Horizontal gene transfer and CRISPR targeting drive phage-bacterial host interactions and co-evolution in pink berry marine microbial aggregates

Running title: Phage-host co-evolution in marine microbial aggregates

James C. Kosmopoulos*[1,2,3], Danielle E. Campbell*[#3,4,5], Rachel J. Whitaker[3,7,8], Elizabeth G. Wilbanks[#3,6]

[1] Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, USA

[2] Microbiology Doctoral Training Program, University of Wisconsin-Madison, Madison, Wisconsin, USA

[3] Microbial Diversity 2020, University of Chicago Marine Biological Laboratory, Woods Hole, MA, USA

[4] Department of Medicine, Division of Infectious Diseases, Washington University School of Medicine, St. Louis, MO, USA

[5] Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

[6] Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, Santa Barbara, California, USA

[7] Department of Microbiology, University of Illinois, Urbana, Illinois, USA

[8] Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana, Illinois, USA

*These authors contributed equally to this work.

[#]Address correspondence to ewilbanks@ucsb.edu or cdanielle@wustl.edu

ABSTRACT

Bacteriophages (phages), viruses that infect bacteria, are the most abundant components of microbial communities and play roles in community dynamics and host evolution. The study of phage-host interactions, however, is made difficult by a paucity of model systems from natural environments and known and cultivable phage-host pairs. Here, we investigate phage-host interactions in the ”pink berry” consortia, naturally-occurring, low-diversity, macroscopic aggregates of bacteria found in the Sippewissett Salt Marsh (Falmouth, MA, USA). We leverage metagenomic sequence data and a comparative genomics approach to identify eight compete phage genomes, infer their bacterial hosts from host-encoded clustered regularly interspaced short palindromic repeats (CRISPR), and observe the potential evolutionary consequences of these interactions. Seven of the eight phages identified infect the known pink berry symbionts *Desulfofustis* sp. PB-SRB1, *Thiohalocapsa* sp. PB-PSB1, and *Rhodobacteraceae* sp. A2, and belong to entirely novel viral taxa, except for one genome which represents the second member of the *Knuthellervirus* genus. We further observed increased nucleotide variation over a region of a conserved phage capsid gene that is commonly targeted by host CRISPR systems, suggesting that CRISPRs may drive phage evolution in pink berries. Finally, we identified a predicted phage lysin gene that was horizontally transferred to its bacterial host, potentially via a transposon intermediary, emphasizing the role of phages in bacterial evolution in pink berries. Taken together, our results demonstrate that pink berry consortia contain diverse and variable phages, and provide evidence for phage-host co-evolution via multiple mechanisms in a natural microbial system.

IMPORTANCE

Phages (viruses that infect bacteria) are important components of all microbial systems, where they drive the turnover of organic matter by lysing host cells, facilitate horizontal gene transfer (HGT), and co-evolve with their bacterial hosts. Bacteria resist phage infection, which is often costly or lethal,

through a diversity of mechanisms. One of these mechanisms are CRISPR systems, which encode arrays of phage-derived sequences from past infections to block subsequent infection with related phages. Here, we investigate bacteria and phage populations from a simple marine microbial community known as "pink berries" found in salt marshes of Falmouth, Massachusetts, as a model of phage-host co-evolution. We identify eight novel phages, and characterize a case of putative CRISPR-driven phage evolution and an instance of HGT between phage and host, together suggesting that phages have large evolutionary impacts in a naturally-occuring microbial community.

INTRODUCTION

Phages, viruses that infect bacteria, occur in all microbial ecosystems, often outnumbering bacteria by 10 to 1, and play pivotal roles in altering community structure (Andersson & Banfield, 2008; Bergh *et al.*, 1989; Breitbart *et al.*, 2018), mediating horizontal gene transfer (HGT) (Breitbart *et al.*, 2018; Hall *et al.*, 2017; Schneider, 2021), and driving bacterial evolution (Campbell *et al.*, 2020; Koskella & Brockhurst, 2014; Martiny *et al.*, 2014). Though some phage-host interactions can be beneficial, phage infection canonically ends with the lysis and death of the host to release progeny phage particles for transmission to new host cells. Thus, there is strong selection for bacteria to evolve mechanisms to resist infection. Likewise, phages must evolve to overcome those resistances to survive. This phage-bacterial host coevolution is often described as an "arms race" (Hampton *et al.*, 2020).

Bacteria have evolved a wide range of phage defense systems, such as clustered regularly interspaced short palindromic repeat (CRISPR) loci, which act as microbial adaptive immune systems. During a new phage infection, CRISPR systems incorporate short segments of phage-derived sequence, known as "protospacers," into CRISPR arrays as "spacers" (Barrangou *et al.*, 2007; Jansen *et al.*, 2002; Jinek *et al.*, 2012). CRISPR systems further encode mechanisms to degrade invading

73 phage DNA that matches an existing spacer, allowing the host to resist infection. CRISPR arrays thus
74 serve as a genetic record of phages a host has encountered and can be leveraged to identify phage
75 hosts from sequence data (Childs *et al.*, 2014; England *et al.*, 2018).

76 In-depth analysis of CRISPR targeting offers insights into phage-host interactions. CRISPR
77 systems often target conserved phage sequences, thus conferring protection from groups of related
78 phages (Barrangou *et al.*, 2007; Deveau *et al.*, 2008; Mojica *et al.*, 2009). Phages can accumulate
79 mutations within protospacers which allow the phage to escape CRISPR defenses (Andersson &
80 Banfield, 2008; Deveau *et al.*, 2008; Sun *et al.*, 2013). Together, these patterns in CRISPR spacer and
81 protospacer nucleotide variation can shed light on phage-host co-evolution.

82 Here, we investigate phage-bacterial interactions in microbial consortia known as "pink
83 berries." Pink berries are macroscopic microbial aggregates found in the Sippewissett Salt Marsh of
84 Falmouth, MA (Seitz *et al*., 1993; Wilbanks *et al.*, 2014). These aggregates are primarily composed of
85 phototrophic, sulfide-oxidizing and sulfate-reducing bacteria that together form a syntrophic sulfur
86 cycle (Wilbanks *et al.*, 2014). Though phage sequences have been found in pink berries previously
87 (Wilbanks *et al.*, 2022), the interactions of phage and their host remain uncharacterized. We show that
88 pink berries host diverse, novel phages that drive bacterial evolution through HGT and putative
89 CRISPR-driven arms race dynamics.

90

91 **RESULTS**

92 ***Pink berries contain novel phages that are variable between individual aggregates***

93 Pink berries were strategically sampled and pooled during sequencing library preparation to
94 observe both the breadth of diversity across individual pink berries, and to deeply sample the total
95 diversity of pink berries. Thus, we sequenced two metagenomes from single pink berries, one with low
96 sequencing depth (LS06-2018-s01) and one with high sequencing depth (LS06-2018-s02), and one

97 metagenome from three pink berries homogenized together and sequenced at very high depth (LS06-
98 2018-s03) (Table 1). Co-assembly of the three pink berry metagenomes yielded 184 contigs totaling
99 4.35 Mb in length (Table 1). Two phage sequence prediction tools, VIBRANT (Kieft *et al.*, 2020) and
100 ViralVerify (Antipov *et al.*, 2020), identified nine full-length, circular phage genomes which were the
101 targets of our downstream analyses (Suppl. Fig. 1). The phages were named according to their hosts
102 predicted by CRISPR spacer-protospacer matches (described in *Pink berry phages are targeted by*
103 *bacterial CRISPR systems).*

104 Functional annotations were obtained for 236 of 632 putative protein-coding genes predicted
105 across the nine complete viral genomes (Suppl. Data 1). One virus was found to have large amounts of
106 homology to eukaryote-infecting Circular Rep-encoding Single Stranded (CRESS) DNA viruses and
107 may infect a nematode host, which are frequently observed grazing on bacteria in pink berries. This
108 virus was excluded from further analysis to focus on the primary pink berry bacterial components and
109 their phages ("Pink berry virus MD00"; Suppl. Data 1). Functional annotations for the remaining 8
110 phage genomes predicted nucleotide metabolism proteins, head and packaging proteins, integration
111 and excision proteins, transcriptional regulators, and various lytic proteins (Fig. 1, Suppl. Data 1).
112 Importantly, all phages of interest here are predicted to exhibit strictly lytic lifecycles, as they lack the
113 genes for a temperate lifecycle, such as an integrase. There were also predictions for several auxiliary
114 metabolic genes (Breitbart *et al.*, 2007), which we inferred based on functional predictions outside the
115 core functions required for the phage lifecycle. One such AMG was a *darB*-like antirestriction gene
116 encoded on the Thiohalocapsa phage MD04 genome (Suppl. Data 1). Interestingly, *darB* has been
117 shown to methylate phage DNA to resist host restriction modification (RM) systems (Iida *et al.*, 1987;
118 Iyer *et al.*, 2017). Prior work found pink berry bacteria employ numerous, diverse RM systems, and
119 revealed that putative pink berry phages have been shown to contain similar methylation profiles as
120 their hosts (Wilbanks *et al.,* 2022). The presence of AMGs such as *darB* suggest that pink berry

121 phages have adapted to increase their fitness in *Thiohalocapsa* hosts, consistent with an armsrace-like
122 process of coevolution.

