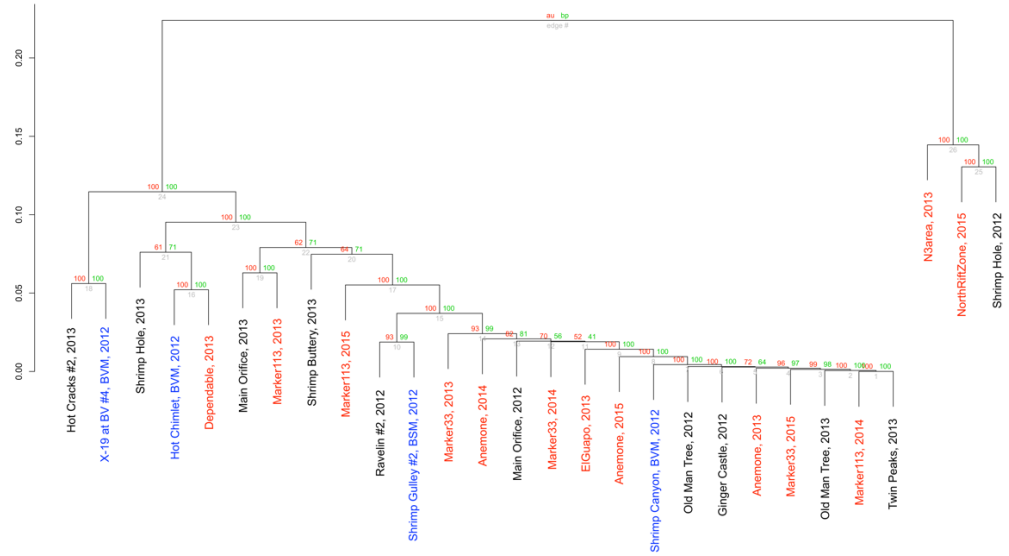
each region, normalized by read abundance (see Methods). Diversity reflects the number of clusters per read, normalized by read abundance (see Methods). Note that the y-axis values are different for each variable. Values for individual samples are indicated with black dots. Violins represent the kernel density estimation of the underlying data distribution. Variables with significant differences between Von Damm, Piccard, and Axial Seamount are indicated with asterisks (based on ANOVA test). For these comparisons, samples from background seawater and plumes were excluded.



A.



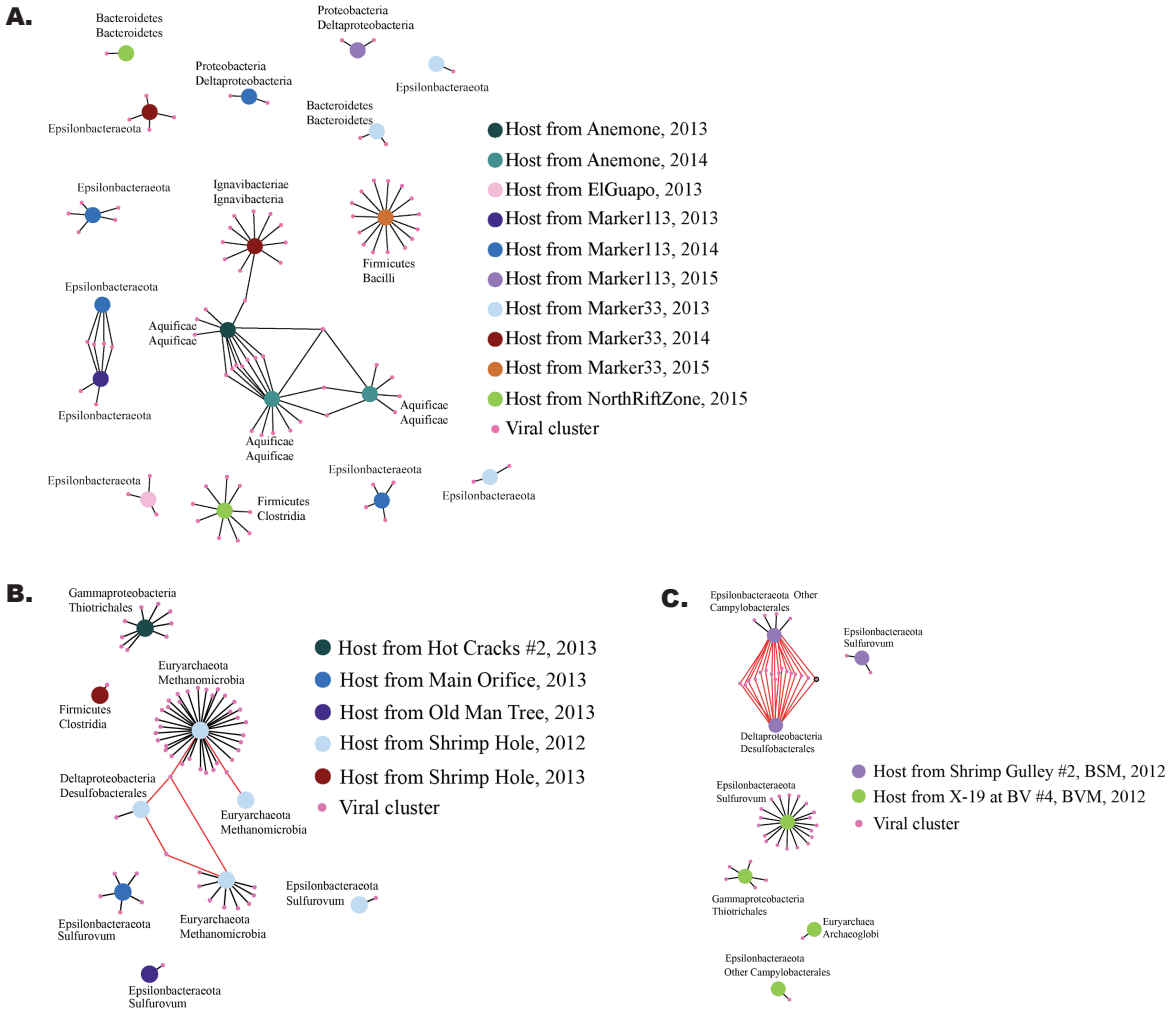
B.



