Genomes of marine cyanopodoviruses reveal multiple origins of diversity

Labrie S.J.[1], K. Frois-Moniz[1], M.S. Osburne[1], L. Kelly[1], S.E. Roggensack[1], M.B. Sullivan[1,3], G. Gearin[2], Q. Zeng[2], M. Fitzgerald[2], M.R. Henn[2] and S.W. Chisholm[1]

[1] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.

[2] Broad Institute, Cambridge, MA, USA.

[3] Current Address: Ecology and Evolutionary Biology Department, University of Arizona, Tucson, AZ, USA

Corresponding author: Sallie W. Chisholm

MIT 48-419

15 Vassar Street

Cambridge, MA 02139

Phone: (617) 253-1771

Fax: (617) 324-0336

Email: Chisholm@mit.edu

Running title: Cyanopodovirus genomics

Summary

The marine cyanobacteria *Prochlorococcus* and *Synechococcus* are highly abundant in the global oceans, as are the cyanophage with which they co-evolve. While genomic analyses have been relatively extensive for cyanomyoviruses, only 3 cyanopodoviruses isolated on marine cyanobacteria have been sequenced. Here we present 9 new cyanopodovirus genomes, and analyze them in the context of the broader group. The genomes range from 42.2 to 47.7 kbp, with G+C contents consistent with those of their hosts. They share 12 core genes, and the pan-genome is not close to being fully sampled. The genomes contain 3 variable island regions, with the most hypervariable genes concentrated at one end of the genome. Concatenated core-gene phylogeny clusters all but one of the phage into three distinct groups (MPP-A and two discrete clades within MPP-B). The outlier, P-RSP2, has the smallest genome and lacks RNA polymerase, a hallmark of the *Autographivirinae* subfamily. The phage in groups MPP-B contain photosynthesis and carbon metabolism associated genes, while group MPP-A and the outlier P-RSP2 do not, suggesting different constraints on their lytic cycles. Four of the phage encode integrases and three have a host integration signature. Metagenomic analyses reveal that cyanopodoviruses may be more abundant in the oceans than previously thought.

Introduction

Viruses are abundant in the oceans, often outnumbering bacterioplankton by an order of magnitude (Bergh et al., 1989; Fuhrman, 1999; Wommack and Colwell, 2000; Weinbauer and Rassoulzadegan, 2004). Among marine bacteria, the cyanobacteria *Prochlorococcus* and *Synechococcus* are the numerically dominant oxygenic phototrophs (Waterbury et al., 1986; Partensky et al., 1999; Scanlan and West, 2002), and contribute significantly to global primary productivity and global biogeochemical cycles (Liu et al., 1997; Liu et al., 1998). They coexist with their specific viruses – cyanophage – which are believed to play a key role in maintaining

diversity by "killing the winner" (Waterbury and Valois, 1993; Suttle and Chan, 1994; Thingstad, 2000). Moreover, cyanophage impact the evolution of their hosts by mediating horizontal gene transfer (Lindell et al., 2004; Zeidner et al., 2005; Sullivan et al., 2006; Yerrapragada et al., 2009).

All cyanophage isolated thus far are Caudovirales – tailed, dsDNA viruses belonging to three families: *Myoviridae*, *Podoviridae* and *Siphoviridae.* Most of the cyanomyoviruses are similar to the archetypal coliphage T4, and have genome sizes ranging from 161 – 252 kb, (Sullivan et al., 2010). Cyanopodoviruses, with genome sizes ranging from 42 kb to 47 kb, are similar in gene content and genome organization to coliphage T7 (Chen and Lu, 2002; Sullivan et al., 2005; Pope et al., 2007). There are fewer examples of cyanosiphoviruses (Sullivan et al., 2009; Huang et al., 2011), which have genome sizes ranging from 30 kb to 108 kb and do not share common features with other bacteriophage (Huang et al., 2011). To date, 18 cyanomyovirus genomes (Sullivan et al., 2005; Weigele et al., 2007; Millard et al., 2009; Sullivan et al., 2010; Sabehi et al., 2012), 5 cyanosiphovirus genomes (Sullivan et al., 2009; Huang et al., 2011), and 5 cyanopodovirus genomes have been published (Chen and Lu, 2002; Sullivan et al., 2005; Pope et al., 2007; Liu et al., 2007; Liu et al., 2008).