123 Phage taxonomy is based on genome similarity (Turner *et al.*, 2021), and forms the basis of
124 inferring the diversity of phages within a community. We hypothesized that pink berry phages would
125 be at least as diverse as their pool of potential bacterial hosts, which span multiple phyla. vConTACT
126 (Bin Jang, *et al.*, 2019), a genome-wide protein similarity-based approach, was used to infer phage
127 taxonomy for the eight phage genomes of interest. Although nominal protein similarity was detected
128 between all phages of interest and the Viral RefSeq protein database during gene annotation (Suppl.
129 Data 1), only Desulfofustis phage MD02, Thiohalocapsa phage MD04, and Rhodobacteraceae phage
130 MD07 are connected to a known phage in the vConTACT network (Fig. 2, Suppl. Data 2). Further,
131 only one phage genome, Desulfofustis phage MD02, had sufficient protein similarity to cluster with a
132 cultured phage reference genome, Pseudomonas phage PMBT14, which is currently the only species
133 of the genus *Knuthellervirus* (Suppl. Data 2). None of the phage genomes of interest clustered with
134 each other. The remaining five genomes of interest lacked sufficient protein similarity for a connection
135 in the vConTACT network, indicating that these phages represent novel and undescribed diversity.

136 Relative bacterial abundances are similar across pink berries (Wilbanks *et al.*, 2014), which is
137 likely due to the constraints of the syntrophic metabolic interactions between constituent bacteria. To
138 assess the distribution of pink berry-associated bacteria and phages, genome-wide read coverages were
139 analyzed for individual pink berry metagenomes (Fig. 3). In agreement with previous observations, we
140 found that the relative abundance of pink berry bacteria is relatively homogenous across samples (Fig.
141 3A). In contrast, phage presence and abundance are highly variable between different pink berry
142 communities (Fig. 3B). Only two phages, Rhodobacteraceae phage MD05 and Pink berry phage
143 MD08, are similarly abundant in each pink berry metagenome, while read mapping to the remaining
144 six phages suggests they are distributed unevenly between individual aggregates (Fig. 3). Additionally,

145 to observe whether the phage genomes of interest are present in pink berry metagenomes from
146 previous years, we mapped reads from a metagenome of 10 pink berries sampled in 2011 (Wilbanks *et*
147 *al*., 2014). This revealed near-complete coverage of Desulfofustis phage MD01 and Rhodobacteraceae
148 phage MD06 (Suppl. Fig. 2), indicating that these phages have persisted over seven years. In contrast,
149 read mapping to Desulfofustis phage MD02 and Thiohalocapsa phage MD04 genome largely occurs at
150 highly conserved regions, and is likely the consequence of non-specific cross-mapping (Suppl. Fig. 2).
151 Likewise, Thiohalocapsa phage MD03, Rhodobacteraceae phages MD05 and MD07, and Pink berry
152 phage MD00 had <1% or no genome coverage. Since neither this study nor the study from 2011
153 (Wilbanks *et al.,* 2014) are exhaustive surveys of pink berry diversity, it is difficult to determine the
154 mechanisms behind the emergence of these six phages. Taken together, these results suggest that
155 although pink berries have relatively simple and conserved bacterial community structures, their
156 phages are highly variable over both space and time.
157
158 ***Pink berry phages are targeted by bacterial CRISPR systems***
159 Bacterial CRISPR arrays serve as a record of past phage infection and can be used to infer
160 hosts for phage genomes (Childs *et al.*, 2014; England *et al.*, 2018). Two independent CRISPR spacer
161 prediction tools identified a total of 48 unique repeat sequences from four reference genomes for
162 known pink berry-associated bacteria: *Desulfofustis* sp. PB-SRB1 (GenBank: JAEQMT010000010.1),
163 *Flavobacteriales* bacterium (GenBank: DNTB01000031.1), *Rhodobacteraceae* sp. A2 (GenBank:
164 JAERIM010000001.1), and *Thiohalocapsa* sp. PB-PSB1 (GenBank: CP050890.1) (Wilbanks *et al.*,
165 2014). Parsing the remaining available pink berry genome, *Oceanicaulis alexandrii* sp. A1 (GenBank:
166 JAERIO010000015.1), did not yield any CRISPRs. Because CRISPR repeat sequences are conserved
167 within bacterial species (Lange *et al.*, 2013; Mojica *et al.*, 2000), they can be used to identify adjacent
168 spacer sequences in unassembled metagenomic short reads (England *et al.*, 2018; Skennerton *et al.*,

169 2013). Using the CRISPR repeat sequences from reference genomes, NARBL (England *et al.*, 2018)
170 identified 2,802 unique CRISPR spacer sequences from the set of merged metagenomic reads from
171 LS06-2018-s01, LS06-2018-s02, and LS06-2018-s03. Of these, 798 spacers were adjacent to repeats
172 associated with *Desulfofustis* sp. PB-SRB1, 71 were adjacent to *Flavobacteriales* repeats, 349 were
173 adjacent to *Rhodobacteraceae* sp. A2 repeats, and 1,584 unique spacers were adjacent to repeats
174 associated with *Thiohalocapsa* sp. PB-PSB1.

175 To look for evidence of previous phage-bacteria interactions, metagenomic CRISPR spacer
176 sequences were aligned to the eight complete phage genome assemblies (Fig. 1 & 4A). Of the 2,802
177 unique spacer sequences extracted from the merged set of metagenomic reads, 163 unique spacers
178 aligned to seven phage contigs of interest with at least 80% identity over the entire spacer length (Fig.
179 1 & 4A, Suppl. Data 3). Spacers from three out of the four potential host taxa aligned to phage
180 genomes, while no spacers from *Flavobacteriales* aligned to any phage genome of interest.
181 *Thiohalocapsa* sp. PB-PSB1 was predicted to be the host of Thiohalocapsa phages MD03 and MD04,
182 *Rhodobacteraceae* sp. A2 was predicted to be the host of Rhodobacteraceae phages MD05, MD06,
183 and MD07, and *Desulfofustis* sp. PB-SRB1 was predicted to be the host of Desulfofustis phages MD01
184 and MD02 (Fig. 4A, Suppl. Data 3). For the two most prevalent and abundant phages identified,
185 Desulfofustis phage MD02 (Suppl. Fig. 3A) and Thiohalocapsa phage MD04 (Suppl. Fig. 3B), phage
186 genome and targeting CRISPR spacer coverages were positively correlated. Moreover, although most
187 host spacers matched to a single virus, two spacers from the *Rhodobacteraceae* host aligned to
188 Thiohalocapsa phage MD04 and Desulfofustis phage MD02 (Fig. 4A). We do not predict these phages
189 to have infected the *Rhodobacteraceae* bacterium, since there was only one alignment each to
190 Thiohalocapsa phage MD04 and Desulfofustis phage MD02 compared to 126 and three alignments
191 from the two phage genomes to *Thiohalocapsa* and *Desulfofustis* spacers, respectively (Suppl. Data 3).
192 Additionally, the alignments from the two *Rhodobacteraceae* spacers to Desulfofustis phage MD02

193 were weaker than the alignments from the *Desulfofustis* spacers (Fig. 4A, Suppl. Data 3). Alignments
194 to multiple host taxa may be due to CRISPR systems targeting motifs that are present in several phage
195 lineages.

196 The vast majority of CRISPR spacer-protospacer matches occurred only once in the dataset.
197 However, four spacers from *Thiohalocapsa* aligned imperfectly to two distinct protospacers on the
198 genome of Thiohalocapsa phage MD04 (*Thiohalocapsa* sp. PB-PSB1 spacers 09-2, 21-2, 21-11, and
199 21-164 in Suppl. Data 3). An additional two spacers from *Desulfofustis* sp. PB-SRB1 are reverse
200 complements of each other and target the same protospacer sequence on both Desulfofustis phage
201 MD01 and MD02 (*Desulfofustis* sp. PB-SRB1 spacers 01-103 and 02-11) (Figs. 1, 4A, 4B, Suppl.
202 Data 3). The shared CRISPR targeting of Desulfofustis phage MD01 and MD02 occurred in a
203 conserved gene predicted to encode a phage capsid protein (Fig. 1, Suppl. Data 1) (ORFs
204 DPMD01_45 and DPMD02_11, respectively), and corresponding to genomic regions with high read
205 coverage relative to the remainder of the phage genomes (Fig. 1, Suppl. Fig. 2). A third spacer (01-64)
206 targets a distinct protospacer within this same capsid gene on Desulfofustis phage MD02. A third
207 capsid gene from an incomplete viral contig was found to be homologous to these two variants from
208 Desulfofustis phage MD01 and MD02 and is 92% identical and of similar length. The capsid gene
209 from this incomplete phage contig aligns with the same three spacers targeting Desulfofustis phage
210 MD01 and MD02 (Fig. 4B, Suppl. Data 4). Nucleotide variation at these protospacers inferred by read
211 mapping shows that other variants of these protospacers likely exist in related pink berry phages not
212 assembled here (Fig. 4B). We obtained metagenome-wide allelic variants spanning the entire capsid
213 gene of Desulfofustis phage MD01, Desulfofustis phage MD02, and Incomplete phage contig 1 and
214 observed a near three-fold increase in the number of variants over CRISPR-targeted regions (Suppl.
215 Fig. 4). Taken together, these results suggest that this conserved phage capsid gene is an active site of
216 diversification.