C.

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 vOTU was calculated. For the host composition, the number of reads that mapped to 16S rRNA gene sequences of each host (classified at the lowest taxonomic level available) in each sample was calculated. Reads that mapped to 16S rRNA gene 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. Grey values at each edge are p-values (%). Red values are

878 approximately unbiased (AU) p-values, computed by multiscale bootstrap resampling, and green values are
879 bootstrap probability (BP) p-values, which are computed by normal bootstrap resampling. All uncertainty
880 metrics in the dendrogram were calculated using the R package pvclust.
881
882
883
884
885



886
887

888 **Figure 4. Infection network showing the links between viral clusters (v**OTUs) **and MAGs at A) Axial**
889 **Seamount, B) Von Damm and C) Piccard. v**OTUs were linked to MAGs via the spacers and direct
890 repeats in CRISPR loci. Edges are colored red when a vOTU is linked to multiple MAGs in the same sample
891 through the same direct repeat type. Due to the Crass algorithm, when the same direct repeat type is found in

892 multiple MAGs in the same sample, it cannot be determined which MAG the spacers associated with direct
893 repeat type came from. Therefore, the red edges are links between vOTUs and MAGs that could not be
894 definitely proven. vOTUs with high relative abundance (top six most abundant) in at least one sample are
895 outlined in black.
896

897 **SUPPLEMENTARY MATERIAL**

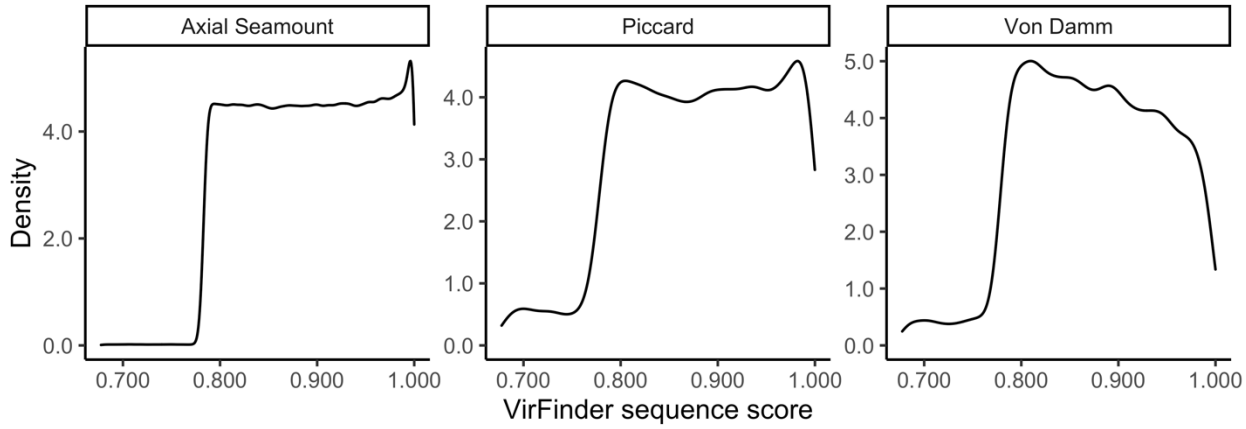
898 *Viruses and prophages in MAGS and viral and prophage genes*