A hallmark characteristic of the cyanomyoviruses and cyanopodoviruses is that they carry homologs to host genes (which we now refer to as phage/host shared genes (Kelly et al, submitted)) whose products are thought to increase phage fitness under certain conditions. A subclass of these genes, referred to as auxiliary metabolic genes ("AMG" (Breitbart et al., 2007)), encode proteins involved in host metabolic pathways such as the light reactions of photosynthesis (PsbA, PsbD, Hli, PsaA, B, C, D, E, K, J/F (Mann, 2003; Lindell et al., 2004; Lindell et al., 2005; Sullivan et al., 2006; Sharon et al., 2009; Béjà et al., 2012)), the pentose phosphate pathway (PPP (Sullivan et al., 2005; Thompson et al., 2011; Zeng and Chisholm, 2012)), phosphate acquisition

74 (Millard et al., 2004; Sullivan et al., 2005; Sullivan et al., 2010; Thompson et al., 2011; Zeng and
75 Chisholm, 2012), nitrogen metabolism (Sullivan et al., 2010) and DNA synthesis (Sullivan et al.,
76 2005), among others. It is thought that the phage carry these homologs to alleviate bottlenecks in
77 these key pathways after host transcription of host homologs has stopped (Thompson et al.,
78 2011).
79
80 Several observations reveal very tight co-evolution of host and cyanophage genomes with regard
81 to these phage/host shared genes. It has been demonstrated, for example, that phage AMGs are
82 expressed simultaneously during infection (Lindell et al., 2007) regardless of their position in the
83 genome, which is striking given the strict genome-order transcription normally associated with
84 such (T7-like) phage. In the case of phage/host shared P-acquisition genes, it has been
85 demonstrated that these genes are carried more frequently by phage in regions of the oceans
86 where cells are P-stressed (Kelly et al., submitted), and expression of the phage version of a high-
87 affinity PO_4-transport protein is actually regulated by the host PhoRB two component regulatory
88 system such that the phage gene is only upregulated when the phage is infecting a P-stressed host
89 cell (Zeng and Chisholm, 2012).
90
91 Phage also carry genes that in the host encode high-light inducible proteins (Hlis – also called
92 small CAB-like proteins (Funk and Vermaas, 1999)) thought to protect the photosynthetic
93 complex, or possibly to be involved in a more general stress response in the host (He et al., 2001).
94 Photosynthesis-associated proteins (Hlis, PsbA and PsbD) found in cyanophage are related to
95 their respective orthologous proteins found in cyanobacterial genomes, indicating that they are of
96 cyanobacterial origin (Lindell et al., 2004; Sullivan et al., 2006). Interestingly, there are two types
97 of *hli* genes found in cyanobacterial genomes, referred to as single- and multi-copy *hlis* (Bhaya et
98 al., 2002). The single-copy *hlis* are part of the *Prochlorococcus* core genome while multi-copy
99 *hlis* contribute to the flexible genome and are found in highly variable genomic islands (Coleman

et al., 2006). Cyanophage *hlis* are homologous to the multi-copy *hlis*, suggesting that cyanophage play a role in horizontal transfer of multi-copy *hlis* (Lindell et al., 2004).

Of the 5 cyanopodoviruses for which complete genomes were available prior to this study, two are of marine origin: P-SSP7 and Syn5 from the Sargasso Sea (Sullivan et al., 2005; Pope et al., 2007), and one is of estuarine environment in Georgia (P60 -Chen and Lu, 2002). The two other isolates, Pf-WMP3 (Liu et al., 2008) and Pf-WMP4 (Liu et al., 2007), are derived from freshwater environment and were isolated on the filamentous cyanobacterium *Leptolyngbya.* The genome of P-SSP7 is organized in three classes, similar to coliphage T7 (Sullivan et al., 2005), the first involved in takeover of host enzymatic machinery, followed by DNA replication and transcription, and finally viral assembly and morphogenesis (Lindell et al., 2007). Interestingly, whereas P60, isolated from a coastal river (Chen and Lu, 2002), has a similar genetic architecture to the freshwater cyanophage, its genes have greater homology to marine cyanopodoviruses (See Note added in proofs).

To expand our understanding of the diversity and evolution of cyanopodoviruses infecting marine cyanobacteria, and to provide more reference genomes for metagenomic analyses, we sequenced 9 additional cyanopodovirus genomes (Table 1) isolated from diverse environments (Red Sea, Sargasso Sea, Gulf Stream, and Subtropical Pacific Gyre) on host strains belonging to four different ecotypes of *Prochlorococcus* (HL I, II and LL I, II), and analyzed them in the context of the entire collection.

Results and discussion

Cyanophage isolation and host range

The cyanopodoviruses reported here were isolated over a period spanning more that a decade (1995-2006; Table 1). Diverse strains of *Prochlorococcus,* including representatives from both high-light and low-light clades*,* were used as hosts to isolate and maintain phage stocks (Table 1). In contrast to cyanomyoviruses, which can typically infect multiple bacterial strains (Sullivan et al., 2003), these cyanopodoviruses have narrow host ranges, infecting only one or two strains under laboratory conditions (Table 2).