217

218 ***Horizontal gene transfer between pink berry phages and their hosts***

219 Desulfofustis phage MD02 and Thiohalocapsa phage MD04 were found to contain discrete
220 regions of high read coverage compared to the rest of the genome (Fig. 1, Suppl. Fig. 2). We
221 hypothesized that these regions may be the result of read mapping from homologous regions of
222 bacterial chromosomes or other phage genomes.

223 The high coverage region on Desulfofustis phage MD02 (genome coordinates 170-1760) did
224 not align to any region of the *Desulfofustis* sp. PB-SRB1 host reference genome (GenBank:
225 JAEQMT000000000.1) or to any other bacterial genomes in RefSeq. This region also did not
226 successfully align to any other contigs in the co-assembly, indicating that the coverage at this region is
227 not the result of conservation of this sequence among other members of the metagenome. Because of
228 the circularity of these genomes, it is possible that the high read coverage at these regions is attributed
229 to terminal redundancy from circular permutation (Garneau *et al.*, 2017; Grossi *et al.*, 1983).

230 The high coverage region of Thiohalocapsa phage MD04 (genome coordinates 21,035-22,077)
231 encodes a glucosaminidase domain-containing protein (ORF TPMD04_36) predicted to function as the
232 phage lysin. TPMD04_36 is homologous to two predicted ORFs (NCBI N838_07070 and
233 N838_07065) in the *Thiohalocapsa* sp. PB-PSB1 genome (GenBank CP050890.1), which are adjacent
234 to a predicted transposase (Fig. 5). This observation prompted an investigation into a possible
235 transposon-mediated HGT event between Thiohalocapsa phage MD04 and its host, *Thiohalocapsa* sp.
236 PB-PSB1.

237 Alignment of the Thiohalocapsa phage MD04 and *Thiohalocapsa* PB-PB1 genomes revealed
238 that the phage lysin gene and the host pseudogene are in frame with each other, except the host ORF
239 N838_07070 contains a single-nucleotide insertion at position 1,668,396 that results in a premature
240 stop codon (Fig. 5). After closer observation of the region surrounding the pseudogene on the

241 *Thiohalocapsa* sp. PB-PSB1 genome, we observed a transposon of the IS4 family with terminal
242 inverted repeats and numerous direct repeats indicative of past transposase activity (Suppl. Fig. 5). The
243 Thiohalocapsa phage MD04 genome contains an imperfect copy of a 17-bp direct repeat, differing by
244 only one nucleotide, at the N-terminal of the TPMD04_36 ORF. Though MD04 was frequently
245 targeted by host CRISPRs (including at protospacers directly adjacent to the lysin gene), no spacer-
246 protospacer alignments were observed within the lysin gene, likely the result of selection against
247 CRISPR self-targeting. Together, this finding suggests a past transposon-mediated HGT event may
248 have resulted in the transfer of the phage lysin gene from Thiohalocapsa phage MD04, or a related
249 ancestral phage, to its host.
250
251 **DISCUSSION**
252 Marine phages are the most numerous biological components of the global ocean,
253 outnumbering their bacterial hosts by tenfold (Breitbart *et al.*, 2018), and play vital ecosystem roles as
254 predators that turn over organic matter through bacterial lysis (Heldal & Bratbak, 1991; Maranger &
255 Bird, 1995; Proctor *et al.*, 1988; Steward *et al.*, 1996) and as agents of HGT impacting bacterial
256 community structure and function (Anantharaman *et al.*, 2014; Breitbart *et al.*, 2018; Kieft *et al.*, 2021;
257 Tuttle & Buchan, 2020). Pink berries are marine microbial aggregates with a microscale sulfur cycle,
258 and have been used as a model system to study cryptic biogeochemical cycling (Wilbanks *et al.*,
259 2014). Though phages have been identified within the pink berry metagenomes (Wilbanks *et al.*,
260 2022), the full diversity of phages and their impacts on pink berry communities remain largely
261 unexplored. Here, we investigated these simple, naturally-occurring microbial communities as a model
262 for phage-host co-evolution.
263 We co-assembled three pink berry samples, recovering eight complete phage genomes
264 spanning a total of 350 Kb and infecting three different bacterial species within the consortia,

265 *Desulfofustis* sp. PB-SBR1, Rhodobacteraceae sp. A2, and *Thiohalocapsa* sp. PSB1. We found that
266 pink berry-associated phages are highly diverse and largely novel, as seven of the eight complete
267 phage genomes analyzed fail to cluster with any known phage sequences. One pink berry phage,
268 Desulfofustis phage MD02, is only the second member of the genus *Knuthellervirus.* The other
269 member of the *Knuthellervirus*, Pseudomonas phage PMBT14, infects *Pseudomonas fluorescens*,
270 another marine organism, suggesting this phage genus infects diverse hosts.

271 Although the composition of the pink berry bacterial community is similar across individual
272 aggregates, we found that phage presence and abundance is highly heterogeneous across samples. Pink
273 berries are free-living microbial aggregates that exist at the sediment-water interface of intertidal
274 ponds, with no obvious physical barrier to phage entry into the system. This raises ecological
275 questions about the mechanisms underlying phage ingress into a pink berry aggregate and their
276 persistence within, loss, or exclusion from the community.

277 CRISPRs are a common phage-resistance system employed by bacteria and archaea
278 (Barrangou *et al.*, 2007; Horvath & Barrangou, 2010; Jansen *et al.*, 2002). We identified an astounding
279 2,731 unique CRISPR spacer sequences from pink berry-associated *Desulfofustis*, *Rhodobacteraceae*,
280 and *Thiohalocapsa* hosts, 163 of which (~6%) target a complete phage genome we assembled. This
281 discrepancy suggests that the true diversity of phages pink berry-associated bacteria encounter is far
282 greater than what we report here. Seven of the eight phages investigated were targeted by described
283 pink berry bacterial CRISPR systems, and the eighth, Pink berry phage MD08, may infect other pink
284 berry-associated taxa that do not encode CRISPR defenses or for which we do not yet have a high-
285 quality reference genome. Finally, we were able to observe diversification of a CRISPR-targeted
286 conserved capsid gene, which is inconsistent with diversification across the rest of the phage genomes.
287 Although we cannot establish a causative relationship between CRISPR targeting and phage variation,
288 these observations are consistent with a model of CRISPR-driven evolution causing positive selection

289 in a phage structural protein. We further observed a positive correlation between CRISPR spacer
290 abundance and target phage abundance for two of the most prevalent and abundant phages identified.
291 This suggests that these CRISPR spacers are positively selected for within individual pink berry
292 consortia, and is consistent with observations from diverse microbial systems (Somerville *et al.*, 2022;
293 Meaden *et al.*, 2021).

294 Phages and other mobile genetic elements are powerful mediators of HGT (Breitbart *et al.*,
295 2018; Hall *et al.*, 2017; Schneider, 2021). HGT between bacterial species within the pink berry
296 consortia has been previously reported (Wilbanks *et al.*, 2022), yet the role of phages in HGT and
297 bacterial genome evolution in this system remains to be explored. We identified a predicted phage
298 lysin gene that was horizontally transferred from Thiohalocapsa phage MD04 to its *Thiohalocapsa*
299 host likely via a transposon intermediary. It is unclear if the lysin gene encoded on the *Thiohalocapsa*
300 sp. PB-PSB1 genome is functional; a nonsense mutation in this ORF suggests it is a pseudogene, and
301 was perhaps selected for to avoid deleterious effects of expressing this potentially lethal protein.
302 Future work should aim to experimentally determine how these phages impact the evolution of both
303 individual bacterial hosts and entire pink berry aggregates through HGT.

304 Taken together, our results demonstrate that pink berry communities contain diverse and
305 variable phage consortia, which are highly targeted by host-encoded CRISPR systems. We leveraged
306 metagenomic sequence data to better understand phage-host co-evolution occurring through CRISPR
307 evasion and HGT. Pink berries offer a simple yet relatively unexplored, naturally-occurring model of
308 phage invasion into and exclusion from microbial communities. The potential roles of phages in pink
309 berry syntrophy and community-wide metabolic exchanges remain to be explored, but it is now clear
310 that phages are notable members of these microbial consortia.