899 In order to identify viral genes and putative auxiliary metabolic genes (AMGs), VirSorter (69)
900 v1.0.3 was used to extract putative viral and prophage sequences from metagenomic assemblies and
901 from individual MAGs. In the case of identifying putative auxiliary metabolic genes, we chose to
902 maximize accuracy rather than recovery. Therefore, we used VirSorter for prophage and AMG
903 identification because VirSorter relies on known viral hallmark genes for viral identification and thus
904 is a more conservative tool for positive viral identification. Category one, category two, category
905 four (prophage), and category five (prophage) VirSorter sequences were used in analyses. Putative
906 viral and prophage sequences from all assembled contigs were annotated via Prokka (69, 70) v1.14
907 with the taxonomic identifier set to "bacteria" as well as to "viruses." We assigned categories to the
908 ORFs that had been identified with the "bacteria" taxonomic identifier using the Clusters of
909 Orthologous Groups of proteins (COG) database (71, 72). We conducted a second round of
910 annotation for ORFs identified by Prokka using the "viruses" taxonomic identifier using VirSorter;
911 we classified these ORFs as having viral function if they were annotated with terms that included
912 "capsid," "tail," "spike," "terminase large subunit," "portal," and "coat."
913
914 Of the 74 Mid-Cayman Rise MAGs, 28 (38%) had at least one putative prophage sequence
915 counting VirSorter categories one, two, four, or five, and 6 (8%) had a putative prophage counting
916 only VirSorter categories four and five (Supplementary Table 4). Of the 98 Axial Seamount MAGs,
917 41 (42%) had at least one putative viral or prophage sequence counting VirSorter categories one,
918 two, four, or five, and 6 (6%) had a putative prophage sequence counting VirSorter categories four
919 and five (Supplementary Table 4). Given the uncertainties associated with binning viral contigs into
920 MAGs, the true number of prophage in hydrothermal vent microbial genomes is likely between
921 these endmember values. VirSorter has not been verified in hydrothermal systems; however, it is
922 commonly used in other non-surface ocean systems (61, 84, 85).
923
924 To determine whether viruses and prophage in these samples carried auxiliary metabolic genes
925 (AMGs), we annotated and characterized viral and prophage gene content in the viral sequences
926 recovered from the metagenomes. The vast majority of genes within the putative viral and prophage
927 sequences were annotated as belonging to the COG categories replication and repair; nucleotide
928 metabolism and transport; and post-translational modification, protein turnover, chaperone
929 functions. Many genes were also annotated as belonging to the cell wall/membrane/envelope
930 biogenesis category. Within the Mid-Cayman Rise, a total of 106 ORFs within putative viral or
931 prophage sequences identified in 10 different metagenomes fell within the broad COG category of
932 "metabolism" (Supplementary Fig. 5). This included genes categorized as energy production and
933 conversion, amino acid transport and metabolism, nucleotide transport and metabolism,
934 carbohydrate transport and metabolism, coenzyme transport and metabolism, lipid transport and
935 metabolism, and inorganic ion transport and metabolism. Samples taken from X-19 and Shrimp
936 Hole (2012) had a high number of metabolism genes (30 and 28, respectively) on putative viral or
937 prophage sequences compared to the other Mid-Cayman Rise samples. A higher proportion of
938 genes on putative viral or prophage sequences were related to metabolism at the Mid-Cayman Rise
939 than Axial Seamount (Supplementary Fig. 5). At Axial Seamount, we identified a total of 185 ORFs
940 on putative viral or prophage sequences from 14 different metagenomes that fell within the broad
941 COG category of "metabolism." However, it is important to note that none of the observed
942 metabolism genes were surrounded by confirmed virus genes (i.e., known viral ORFs both upstream