General features of cyanopodovirus genomes

The general features of the cyanopodovirus genomes are shown in Table 1, and include 9 genomes reported for the first time, along with 5 existing genomes that were used for comparative analyses. The genomes of cyanopodovirus P-SSP7 and Syn5 are known to be linear, with direct terminal repeats (Pope et al., 2007; Sabehi and Lindell, 2012), and we assume that the new genomes are linear as well. The marine cyanopodovirus genomes range from 42.2 kbp to 47.7 kbp, and code for 48 to 68 putative open reading frames (ORFs). The majority of the putative genes are encoded on the same strand, but phage P-RSP2 and P60 that contain an inverted region of 1.5 kb and multiple genome rearrangements, respectively (ORF15-17$_{\text{P-RSP2}}$ – Fig. 3) (See Note added in proofs). Phage isolated on *Prochlorococcus* have a G+C content of 34% to 40.5%, while those isolated on *Synechococcus* range from 53% to 55% (Table 1) reflecting the different G+C content of the two hosts and the selective pressure for the phage to adapt their codon usage to that of their hosts (Krakauer and Jansen, 2002; Limor-Waisberg et al., 2011). The ability of cyanomyoviruses to cross-infect both *Prochlorococcus* and *Synechococcus*, despite their different G+C content, is thought to be facilitated by the tRNAs encoded by this group of phage (Enav et al., 2012). Only two tRNAs were identified in the cyanopodoviruses, however – one partial tRNA

148 in P-SSP7 (Sullivan et al., 2005) and one glycine tRNA in P-RSP5. The latter does not
149 correspond to a rare codon in its host genome or to a highly used codon in the P-RSP5 genome
150 (data not shown), suggesting that the G+C content difference between the genomes of
151 cyanopodoviruses that infect *Synechococcus* and *Prochlorococcus* is probably a significant
152 barrier to cross-infectivity (Enav et al., 2012).
153
154 *DNA Polymerase Phylogeny and the Core and Pan Genomes*
155
156 As a foundation for the analyses that follow, we wanted to identify the core genes shared by a
157 defined set of cyanopodoviruses, as well as their flexible gene set. Previous work on *Podoviridae*
158 DNA polymerase diversity suggests that this gene could be an acceptable phylogenetic tracer for
159 *Podoviridae* because it is conserved among different groups of phage and shows signs of vertical
160 inheritance (Chen et al. 2009; Labonté et al. 2009). Thus we used the phylogeny of this gene to
161 define sets of phage for the core and pan-genome analysis, and to guide our analysis of
162 relatedness among the phage. We first cast a broad net, including 71 DNA polymerase genes
163 from phage of different genera and families according to current International Committee on
164 Taxonomy of Viruses (ICTV) classification (Fig. 1). All cyanopodoviruses fell into the same
165 clade – designated the P60-like genus (Lavigne et al., 2008) – with the exception of two
166 freshwater cyanopodoviruses (indicated by three blue dots in Fig 1, as DNA polymerase is
167 encoded by two genes in one of the phage). The P60-like clade can be divided into three
168 subclades, supported by bootstrap values greater than 95% which exclude an outlier – P-RSP2 .
169 The first clade corresponds to the clade MPP-A (marine picocyanopodovirus A) established by
170 Chen and colleagues (2009), while the other two fall within clade MPP-B and form two discrete
171 clades (B1 and B2) (see the core genome phylogeny analysis section below – Figs 1 & 3).
172

Using an analysis similar to that described in Tettelin *et al* (2005) and used in our analysis of cyanomyoviruses (Sullivan et al., 2010), we first defined a set of core genes using only the 10 cyanopodoviruses isolated on *Prochlorococcus* (P-RSP2, P-HP1, P-SSP11, P-SSP10, P-GSP1, P-SSP2, P-SSP3, P-SSP7, P-RSP5 and P-SSP9 – Table 1). This core is composed of 19 genes (Fig. 2A); adding *Synechococcus*-specific phage Syn5 to the analysis reduces this number to 17 (Fig. 2B), and if *Synechococcus* phage P60 is added, the shared gene set drops to 12 (Table 3 – Fig. 2C). The significant impact of adding P60 is perhaps not surprising given its estuarine habitat. P60's genome also includes several frameshifts (see below) and incomplete proteins (Table 3) (See Note added in proofs). Finally, adding the two freshwater cyanopodoviruses to the analysis causes a precipitous drop to 3 core genes: primase/helicase, DNA polymerase, and terminase (Fig. 2D) – consistent with the divergence of these phage seen in the DNA polymerase tree (Fig. 1).

Of the 17 core genes shared by the 10 *Prochlorococcus* cyanopodoviruses and Syn5, 9 are involved in DNA metabolism and assembly of virions, 6 encode phage structural proteins (portal protein, MCP, tail tube proteins A and B, internal core protein, tail fiber), one encodes the terminase, and one codes for an hypothetical protein of unknown function (Table 3; Fig. 3, blue shading). The pan-genome of this set of cyanopodoviruses is composed of 241 clustered orthologous groups (COGs), and the cumulative curve of unique genes is nowhere near saturation, suggesting that vast diversity remains (Fig. 2). Each new genome contributed an average of 15 unique genes to the pan-genome, representing 22.0% to 31.6% of the genes in each genome. In a similar analysis of 16 cyanomyoviruses, each genome adds approximately 90 new genes, or 27.5% to 42.8% of their gene content (Sullivan et al., 2010). In both, the percentage is significantly higher than that observed for host strains, where each new sequenced genome added approximately 7.3% to 11.8% of their gene content to the pan-genome (Kettler et al., 2007).