311

312 **METHODS**

313 ***Sampling***

314 Pink berries, their surrounding sediment, and seawater were collected from pond LS06
315 (41.57587, -70.63781) in the Little Sippewissett Salt Marsh in Falmouth, MA, on July 17, 2018, using
316 sterile 50-mL conical tubes. Samples LS06-2018-s01 and LS06-2018-s02 each contained one pink
317 berry aggregate, and sample LS06-2018-s03 contained three pink berry aggregates. These samples
318 were transported to the lab and immediately processed for DNA extraction.

319

320 ***DNA isolation and sequencing***

321 Pink berry samples were each mechanically homogenized in 1 mL of TE buffer and
322 centrifuged at 1,000 x*g* for 1 min to pellet particulate matter. The supernatant was removed and
323 subjected to a Wizard Genomic DNA Purification Kit (Promega Catalog No. A1120) according to the
324 manufacturer's instructions. Purified DNA was fragmented, and sequencing adapters and barcodes
325 were ligated with the Nextera DNA Flex Library Prep Kit (Illumina Catalog No. 20018705) using
326 Nextera DNA CD indexes (Illumina Catalog No. 20018708). DNA yield was measured with a Qubit
327 High Sensitivity dsDNA assay kit (ThermoFisher Catalog No. Q32851), and DNA from LS06-2018-
328 s01, LS06-2018-s02, and LS06-2018-s03 were pooled at a ratio of 1:1:10. After pooling, DNA was
329 purified with Ampure XP beads (Beckman Coulter Catalog No. A63881) according to the
330 manufacturer's instructions. DNA sequencing was performed on an Illumina HiSeq 2500 using a
331 2x250nt protocol at the University of Illinois at Urbana-Champaign Roy J. Carver Biotechnology
332 Center.

333

334 ***Metagenome co-assembly & read mapping***

335 For each sample, reads were quality checked with FASTQC v0.11.9 (Andrews, 2010), trimmed
336 using Trimmomatic v0.39 (Bolger *et al.*, 2014), and adapter sequences were removed. Trimmomatic

filtered reads using a sliding window of 4 base pairs, a minimum average base quality score of 15, a minimum quality score for retention on the leading and trailing ends of 2, and a minimum read length of 100 bases. The resulting trimmed and filtered reads were merged into a single set of reads and co-assembled using Metaviral SPAdes v3.15.2 (Antipov *et al.*, 2020) with default parameters to maximize the recovery of complete, circular phage genomes. To estimate phage and bacterial abundances, Bowtie2 v2.4.5 (Langmead & Salzberg, 2012) was used to map reads from the set of merged reads or from each pink berry sample to assembled phage contigs and to representative host genomes from NCBI BioProject PRJNA684324: *Desulfofustis* sp. PB-SRB1 (GenBank: JAEQMT010000010.1), *Rhodobacteraceae* sp. A2 (GenBank: JAERIM010000001.1), and *Thiohalocapsa* sp. PB-PSB1 (GenBank: CP050890.1) (Wilbanks *et al*., 2022). Read mapping statistics were obtained from Bowtie2 alignments using samtools v1.15.1 (Danecek *et al.*, 2021).

Phage sequence identification, binning, and annotation

ViralVerify v1.1 (Antipov *et al.*, 2020) was used with default settings to categorize the metagenome contigs as putatively bacterial or viral. ViralComplete v1.1 (Antipov *et al.*, 2020) was used with default settings to identify viral contigs that represent complete phage genomes. To verify these predictions, VIBRANT v1.2.1 (Kieft *et al.*, 2020) was used with default settings on the same metagenome contigs. All resulting putative viral contigs that were estimated to be complete viral genomes by both prediction tools were targeted for downstream analyses and annotation. vConTACT v2.0 (Bin Jang *et al.*, 2019) was used with the RefSeq v211 Viral database (Brister *et al.*, 2015) to cluster the viral contigs of interest with existing phage genomes and to approximate phage taxonomy.

The viral contigs of interest were passed through Pharokka v1.0.1 (github.com/gbouras13/pharokka), using PHANOTATE v1.5.0 (McNair *et al.*, 2019) to predict genes and PHROGs v3 (Terzian *et al.*, 2021) to provide initial protein annotations. Protein functions were

361 also predicted using Phyre2 (Kelley *et al.*, 2015), BLASTp v2.11.0 (Altschul *et al.*, 1990) with the
362 NCBI RefSeq v211 virus amino acid database and the non-redundant amino acid database (Brister *et*
363 *al.*, 2015; O'Leary *et al.*, 2016), and HMMER v3.2.1 (Eddy, 2011) with Pfam-a v35.0 (Mistry *et al.*,
364 2021) and TIGRFAMs v15.0 (Li *et al.*, 2021). The resulting predictions from each method were
365 manually reviewed for each protein, and a consensus annotation was inferred (Suppl. Data 1). Clinker
366 v0.0.23 (Gilchrist & Chooi, 2021) was used to identify conserved genes among phage genomes, which
367 were then aligned with tBLASTx v2.11.0 (Camacho *et al.*, 2009) against all the contigs in the
368 metagenome co-assembly. Metagenomic reads from pink berries sampled in 2011 (SRA:
369 SRR13297012) were mapped to the viral contigs of interest with Bowtie2 using the same methods
370 described above.
371
372 ***CRISPR spacer-protospacer analysis and host prediction***
373 CRISPRclassify v1.1.0 (Nethery *et al.*, 2021) and MinCEd v0.4.2 (Bland *et al.*, 2007) were
374 used to identify CRISPR repeat sequences from representative genomes of pink berry taxa from NCBI
375 BioProject PRJNA684324: *Desulfofustis* sp. PB-SRB1 (GenBank: JAEQMT010000010.1),
376 *Oceanicaulis alexandrii* (GenBank: JAERIO000000000.1), *Rhodobacteraceae* sp. A2 (GenBank:
377 JAERIM010000001.1), *Thiohalocapsa* sp. PB-PSB1 (GenBank: CP050890.1) (Sayers *et al*., 2020;
378 Wilbanks *et al*., 2022), and *Flavobacteriales* bacterium (GenBank: DNTB01000031.1). The identified
379 repeats from each tool were combined and dereplicated to obtain a list of repeats found in the genomes
380 of pink berry taxa. Since CRISPR arrays are often misassembled with short-read data, putative spacer
381 sequences from the trimmed and filtered reads of each LS06-2018 metagenome were identified with
382 NARBL (England *et al.*, 2018) using the dereplicated set of repeats identified from the reference
383 genomes, an approximate repeat size of 36, and a minimum coverage of supporting neighbor spacers
384 of 2. Since repeat sequences are highly conserved between bacterial species (Kunin *et al.*, 2007), any

385 spacer identified by NARBL was inferred to belong to the same species as the reference genome from
386 which its associated repeat came. The resulting spacers were aligned to the viral contigs of interest
387 using Spacerblast v0.7.7 (Collins & Whitaker, 2022). Viral contigs that aligned to spacers with at least
388 80% identity over the full length of the spacer were considered to have a host match with the
389 bacterium whose genome contained the spacer. The merged set of filtered and trimmed reads were
390 aligned to spacers identified by NARBL with Bowtie2 and read coverage statistics were obtained with
391 samtools as described above.

392 Metagenome reads were mapped to the regions containing conserved phage genes identified
393 with Clinker or tBLASTx, above, with spacer-protospacer alignments with Bowtie2. Mapped reads
394 were converted to multiple sequence alignments using SAM4WebLogo in JVarkit v2021.10.13
395 (Lindenbaum, 2015) and sequence logos were visualized with WebLogo (Crooks *et al.*, 2004). Allele
396 variants for conserved phage genes were called by using Snippy v3.2 (github.com/tseemann/snippy)
397 with default settings on metagenome reads. Variants that resulted before the "--mincov" and "--
398 minfrac" filters were applied were used downstream to maximize the number of possible variants
399 recovered. Variant statistics were obtained and visualized using vcfR v1.13.0 (Knaus & Grünwald,
400 2017) with default settings, except 100-bp size windows were used instead of 1000-bp.

401

402 ***Horizontal gene transfer analysis***

403 Uneven sequence coverage patterns on phage genomes are sometimes attributed to HGT
404 between the phage and its host genome, especially if the region aligns to the host genome and/or
405 contains genes that facilitate HGT (Kleiner *et al.*, 2020). Viral genomes of interest and their coverages
406 from the merged set of metagenome reads were visualized in IGV v2.11.4 (Thorvaldsdóttir *et al.*,
407 2013). For any discrete regions of the viral genomes of interest that had much higher read coverage
408 (>3x) than the surrounding region, those regions were aligned to their predicted host genomes, if

available, using BLASTn v2.11.0 (Camacho *et al.*, 2009). Predicted regions of HGT between phages and their hosts were aligned with Mauve v2015-02-13 (Darling *et al.*, 2004). Conserved repeats flanking a putative transposon were initially identified by reciprocal BLASTn using "blastn-short". The transposon region with its inverted repeats in the *Thiohalocapsa* sp. PB-PSB1 genome was identified with ISEScan v1.7.2.3 (Xie & Tang, 2017), and other repeats were identified and annotated manually in Geneious Prime v2022.1.1 (www.geneious.com/prime).