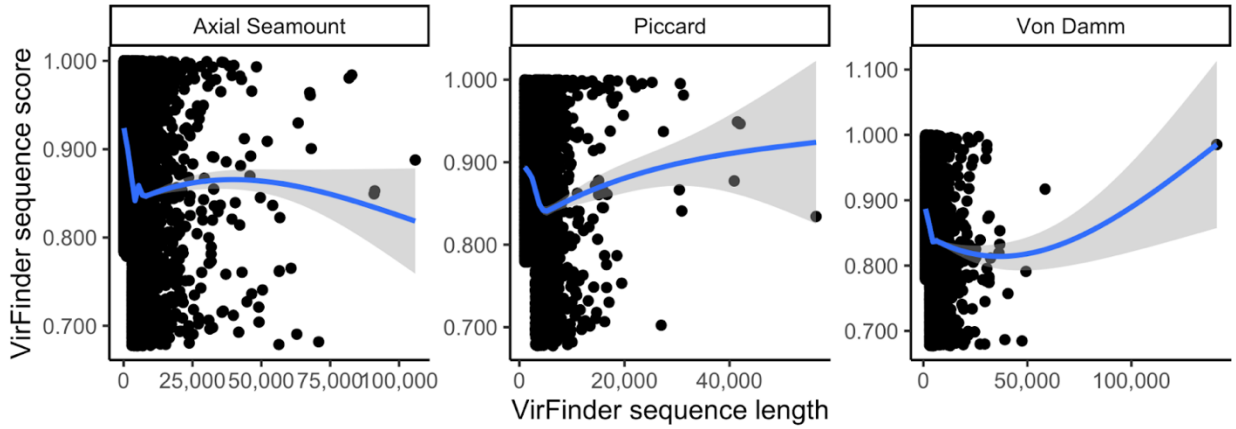
943 and downstream of the metabolic ORF), and therefore we cannot rule out the possibility that some
944 of these genes were microbial in origin.
945
946 *Auxiliary metabolic genes and the role of virus-driven horizontal gene transfer in these systems*
947 Previous work has suggested that viruses in hydrothermal systems have the genomic capacity
948 to alter their hosts' microbial metabolism through harboring auxiliary metabolic genes (AMGs) (17–
949 19). We found a higher percentage of microbial genomes with putative viral sequences (40%) than
950 has been previously reported for single-cell genomes (SAGs, 10%) in diffuse flow hydrothermal
951 fluids (15). The high abundance of prophage identified here is consistent with previous work
952 documenting a high incidence of prophage in diffuse flow systems compared to deep seawater (14).
953 The ORFs on the viral contigs identified in these hydrothermal vent metagenomes encoded
954 a wide range of functions, some of which may function as auxiliary metabolic genes. Many viral
955 contigs contained genes related to cell wall or membrane proteins. The function of these genes is
956 unknown; it is possible that these genes are involved in the synthesis of membranes surrounding
957 viral capsids. Genes related to outer membrane proteins and protein glycosylation are commonly
958 observed as variable genes within microbial pangenomes (94–100), possibly as a means to vary
959 membrane proteins to evade viral infection. Viruses often act as a source of genes for horizontal
960 gene transfer via transduction, and it is possible that the introduction of these novel genes may
961 enable some microbial strains to avoid infection by other viruses. We also observed several genes
962 with functions related to energy metabolism, inorganic ion transport, and signal transduction.
963 Previous work has found that genes related to inorganic ion transport are differentially distributed in
964 the *Sulfurovum* pangenome according to nutrient availability (39) and further work on viral sequences
965 across vent sites is needed to determine whether these genes are carried by viruses to benefit their
966 hosts. Moreover, work on viruses in hydrothermal plumes (19) and in diffuse flow vent fluid (18)
967 has indicated that viruses in these systems encode energy-metabolizing AMGs, potentially to
968 supplement the host's ability to generate sufficient energy for cellular processes during the course of
969 infection. The abundance and diversity of metabolic genes observed on these contigs provide
970 further evidence that viral-encoded AMGs are widespread and diverse within hydrothermal vent
971 ecosystems. These genes also have the potential to be horizontally transferred via viral transduction.
972 However, our analyses of viral biogeography in hydrothermal vents has revealed virus-host
973 interactions to have limited distributions because viruses are spatially restricted and highly host-
974 specific. Thus, the potential for viruses to act as mechanisms for horizontal gene transfer between
975 spatially and phylogenetically distant hosts is likely limited.
976
977
978
979
980
981
982
983
984
985