Genome organization

With the exception of P60 (See Note added in proofs) and the two freshwater cyanophage (Pf-WMP3 and Pf-WMP4) gene order in these genomes is roughly consistent with their relatedness in the DNA polymerase tree and core genome analysis (Fig. 3). As in P-SSP7 (Sullivan et al., 2005), order is highly conserved, and strikingly similar to the distantly related prototype enterophage T7 (Dunn et al., 1983), supporting the hypothesis that T7-like enterophage and cyanopodoviruses evolved from a common ancestor, diverging at the protein sequence level (Sullivan et al., 2005; Lavigne et al., 2008) while keeping a similar genome organization. The exception is P60, which has multiple inversions (Fig. 3), rendering its genome architecture more similar to the freshwater cyanopodoviruses Pf-WMP3 and Pf-WMP4 (Liu et al., 2007; Liu et al., 2008), while its protein sequences are more similar to those of marine cyanophage (Liu et al., 2007). That is, P60 evolved with the other marine cyanopodoviruses in terms of protein sequences, but underwent multiple genomic rearrangements altering the T7-like genome architecture (See Note added in proofs). We note again, that P60 was isolated from an estuarine environment – quite distinct from the open ocean habitat of the other marine phages.

Similar to T7 (Molineux, 2006), P-SSP7 genes are grouped into three ordered classes of genes that are sequentially expressed over the course of infection – marked in red, green, and blue along the P-SSP7 genome in Fig. 3 (Lindell et al., 2007). Class I genes encode primarily small proteins, including MarR and gp0.7, thought to be involved in redirecting transcription from the host to the phage (Lindell et al., 2007). This region is highly variable and does not include core genes (see below). Class II includes genes from the RNA polymerase gene up to, but not including the major capsid protein (MCP) gene and is involved in transcription, DNA metabolism and replication, and code for phage scaffolding proteins and structural components. Class III consists of genes involved in phage assembly and DNA maturation (Molineux, 2006) and spans the rest of the genome (Lindell et al., 2007).

225

226 Since P60 was the first cyanopodovirus sequenced (Chen and Lu, 2002) we are upholding naming
227 conventions for phage and referring to this as the "P60-like genus" (Lavigne et al., 2008), even
228 though P60 is not a 'typical' phage in this group with respect to gene content and organization
229 (See Note added in proofs).

230

231 *Phylogeny and classification based on core genomes*

232 To further examine the phylogenetic groupings established above, the amino acid sequences of
233 the core genes shared by the marine cyanopodovirus genomes (Fig. 2C) were concatenated and
234 aligned, and a maximum likelihood analysis was applied (Fig. 3, tree on the left). Three distinct
235 subgroups (MPP-A, MPP-B1 and B2) emerged with a topology consistent with the DNA
236 polymerase tree above (compare Fig. 1 and Fig. 3), with P-RSP2 as an outlier, but still belonging
237 to the group. The two divergent freshwater cyanopodoviruses (Fig. 1) were excluded from this
238 core phylogeny analysis since they are missing most of the core genes (Fig. 2D).

239

240 Based on the sequence analysis of the concatenated core genomes (Fig. 3), and its congruence
241 with the DNA polymerase tree (Fig. 1), the 12 marine cyanopodoviruses in Fig. 3 belong to the
242 same genus – the P60-like genus of the subfamily of the *Autographivirinae.* Even though P-RSP2
243 is divergent from the other members of the group, it clearly falls within this clade. Because P-
244 RSP2 lacks an RNA polymerase gene, however, it would normally be excluded from the
245 *Autographivirinae* subfamily – which currently includes even very distantly related *Podoviridae*
246 (eg. T7 and phiKMV – Fig. 1, middle ring) – based on this single criterion. Although the presence
247 of RNA polymerase has been considered a hallmark gene for assignment of a phage to the
248 *Autographivirinae,* we argue that P-RSP2 should be included based on its similarities to other
249 phage in the P60-like genus (Figs. 1 & 3).

250

P-RSP2 – the outlier

P-RSP2 shares the same genome organization as the other cyanopodoviruses (with the exception of an inverted region in the class III genes), and has the same set of core genes, but it is highly divergent (Figs. 1 & 3). In fact, only one of its core genes (DNA polymerase – Fig. 1) shares more than 60% amino acid identity with the other phage. That it is the only phage in the group that was isolated on *Prochlorococcus* strain MIT9302 raises the question of whether there is something unique about this phage/host relationship. As discussed above, P-RSP2 is also the only phage in this group that lacks an RNA polymerase gene, essential for inclusion in the *Autographivirinae* (Lavigne et al., 2008), which in the canonical podovirus coliphage T7 is required for efficient transcription of class II and class III phage genes (Summers and Szybalski, 1968; Studier and Maizel, 1969; Studier, 1972).