Data availability

DNA sequencing reads from this study are deposited in the NCBI SRA under PRJNA907316. Assembled phage genomes are deposited in NCBI GenBank under the following accession numbers: OP947158.1 (Desulfofustis phage MD01), OP947159.1 (Desulfofustis phage MD02), OP947165.1 (Thiohalocapsa phage MD03), OP947166.1 (Thiohalocapsa phage MD04), OP947161.1 (Rhodobacteraceae phage MD05), OP947162.1 (Rhodobacteraceae phage MD06), OP947163.1 (Rhodobacteraceae phage MD07), OP47164.1 Pink berry phage MD08, OP947160.1 (Pink berry virus MD00).

ACKNOWLEDGMENTS

This research was conducted as a part of the 2020 Microbial Diversity course at the Marine Biological Laboratory (MBL), which was funded by NSF Award ID 1822263, DOE Award ID DE-SC0016127, and the Simons Foundation. We gratefully acknowledge the insights and support provided by all participants and instructors, especially Whitney England, Laura Suttenfield, and George O'Toole. D.E.C. was supported by NIH T32 DK077653-29 and Crohn's & Colitis Foundation Research Fellowship Award #935619. E.G.W. gratefully acknowledges support from the Whitman Fellowship at the MBL. This research was sponsored by the U.S. Army Research Office and

accomplished under cooperative agreement W911NF-19-2-0026 for the Institute for Collaborative Biotechnologies.

Conceptualization, J.C.K., D.E.C., E.G.W., and R.J.W.; Data Curation, J.C.K., D.E.C. and R.J.W.; Analysis, J.C.K. and D.E.C.; Funding Acquisition, R.J.W.; Investigation, J.C.K. and D.E.C.; Methodology, J.C.K., D.E.C., E.G.W., and R.J.W.; Project Administration, J.C.K., D.E.C., E.G.W., and R.J.W.; Resources, R.J.W.; Software, J.C.K. and D.E.C.; Supervision, D.E.C., E.G.W., R.J.W.; Validation, J.C.K. and D.E.C.; Visualization, J.C.K. and D.E.C.; Writing—Original Draft, J.C.K. and D.E.C.; Writing—Review and Editing, J.C.K., D.E.C., E.G.W., and R.J.W.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. J*ournal of molecular biology, 215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M., & Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science*, *344*(6185), 757-760.

Andersson, A. F., & Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, 320(5879), 1047-1050.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2020). Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics (Oxford, England)*, *36*(14), 4126–4129. https://doi.org/10.1093/bioinformatics/btaa490.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., ... & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), 1709-1712.

Bergh, Ø., Børsheim, K. Y., Bratbak, G., & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467-468.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D., & Sullivan, M. B. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology*, *37*(6), 632–639. https://doi.org/10.1038/s41587-019-0100-8

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition Tool (CRT): A tool for automatic detection of Clustered Regularly Interspaced Palindromic repeats. *BMC Bioinformatics, 8*(1). https://doi.org/10.1186/1471-2105-8-209

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer FOR Illumina sequence data. *Bioinformatics, 30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. *Nature Microbiology* 3, 754–766. https://doi.org/10.1038/s41564-018-0166-y

Breitbart, M., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Oceanography (Wash DC)* 20, 135–139.

Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI viral genomes resource. *Nucleic Acids Research, 43*. https://doi.org/10.1093/nar/gku1207

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*(421), https://doi.org/10.1186/1471-2105-10-421

Campbell, D. E., Ly, L. K., Ridlon, J. M., Hsiao, A., Whitaker, R. J., & Degnan, P. H. (2020). Infection with Bacteroides phage BV01 alters the host transcriptome and bile acid metabolism in a common human gut microbe. *Cell reports, 32*(11), 108142.

Childs, L. M., England, W. E., Young, M. J., Weitz, J. S., & Whitaker, R. J. (2014). CRISPR-Induced Distributed Immunity in Microbial Populations. *PLoS ONE*, *9*(7). https://doi.org/10.1371/journal.pone.0101710

Collins, A. J., & Whitaker, R. J. (2022). CRISPR Comparison Toolkit (CCTK): Rapid Identification, Visualization, and Analysis of CRISPR Array Diversity. *BioRxiv* 2022.07.31.502198. Advance online publication. https://doi.org/10.1101/2022.07.31.502198

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome research, 14(6), 1188–1190. https://doi.org/10.1101/gr.849004

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of samtools and BCFtools. GigaScience, 10(2). https://doi.org/10.1093/gigascience/giab008

Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, *14*(7), 1394–1403. https://doi.org/10.1101/gr.2289704

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., & Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology*, 190(4), 1390-1400.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10). https://doi.org/10.1371/journal.pcbi.1002195

England, W. E., Kim, T., & Whitaker, R. J. (2018). Metapopulation Structure of CRISPR-Cas Immunity in Pseudomonas aeruginosa and Its Viruses. *MSystems*, *3*(5). https://doi.org/10.1128/msystems.00075-18

Garneau, J. R., Depardieu, F., Fortier, L. C., Bikard, D., & Monot, M. (2017). PhageTerm: a tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. *Scientific reports*, *7*(1), 8292. https://doi.org/10.1038/s41598-017-07910-5

Gilchrist, C. L., & Chooi, Y. H. (2021). Clinker & clustermap. js: Automatic generation of gene cluster comparison figures. *Bioinformatics, 37*(16), 2473-2475.

Grossi, G. F., Macchiato, M. F., & Gialanella, G. (1983). Circular permutation analysis of phage T4 DNA by electron microscopy. *Zeitschrift fur Naturforschung. Section C, Biosciences*, *38*(3-4), 294–296. https://doi.org/10.1515/znc-1983-3-422

Hall, J. P., Brockhurst, M. A., & Harrison, E. (2017). Sampling the mobile gene pool: Innovation via horizontal gene transfer in bacteria. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1735), 20160424. https://doi.org/10.1098/rstb.2016.0424

Hampton, H. G., Watson, B. N., & Fineran, P. C. (2020). The arms race between bacteria and their phage foes. *Nature*, *577*(7790), 327–336. https://doi.org/10.1038/s41586-019-1894-8

Heldal, M., & Bratbak, G. (1991). Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog. Ser*, *72*(3), 205-212.

Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science, 327*(5962), 167-170.

523 Iida, S., Streiff, M. B., Bickle, T. A., & Arber, W. (1987). Two DNA antirestriction systems of
524 bacteriophage P1, darA, and darB: characterization of darA phages. *Virology*, *157*(1), 156-
525 166.
526 Iyer, L. M., Burroughs, A. M., Anand, S., de Souza, R. F., & Aravind, L. (2017). Polyvalent proteins,
527 a pervasive theme in the intergenomic biological conflicts of bacteriophages and conjugative
528 elements. *Journal of bacteriology*, *199*(15), e00245-17.
529 Jansen, R., Embden, J. D. V., Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are
530 associated with DNA repeats in prokaryotes. *Molecular microbiology*, *43*(6), 1565-1575.
531 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A
532 programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *science*,
533 *337*(6096), 816-821.
534 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web
535 portal for protein modeling, prediction and analysis. Nature Protocols 10, 845-858.
536 Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and
537 curation of microbial viruses, and evaluation of viral community function from genomic
538 sequences. *Microbiome*, *8*(1), 1-23.
539 Kieft, K., Zhou, Z., Anderson, R. E., Buchan, A., Campbell, B. J., Hallam, S. J., Hess, M., Sullivan,
540 M. B., Walsh, D. A., Roux, S., & Anantharaman, K. (2021). Ecology of inorganic sulfur
541 auxiliary metabolism in widespread bacteriophages. *Nature Communications, 12*(1).
542 https://doi.org/10.1038/s41467-021-23698-5
543 Kleiner, M., Bushnell, B., Sanderson, K. E., Hooper, L. V., & Duerkop, B. A. (2020). Transductomics:
544 sequencing-based detection and analysis of transduced DNA in pure cultures and microbial
545 communities. *Microbiome*, *8*(1), 1-17.

Koskella, B., & Brockhurst, M. A. (2014). Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, *38*(5), 916–931. https://doi.org/10.1111/1574-6976.12072

Kunin, V., Sorek, R., & Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome biology*, *8*(4), 1-7.

Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S., & Backofen, R. (2013). CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic acids research*, *41*(17), 8034-8044.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., Gonzales, N. R., Gwadz, M., Lanczycki, C. J., Song, J. S., Thanki, N., Wang, J., Yamashita, R. A., Yang, M., Zheng, C., … Thibaud-Nissen, F. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic acids research*, *49*(D1), D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.

Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics. *figshare*. Journal contribution. https://doi.org/10.6084/m9.figshare.1425030.v1

Maranger, R., & Bird, D. F. (1995). Viral abundance in aquatic systems: a comparison between marine and fresh waters. *Marine Ecology Progress Series*, *121*, 217-226.

Martiny, J. B., Riemann, L., Marston, M. F., & Middelboe, M. (2014). Antagonistic coevolution of marine planktonic viruses and their hosts. *Annual review of marine science*, *6*, 393-414.

McNair, K., Zhou, C., Dinsdale, E. A., Souza, B., & Edwards, R. A. (2019). PHANOTATE: A novel approach to gene identification in phage genomes. *Bioinformatics*, *35*(22), 4537–4542. https://doi.org/10.1093/bioinformatics/btz265

Meaden, S., Biswas, A., Arkhipova, K., Morales, S., Dutilh, B., Westra, E., Fineran, P. (2021). High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. *Current bioloigy, 32*(1), P220-227.E5. https://doi.org/10.1016/j.cub.2021.10.038

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E., Tosatto, S., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mojica, F. J., Díez-Villaseñor, C., Soria, E., & Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular microbiology*, *36*(1), 244-246.

Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, *155*(3), 733-740.

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular ecology resources*, *17*(1), 44-53.

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2020). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nature Biotechnology, 39(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nethery, M. A., Korvink, M., Makarova, K. S., Wolf, Y. I., Koonin, E. V., & Barrangou, R. (2021). CRISPRclassify: Repeat-Based Classification of CRISPR Loci. *The CRISPR journal*, *4*(4), 558–574. https://doi.org/10.1089/crispr.2021.0021

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Proctor, L. M., Fuhrman, J. A., & Ledbetter, M. C. (1988). Marine bacteriophages and bacterial mortality. *Eos*, *69*, 1111-1112.

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., & Karsch-Mizrachi, I. (2020). GenBank. *Nucleic Acids Research*.

Schneider, C. L. (2021). Bacteriophage-mediated horizontal gene transfer: transduction. *Bacteriophages: Biology, Technology, Therapy*, 151-192.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, *18*(20), 6097-6100. https://doi.org/10.1093/nar/18.20.6097

Seitz, A. P., Nielsen, T. H., & Overmann, J. (1993). Physiology of purple sulfur bacteria forming macroscopic aggregates in Great Sippewissett Salt Marsh, Massachusetts. *FEMS microbiology ecology*, *12*(4), 225-235.

Skennerton, C. T., Imelfort, M., & Tyson, G. W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic acids research, 41*(10), e105-e105.

Somerville, V., Schowing, T., Chabas, H., Schmidt, R., Ah, U., Bruggmann, R., Engel, P. (2022). Extensive diversity and rapid turnover of phage defense repertoires in cheese-associated bacterial communities. *Microbiome, 10*(1), 137. https://doi.org/10.1186/s40168-022-01328-6

Steward, G. F., Smith, D. C., & Azam, F. (1996). Abundance and production of bacteria and viruses in the Bering and Chukchi Seas. *Marine Ecology Progress Series*, *131*, 287-300.

Su, G., Morris, J. H., Demchak, B., & Bader, G. D. (2014). Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics*, *47*(1), 8-13.

Sun, C. L., Barrangou, R., Thomas, B. C., Horvath, P., Fremaux, C., & Banfield, J. F. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. *Environmental microbiology*, 15(2), 463-470.

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M. A., & Enault, F. (2021). PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics*, *3*(3).

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, *14*(2), 178-192.

Turner, D., Kropinski, A. M., & Adriaenssens, E. M. (2021). A Roadmap for Genome-Based Phage Taxonomy. *Viruses*, *13*(3), 506. https://doi.org/10.3390/v13030506

Tuttle, M. J., & Buchan, A. (2020). Lysogeny in the oceans: lessons from cultivated model systems and a reanalysis of its prevalence. *Environmental microbiology*, *22*(12), 4919-4933.

Wilbanks, E. G., Doré, H., Ashby, M. H., Heiner, C., Roberts, R. J., & Eisen, J. A. (2022). Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *The ISME Journal*, 1-11.

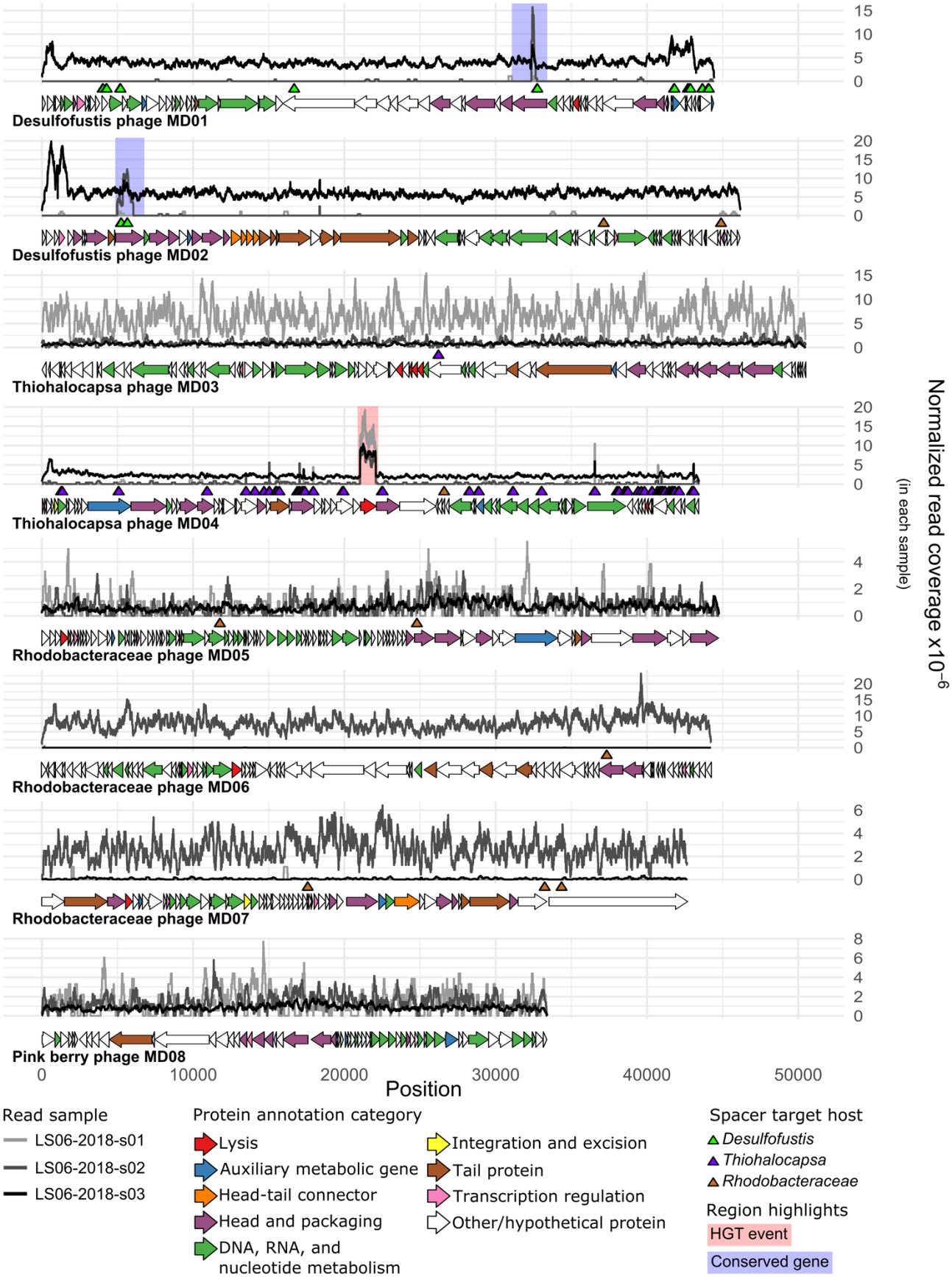
Wilbanks, E. G., Jaekel, U., Salman, V., Humphrey, P. T., Eisen, J. A., Facciotti, M. T., Buckley, D. H., Zinder, S. H., Druschel, G. K., Fike, D. A., & Orphan, V. J. (2014). Microscale sulfur cycling in the phototrophic pink berry consortia of the Sippewissett Salt Marsh. *Environmental Microbiology, 16*(11), 3398–3415. https://doi.org/10.1111/1462-2920.12388

Xie, Z., & Tang, H. (2017). ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*, *33*(21), 3340-3347.