986 **SUPPLEMENTARY FIGURES**

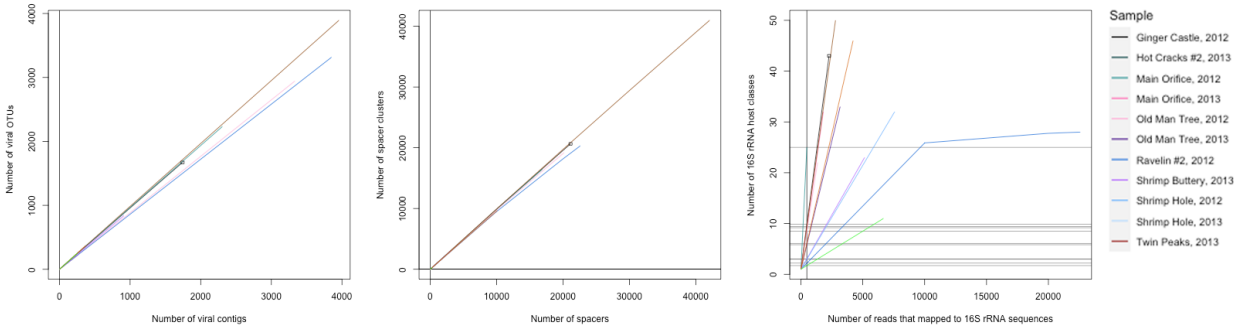
987



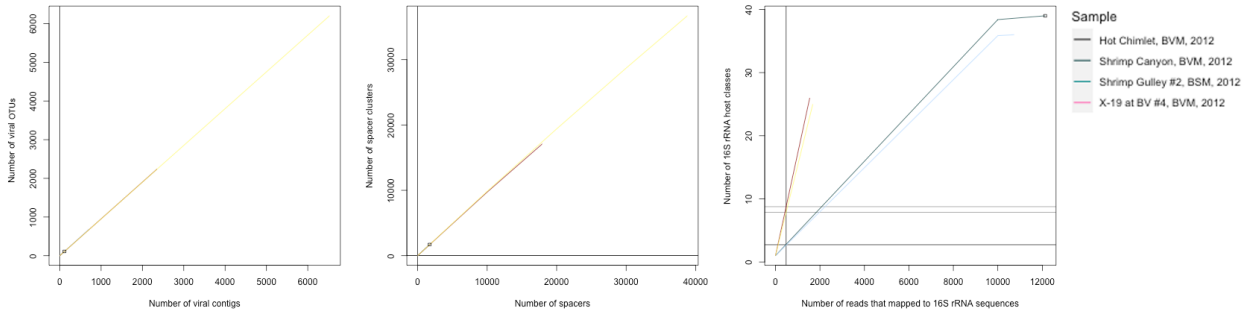
988
989 **Supplementary Figure 1**. Distributions of the VirFinder score of putatively viral sequences identified
990 at the Axial Seamount, Piccard, and Von Damm vent fields. The majority of VirFinder sequences are
991 assigned scores above 0.8 at all three vent fields.
992



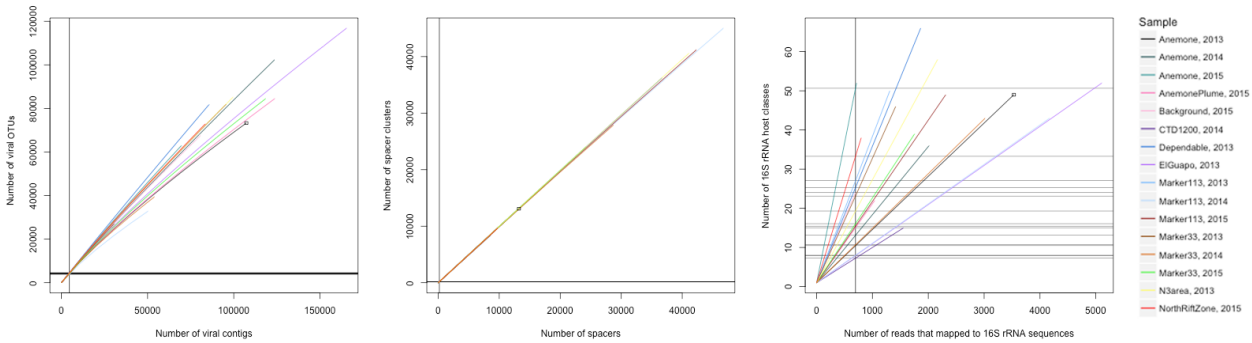
993
994 **Supplementary Figure 2**. VirFinder sequence score vs. length at the Axial Seamount, Piccard, and Von
995 Damm vent fields. The VirFinder score does not increase with sequence length for the majority of
996 sequences.
997
998
999 **A)**