Since P- RSP2 does not encode its own RNA polymerase, it likely has evolved mechanisms to use host transcriptional machinery to transcribe class II-III genes, such as additional host-like promoters or modulation of host RNA polymerase with transcriptional regulators such as sigma factors (Sullivan et al., 2009; Pavlova et al., 2012). In T4, for example, middle and late gene expression is coordinated by two transcriptional activators (Brody et al., 1995), but a search for similar activators in P-RSP2 yielded nothing. The G+C content of cyanopodoviruses prohibits the use of computational approaches like those of Vogel et al. (2003) to search for host-like promoters, thus the mechanism by which P-RSP2 transcribes Class II and III genes remains a mystery.

Comparative genomics

The Class I gene set (Fig. 3 – red under the P-SSP7 genome), is composed of very short genes that are highly variable. The set is most conserved in the MPP-B1 group relative to MPP-B2 and MPP-A, and consists of a genetic module of 10-13 genes that code for putative proteins mostly of

277 unknown function (Fig. 3). Genes of interest include an integrase (in 4 genomes), and a protein
278 similar to T7 gp0.7 (a transcriptional regulator involved in the takeover of the cellular metabolism
279 by the phage (Molineux, 2006), found in 3 genomes). Three of the 4 genomes that have the
280 integrase gene have a downstream integration signature sequence, suggestive of the potential for
281 lysogeny (discussed in more detail below).
282
283 Class II genes (Fig. 3 – green under the P-SSP7 genome) were among the most conserved (Table
284 3) across all three MPP groups. In addition to core genes, Class II also includes genes encoding
285 RNA polymerase (11/12 genomes), high light inducible proteins (Hli – 9/12 genomes),
286 photosystem II D1 protein (PsbA – 8/12 genomes) and transaldolase (TalC – 8/12 genomes).
287 These genes have orthologs in bacterial genomes (phage/host shared genes), and while
288 photosynthesis-associated genes are thought to have been derived from the host, the origin of *talC*
289 is not clear (Ignacio-Espinoza and Sullivan, 2012) (see discussion below). The genes *hli, psbA*
290 and *talC*, only found in MPP-B1 and MPP-B2, are common in cyanophage (Lindell et al., 2004;
291 Sullivan et al., 2005; Lindell et al., 2005; Sullivan et al., 2006; Chenard and Suttle, 2008; Sullivan
292 et al., 2010; Thompson et al., 2011; Sabehi et al.,2012) and are thought to increase phage fitness
293 during infection (Bragg and Chisholm, 2008; Thompson et al., 2011).
294
295 Class III genes (Fig. 3 – blue under the P-SSP7 genome) mainly consist of genes coding for
296 structural components of mature virions. This class contains a highly variable region that encodes
297 host specificity determinants, including genes in the region downstream of the tail tube protein B
298 ($gp31_{P\text{-}SSP7}$) and through the tail fiber protein ($gp36_{P\text{-}SSP7}$).
299
300 *P-SSP2 and P-SSP3: two co-isolated phage reveal a hypervariable genomic region*
301 Phage P-SSP2 and P-SSP3 were isolated on the same day, at the same station, from proximate
302 depths (120m and 100m respectively), using *Prochlorococcus* MIT9312 as the host. Their

303 genomes share 95% overall nucleotide sequence identity, and most proteins are 100% identical
304 (Fig. 4). They differ in only 7 genes (Table 5), each being either significantly divergent, or absent
305 in one or the other. The Class I module in the two genomes includes 2 pairs of divergent genes:
306 *gp14*$_{\text{P-SSP2}}$/*gp55*$_{\text{P-SSP3}}$ and *gp18*$_{\text{P-SSP2}}$/*gp52*$_{\text{P-SSP3}}$, whose gene products share 76% and 66% identity,
307 respectively. Immediately adjacent to the latter pair, P-SSP2 encodes an additional orphan gene
308 (*gp17*$_{\text{P-SSP2}}$) (Fig. 4) that does not share similarity with proteins in public databases. A second
309 divergent region is located at the C-terminus of the tail fiber(*gp16*$_{\text{P-SSP3}}$ and *gp57*$_{\text{P-SSP2}}$) (Fig. 4;
310 Table 5) involved in host recognition,. The P-SSP3 tail fiber gene (*gp16*$_{\text{P-SSP3}}$) is smaller than that
311 of (*gp57*$_{\text{P-SSP2}}$). Downstream of *gp16*$_{\text{P-SSP3}}$ are two small genes - *gp15*$_{\text{P-SSP3}}$ and *gp14*$_{\text{P-SSP3}}$ – that are
312 absent in the P-SSP2 genome. The former is an orphan while the latter shares 29% amino acid
313 identity with genes *gp40*$_{\text{P-SSP7}}$ (Figs. 1 & 3) – and 20% amino acid identity with *gp28*$_{\text{P-RSM4}}$ in a
314 cyanomyovirus isolated on *Prochlorococcus* MIT9303 (Sullivan et al., 2010). Genes *gp40*$_{\text{P-SSP7}}$
315 and *gp14*$_{\text{P-SSP3}}$ are located in the same genomic region (Fig. 3).
316
317 The N-terminal regions of all marine cyanopodoviruses tail fiber proteins are more conserved
318 than the C-terminal regions (data not shown). The hypervariable C-terminal regions likely help
319 phage adapt to host receptor diversity, and could either result from random
320 mutation/recombination events or through an active mechanism. The latter has been reported in
321 podoviruses that infect the pathogen *Bordetella* (Uhl and Miller, 1996), which encode a template-
322 dependent, reverse transcriptase-mediated diversity generating mechanism (Liu et al., 2002; Liu
323 et al., 2004; Doulatov et al., 2004), but we could find no evidence of this in our genomes. The
324 counterpart of this phage hypervariable region in their hosts was studied by Avrani et al. (2011).
325 They found that phage resistance in *Prochlorococcus* was acquired by accumulating mutations in
326 hypervariable genomic islands coding for cell surface receptors, among others. Together, these
327 recent findings beautifully illustrate the ongoing evolutionary arms race between phage and their
328 hosts.