641 **Table 1. Summary of pink berry metagenome sequencing and co-assembly.**

	LS06-2018-s01	LS06-2018-s02	LS06-2018-s03
Total read length (Gb)	0.41	1.07	7.19
Quality trimmed total read length (Gb)	0.35	0.91	6.94
Perc. of total read length after quality trimming	87%	85%	97%
		Co-assembly	
Contigs		184	
N50 (bp)		50,490	
L50		23	
Avg. read coverage (±SD)		12.89 ± 39.63	

642


Desulfofustis phage MD01
Desulfofustis phage MD02
Thiohalocapsa phage MD03
Thiohalocapsa phage MD04
Rhodobacteraceae phage MD05
Rhodobacteraceae phage MD06
Rhodobacteraceae phage MD07
Pink berry phage MD08
Normalized read coverage x10^{-6}
(in each sample)
Position
0
10000
20000
30000
40000
50000
Read sample
LS06-2018-s01
LS06-2018-s02
LS06-2018-s03
Protein annotation category
Lysis
Auxiliary metabolic gene
Head-tail connector
Head and packaging
DNA, RNA, and nucleotide metabolism
Integration and excision
Tail protein
Transcription regulation
Other/hypothetical protein
Spacer target host
Desulfofustis
Thiohalocapsa
Rhodobacteraceae
Region highlights
HGT event
Conserved gene

643

644 **Figure 1. Complete phage genomes vary in abundance across samples and are targeted by**
645 **bacterial CRISPR spacers.** Normalized read coverage by position for each sample are given.
646 Coverage values were normalized to the total number of trimmed and filtered reads for each sample.
647 Horizontal arrows indicate ORFs predicted by PHANOTATE (McNair *et al.*, 2019), and their colors
648 correspond to predicted functional categories. Triangles indicate genome positions of protospacers,
649 colored by host taxonomy of the corresponding spacer. Regions highlighted with a blue background
650 indicate a conserved gene between phage genomes inferred by Clinker (Gilchrist & Chooi, 2021).
651 Region highlighted with a red background was found to be an HGT event between the phage and host.

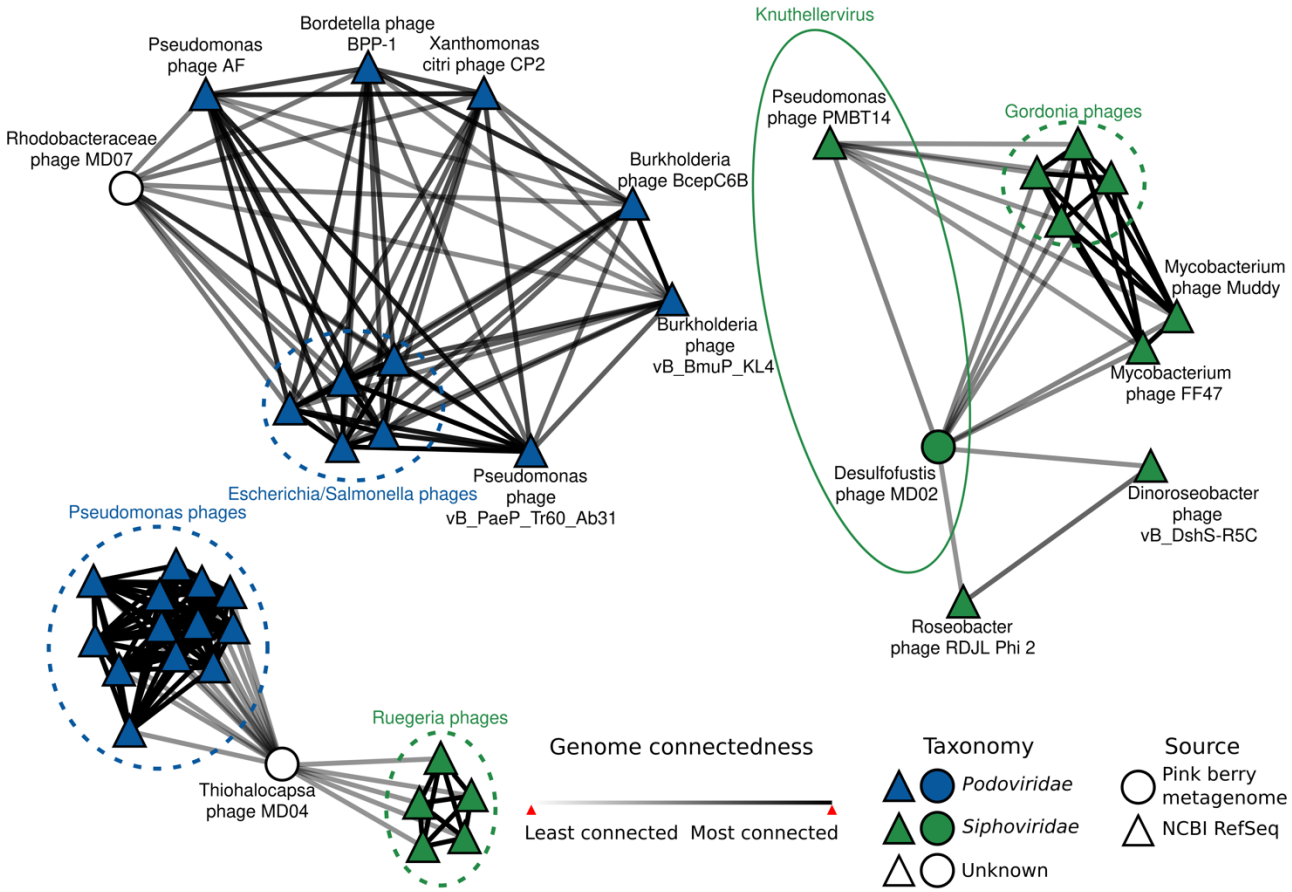


Figure 2. Pink berry consortia host largely novel phages. Each node represents a phage genome and edges are genome relatedness inferred by vConTACT (Bin Jang *et al.*, 2019). Reference genomes are from the RefSeq Prokaryotic virus database v211 (Brister *et al.*, 2015). The group circled with a solid line shows the phage genus of interest, *Knuthellervirus*. Groups circled in dashed lines show related phages infecting the same host. The network was visualized in Cytoscape v3.9.0 (Su *et al.*, 2014). Only phages which are first-neighbors to pink berry phage genomes are shown.

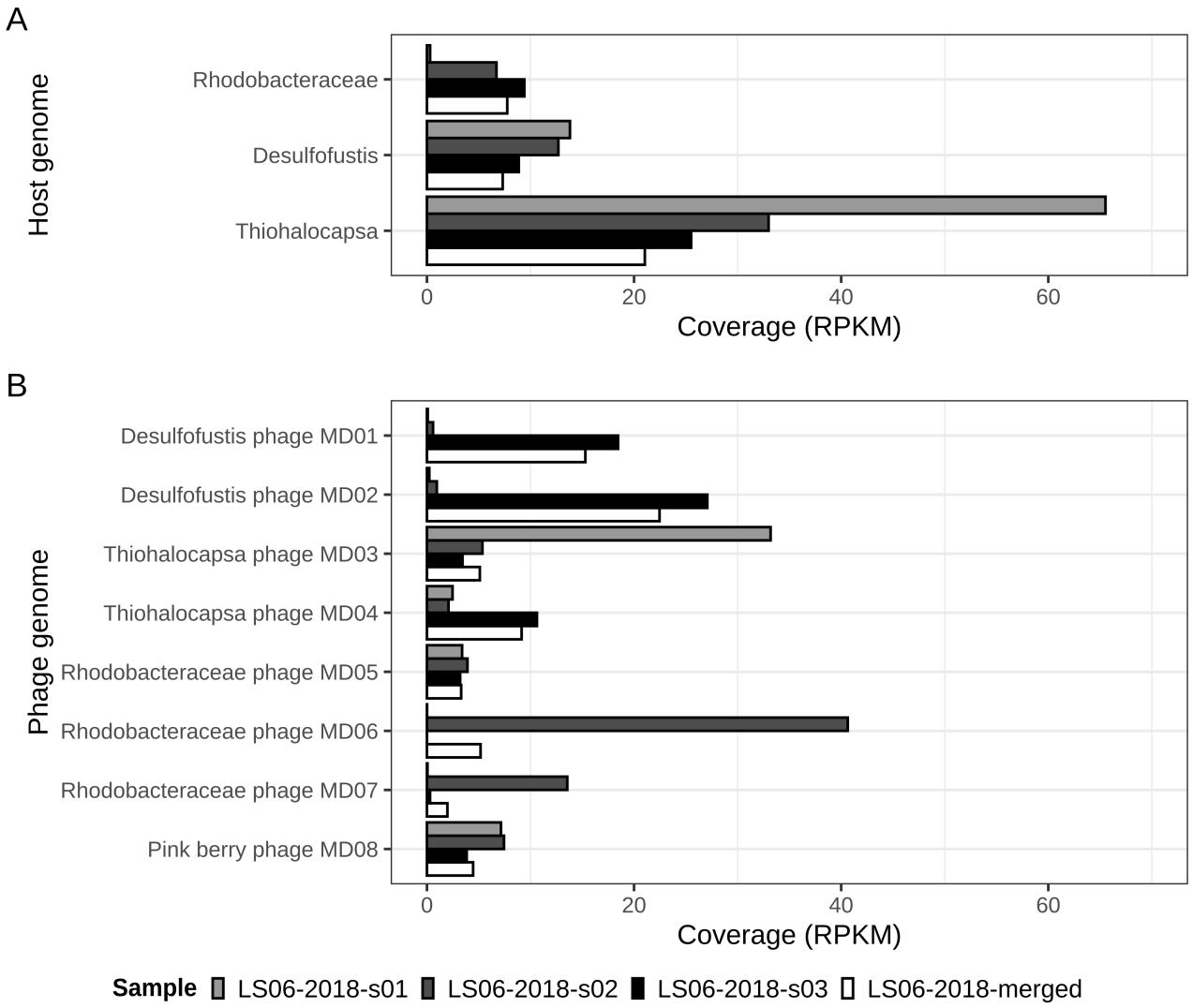


Figure 3. Phage presence and abundance are highly variable between pink berry communities. Average read coverages for host (A) and phage (B) genomes were converted to reads per kilobase million (RPKM) using the total number of filtered and trimmed reads per sample.