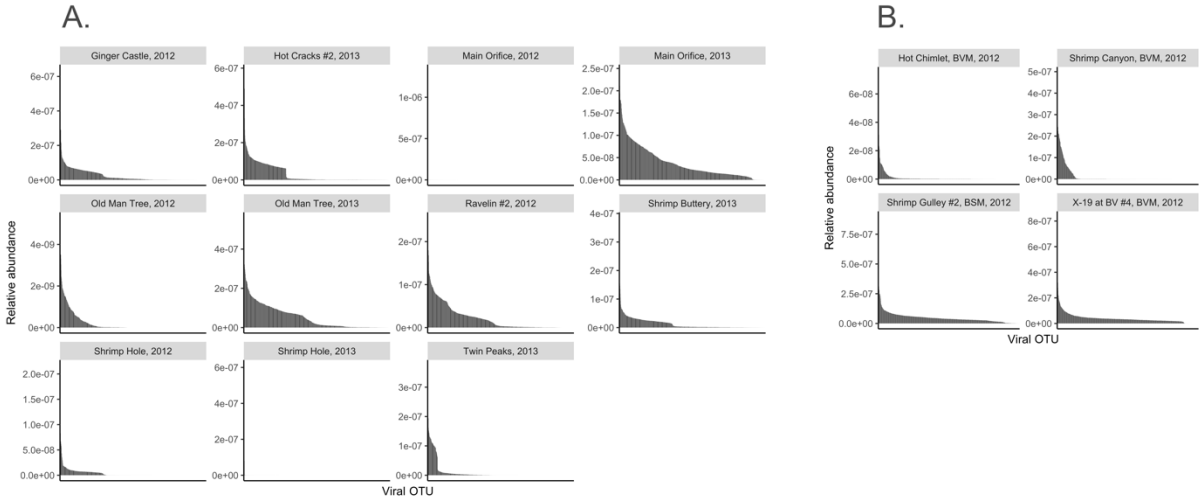
1000
1001 **B)**



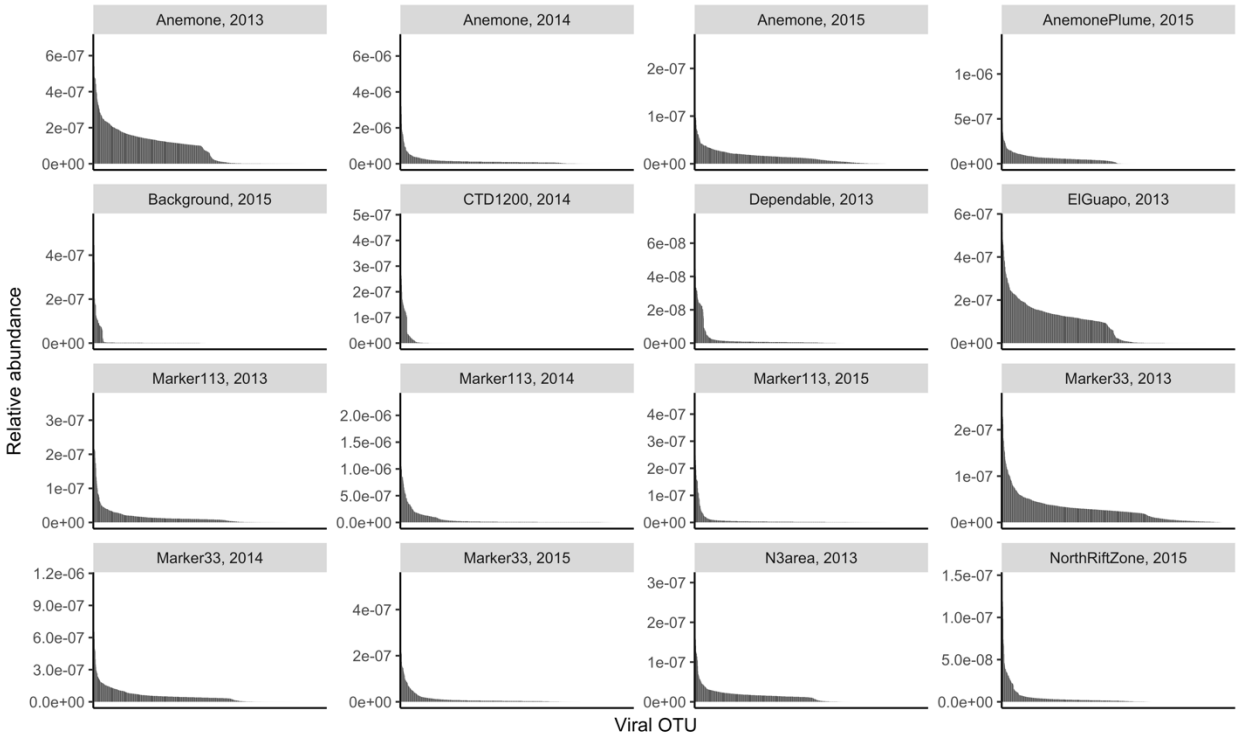
C)