329

330 *Phage/host shared genes, myo/podo shared genes, and genomic islands*

331 One of the most interesting features of some cyanophage is the set of genes they carry that appear
332 to be of bacterial origin (Mann, 2003; Lindell et al., 2004; Millard et al., 2004; Sullivan et al.,
333 2005; Lindell et al., 2005; Sullivan et al., 2006; Sullivan et al., 2010; Thompson et al., 2011;
334 Zeng and Chisholm, 2012) – 'phage/host shared genes' (Kelly et al., submitted) – 3 of the most
335 well studied examples being *psbA*, *talC*, and *hlis.* There are 66 genes in these cyanopodovirus
336 genomes with orthologs in *Prochlorococcus* and *Synechococcus* (Proportal
337 http://proportal.mit.edu/ - (Kelly et al., 2012)). They group into 12 COGs and are localized in
338 three regions of the phage genomes (Fig. 5A - diamonds). The first includes genes involved in
339 nucleotide metabolism that are found in all branches of the tree of life, and as such we don't
340 consider it an island. The second contains the *psbA* and *hli* genes, and the third includes *talC*,
341 which is involved in host carbon metabolism, a nuclease-encoding gene, and a gene of unknown
342 function – all genes likely acquired by horizontal gene transfer. These regions, which have some
343 similarity to the genomic islands found in cyanomyoviruses (Millard et al., 2009), are referred to
344 as Island II and III (Fig. 5A).

345

346 Island II (Fig. 3, pink shading), surrounded by core genes, is composed of up to 6 genes,
347 including *psbA* and *hli* and additional genes of unknown function (Table 4). Island II genes are
348 not present in the Syn5, and P-RSP2 genomes, and P-SSP9 has only the hli gene (Figs. 3 and
349 5A). The *psbA* and *hli* genes in this island have orthologs in cyanomyoviruses and hosts (Mann,
350 2003; Lindell et al., 2004; Lindell et al., 2005; Sullivan et al., 2006), so we wondered whether the
351 rest of the genes in this island did as well (Table 4). gp222_COG and gp30_COG, clusters of
352 genes coding for hypothetical proteins, have orthologs in cyanomyoviruses but not in
353 picocyanobacteria, while gp32_COG has orthologs only in host genomes (Table 4). While the
354 synteny of Island II is not present in the hosts or cyanomyoviruses (data not shown), orthologous

355 genes in cyanomyovirus were often located within 15-20 genes of each other suggesting that
356 Island II was likely acquired in small pieces via multiple gene gain events, or as a larger insert
357 that underwent a series of deletions and reorganizations.
358
359 Analysis of the phylogeny of the *psbA* and *talC* genes in this expanded set of phage genomes (Fig
360 S1 and S2) generally confirms the conclusions of other reports (Lindell et al., 2004; Millard et al.,
361 2004; Sullivan et al., 2006; Ignacio-Espinoza and Sullivan, 2012) that phage *psbA* was not
362 recently acquired from picocyanobacteria (Fig. S1) and was likely acquired multiple times
363 (Ignacio-Espinoza *et al.* 2012). But while the cyanomyovirus *psbA* genes are closely related to
364 their specific hosts (Fig. S1), cyanopodovirus *psbA* genes form a clade distinct from those from
365 both cyanomyoviruses and hosts (Fig. S1). Further, cyanopodovirus *psbA* genes appear more
366 diverse than those of cyanomyoviruses, as indicated by the long branch lengths. As for *talC*, we
367 confirm that the origin of phage *talC* is less clear, as it differs significantly from
368 picocyanobacterial versions of this gene (Ignacio-Espinoza and Sullivan, 2012). In fact, phage
369 *talC* genes are more related to organisms from different phyla (Gammaproteobacteria, Firmicute
370 and Actinobacteria – Fig. S2). In contrast to *psbA*, cyanophage *talC* genes are highly conserved,
371 form a monophyletic clade, and likely were only acquired once and then diverged (Ignacio-
372 Espinoza and Sullivan, 2012).
373
374 It is intriguing that if a genome has any of the three genes, *psbA*, *hli* or *talC,* it has them all - with
375 the exception of P-SSP9 which has only one hli gene (Table 3). While Island II contains *psbA*
376 and *hli,* and is in the middle of the genome, *talC* is at the extreme downstream end, making it
377 unlikely that this set of genes could be simultaneously acquired or lost. Yet they are linked in the
378 observed gene gain/loss pattern (Fig. 5A – green and turquoise diamonds in Island II, and red
379 diamonds in Island III) and their co-expression, despite their separation in the genome, led
380 Lindell et al. (2007) to argue that their physical separation might reflect "evolution in progress"