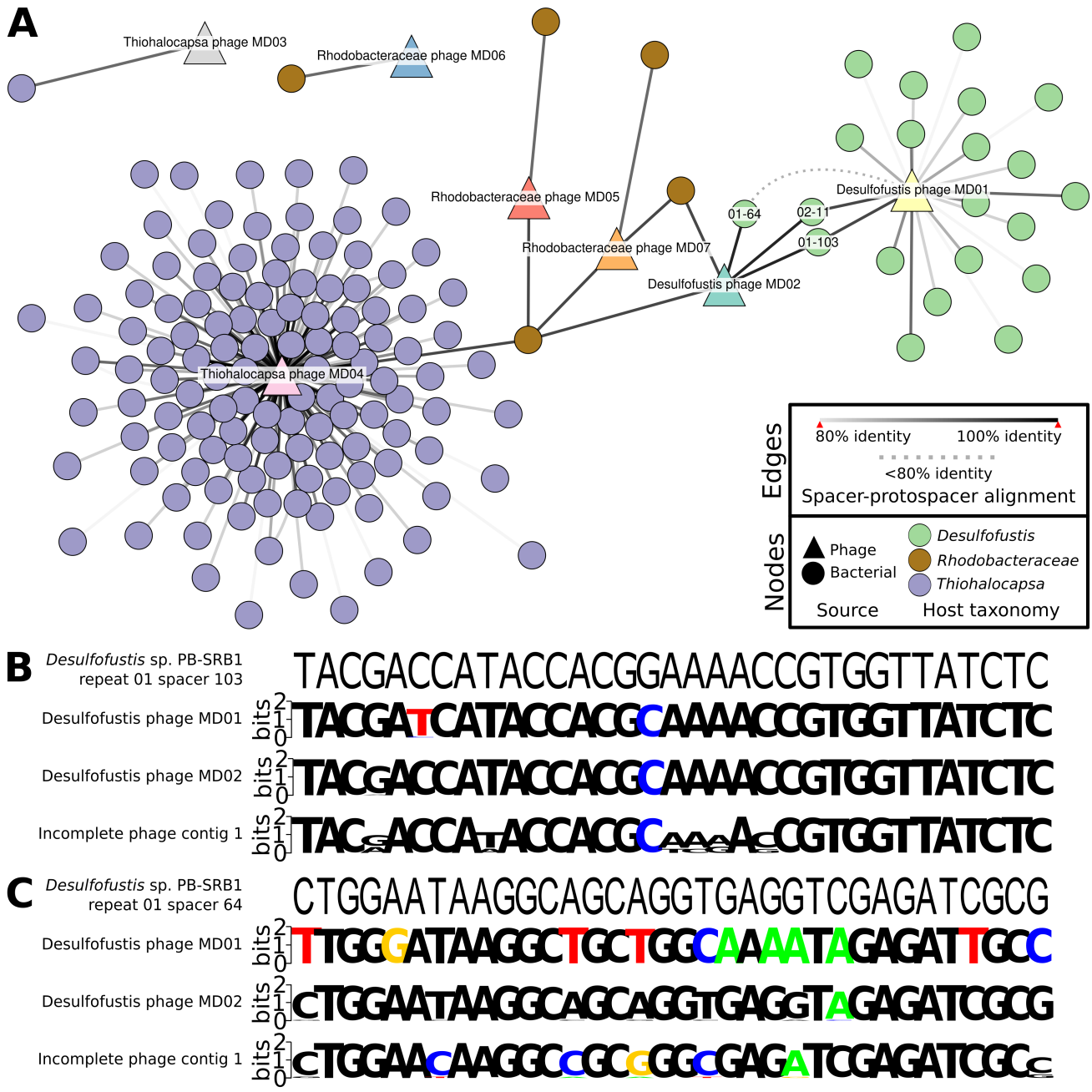


Figure 4. CRISPR spacer-phage genome alignments reveal hosts and a conserved protospacer. (A) Spacerblast (Collins & Whitaker, 2022) alignment results with at least 80% nucleotide identity over the entire spacer length were visualized in Cytoscape v3.9.0 (Su *et al.*, 2014). Circular nodes represent unique spacers from bacterial contigs and are colored by taxonomy. Triangular nodes are phage contigs. Solid edges represent percent nucleotide identity over the entire spacer length. The dashed edge shows the connection between Desulfofustis phage MD01 and *Desulfofustis* sp. PB-SRB1

670 repeat 01 spacer 64, which is below 80% identity and included in part (C). The nucleotide sequences
671 of phage protospacers within conserved capsid genes that were identified on phage contigs were
672 aligned to (B) *Desulfofustis* sp. PB-SRB1 repeat 01 spacer 103 and (C) *Desulfofustis* sp. PB-SRB1
673 repeat 01 spacer 64. Spacer 02-11 is the reverse complement of 01-103 and is not shown. The resulting
674 metagenome-wide variation in protospacer sequences from mapping reads to protospacers are shown
675 as sequence logos (Crooks *et al.*, 2004; Schneider & Stephens, 1990).

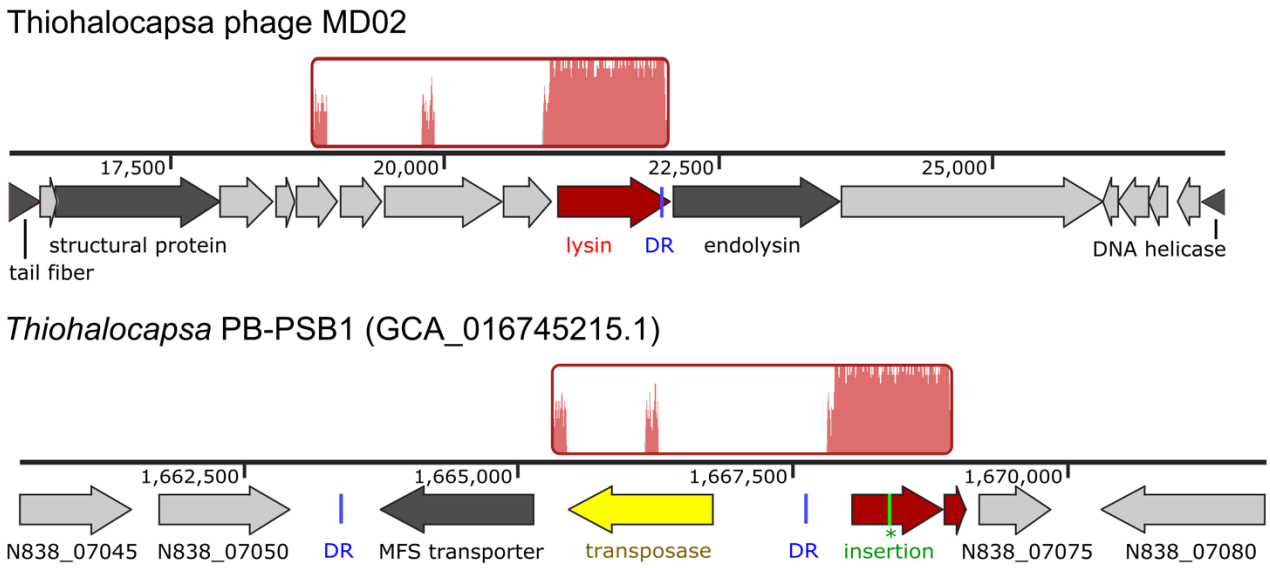


Figure 5. Region of homology between Thiohalocapsa phage MD04 and its host. Locally colinear blocks aligned with Mauve (Darling *et al.*, 2004) are shown in red, with traces inside representing nucleotide similarity. Genome tracks show genome coordinates and ORFs. Conserved 17-bp direct repeat sequences (DR) are shown in blue. A single-nucleotide insertion (green) in the *Thiohalocapsa* ORF N838_07070 results in a premature stop codon and a pseudogene annotation.