Supplementary Figure 3. Rarefaction curves for viral OTUs, spacer clusters, and 16S rRNA host classes from samples collected at A) Von Damm, B) Piccard and C) Axial Seamount. There is no viral rarefaction curve for Old Man Tree, 2012 because there were no viral OTUs in this sample.

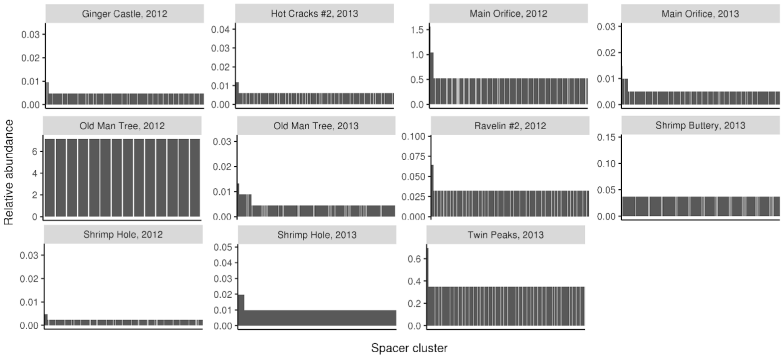


C.

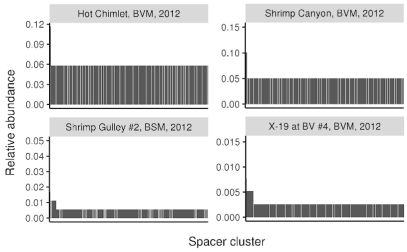


Supplementary Figure 4. Rank abundance curves of vOTUs at A) Von Damm, B) Piccard and C) Axial Seamount. The y-axis represents the number of reads that mapped to each vOTU divided by the total length of the viral contigs in the cluster and the number of reads in the sample. The x-axis represents the vOTUs ordered according to relative abundance, with the most abundant vOTUs at the left. For each vent field, the 1000 most abundant viral OTUs are shown.

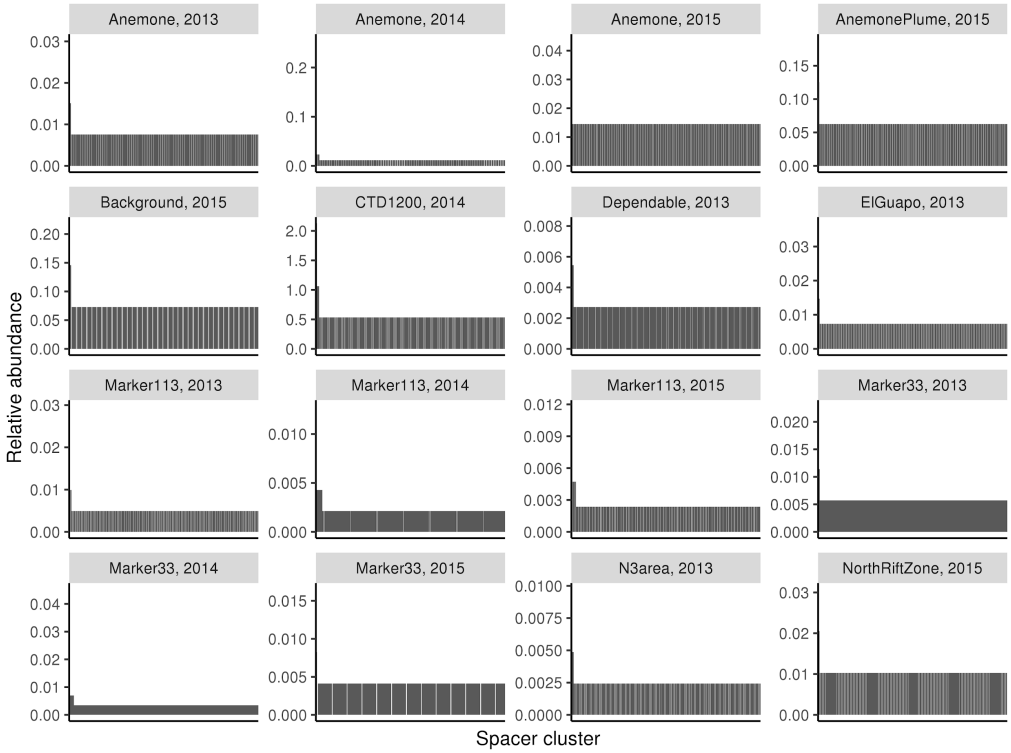
A.