381 i.e. an initial step toward the co-localization of these co-transcribed genes (Molineux, 2006;
382 Lindell et al., 2007). The fact that *talC* lies at the end of all of the cyanopodoviruses now in our
383 collection, however, argues against this, and suggests that there is something significant about
384 this positioning that still eludes us.
385
386 We found 59 proteins (grouped into 16 COGs) shared only by cyanopodo and cyanomyoviruses –
387 i.e. not present in hosts – and all are of unknown function (Table 6). The majority are in Islands II
388 and III (Fig. 5A; Table 6) – also the location of all of the phage/host shared genes.
389
390 The mechanisms underlying the genetic variability in islands in cyanopodoviruses are not clear.
391 In small lambda-like siphoviruses, rapid evolution is facilitated by structural simplicity, a small
392 set of core genes, and the exchange of compatible genetic modules (Botstein, 1980; Hendrix et
393 al., 1999; Comeau et al., 2007). T4-like myoviruses, on the other hand, have a significantly
394 larger, and syntenic, set of core genes, that are for the most part vertically inherited (Filée et al.,
395 2006; Comeau et al., 2007; Ignacio-Espinoza and Sullivan, 2012). This core is involved in
396 replication and assembly of the viruses, often requiring complex protein-protein interactions
397 (Leiman et al., 2003), which reduces the probability of acquiring functional orthologs. Thus in
398 T4-like phage, horizontal gene transfer events are concentrated in hypervariable islands (Comeau
399 et al., 2007; Millard et al., 2009), while the optimal core genome is kept intact (Comeau et al.,
400 2007). Cyanopodoviruses appear to use a strategy similar to T4-like phages, accessing the genetic
401 diversity thought to be involved in adaptation to their host's metabolism and ecological niche
402 through genomic islands (Filée et al., 2006; Comeau et al., 2007), while conserving an optimal
403 core genome.
404
405 *The flexible genome positioning reveals more islands*

406 We explored whether the frequency of occurrence of a gene in this set of phage (Fig. 2) would be
407 reflected in the position of that gene in a genome, hoping that this might ultimately yield insights
408 into gene gain and loss mechanisms. We divided the flexible COGs into 3 groups for this
409 analysis: i) hyperflexible genes (found in 1-3 genomes – Fig. 5B, red diamonds), ii) flexible genes
410 (found in 4-6 genomes – Fig. 5B, green diamonds), and iii) conserved flexible genes (found in 7-
411 10 genomes – Fig. 5B, blue diamonds). The hyperflexible genes are concentrated in the left
412 extremity of the genomes, which we name Island I, while the flexible genes are more
413 concentrated in Island II and the right arm of the genome (Island III). Finally, the core and the
414 conserved flexible genes appear more distributed along the middle, and slightly in the right arm
415 of the genomes.
416
417 Assuming that these cyanopodoviruses reproduce similarly to T7 (Wolfson et al., 1972;
418 Molineux, 2006), in which the genome replicates as linear concatemers that are cleaved before
419 encapsidation, the propensity of hypervariable genes to be located in Island I could suggest that
420 gene gain/loss events occur primarily at the extremities of the linear genomes. An alternative
421 explanation is lysogeny, in which the temperate phage integrates into the host genome as a linear
422 fragment, and the excision of the phage genome from host chromosome may be imprecise. Two
423 published cyanopodovirus genomes (P-SSP7 (Sullivan et al., 2005) and Syn5 (Pope et al., 2007))
424 and three reported here (P-SSP2, P-SSP3 and P-SSP9) encode a phage-like integrase gene.
425 Furthermore, a 40-50 bp sequence with a perfect match to a cyanobacterial host sequence is found
426 downstream – suggesting a possible host integration site (Sullivan et al., 2005).
427
428 Despite indirect evidence for lysogeny in picocyanobacteria (McDaniel et al., 2002; Ortmann et
429 al., 2002), none of the complete marine cyanobacterial genomes examined contains an intact
430 prophage. This is perhaps not surprising as it is thought that lysogeny is favored when the
431 environment is not optimal for growth of host cells, the opposite of optimally growing laboratory

cultures (Waterbury and Valois, 1993). Recently, however, a partial prophage sequence, highly similar to P-SSP7, was found in a genome fragment from a wild *Prochlorococcus* single-cell (Malmstrom et al., 2012).