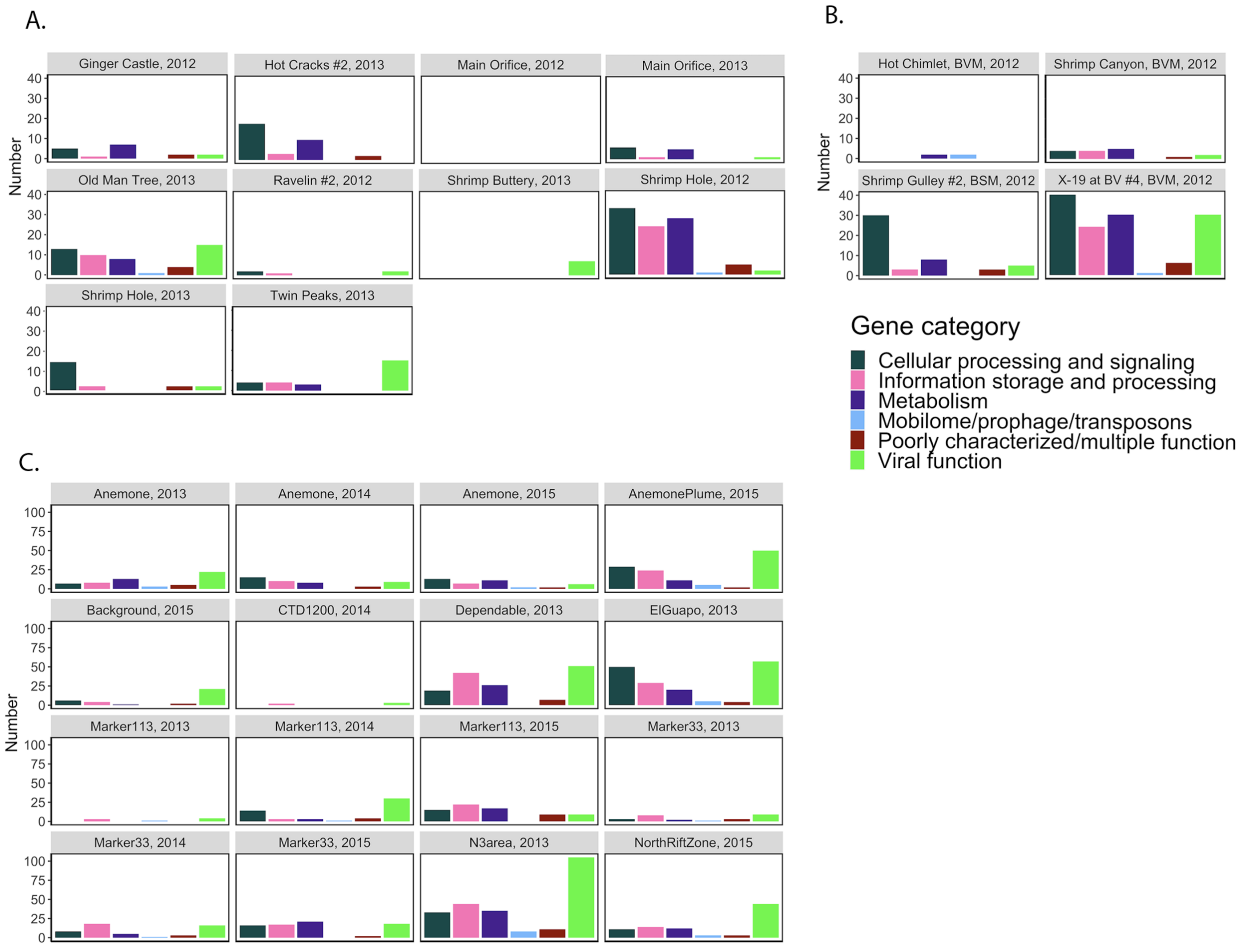
B.



C.



Supplementary Figure 5. 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 6. 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.