Biogeography of cyanopodoviruses

To analyze the distribution of the cyanopodoviruses in the oceans and place it in the context of their hosts and other cyanophage, we recruited reads from marine metagenomic datasets using all the cyanophage genomes available (see methods) (Fig. 6-7). We first examined the relative number of metagenomic reads recruited by cyanosipho-, podo-, and myovirus genomes in the viral metagenome samples from the HOT212 sample (N. Pacific) and "Marine Virome". Using only the 3 previously published cyanopodovirus genomes to recruit, cyanopodoviruses represent 22% of all recruited reads in the HOT212 sample (Fig. 6). This jumps to 50% if all 12 genomes are used for recruitment, and a similar proportion emerges from the analysis of the MarineVirome database (Fig. 6).

Analysis of the relative abundance of the three viral groups in the bacterial-fraction metagenomes from the North Pacific (HOT), Bermuda (BATS), Mediterranean (MedDCM), and the Global Ocean Survey (GOS) (Fig. 6) revealed the dominance of cyanomyoviruses in all samples, consistent with the observations of others for GOS and MedDCM databases (Williamson et al., 2008; Huang et al., 2011). The significant overabundance of cyanomyoviruses in these samples relative to those from the viral fraction ("Marine Virome and HOT212") samples is likely due to the larger size of cyanomyoviruses, which would cause them to be preferentially retained by filters, either attached to cells or freely floating.

We analyzed the geographic distribution of cyanopodo- and cyanomyoviruses in the Global Ocean Survey (GOS) and found that cyanopodoviruses are widespread but appear to be more

458 abundant in the Caribbean Sea, the Gulf of Mexico, the Eastern Tropical Pacific Ocean and the
459 Indian Ocean (Fig. 7B). Interestingly, abundance of *Prochlorococcus* recruited reads also
460 qualitatively corresponds to areas of relatively high cyanopodovirus counts (Fig. 7C). Thus
461 although quantitative assessments are not possible, the additional reference genomes for
462 cyanopodoviruses help document their widespread distribution, and point to some hotspots of
463 abundance.
464
465 **Conclusions and future directions**
466
467 The growing number of cyanophage genomes is helping us better understand their relatedness
468 and evolution, and their interactions with their host cells. Here we used four approaches to
469 explore the similarities and differences among cyanopodoviruses: DNA polymerase phylogeny,
470 concatenated core genome phylogeny, the presence or absence of RNA polymerase, and genome
471 architecture. All but the extremely divergent freshwater cyanopodoviruses would fall into the
472 "P60-like genus" by these criteria, except for P-RSP2, which is an outlier in the concatenated
473 core genome tree, and lacks the hallmark RNA polymerase gene for this group. It is also the only
474 phage isolated on *Prochlorococcus* MIT9302. Bcause its core genome architecture is similar to
475 the others over much of the genome, and its position in the DNA polymerase tree assigns it to the
476 "P60-like genus" group, we include it here.
477
478 Cyanopodoviruses have two hypervariable island regions in which genes shared with their hosts,
479 and/or with cyanomyoviruses, are concentrated. The positions of hyperflexible genes – i.e. those
480 found in only 1 to 3 genomes – are highly concentrated in a third island at one extremity of the
481 genome. These islands point to interesting regions for unveiling gene acquisition and loss
482 mechanisms. Another hypervariable region, at a finer evolutionary scale, encompasses the C-
483 terminal part of the tail fiber gene in the two very closely related phage, P-SSP2 and P-SSP3.

484 This region may indicate an underlying diversity-generating mechanism, helping phage to adapt
485 to the vast diversity of host receptors found in marine environments.
486
487 Our analysis contributes to the growing appreciation of the complexity of phage diversity in the
488 oceans, and the degree to which it is under-sampled.
489
490 **Materials and Methods**
491
492 *Bacteriophage isolation, characterization, DNA extraction*
493 Phage were isolated as previously described (Waterbury and Valois, 1993; Sullivan et al., 2003).
494 All phage used in this study were isolated by triple (or greater) plaque purification, followed by
495 two rounds of dilution to extinction. The phage stocks were filtered through 0.2μm and stored at
496 4˚C in the dark. For each phage, we used the earliest sample in our collection that still retained
497 infectivity, to minimize the number of infectious cycles the phage went through – and therefore,
498 the accumulation of mutations in the genome. Nonetheless, all of these phage went through
499 multiple transfers on serially transferred host cultures before the final stock was collected for
500 sequencing. The DNA was extracted as previously described (Henn et al., 2010).
501
502 *Genome sequencing, assembly and annotation*
503 The genomes were sequenced by 454 pyrosequencing, and assembled and annotated at the Broad
504 Institute as previously described (Henn et al., 2010). The protein sequences were clustered into
505 orthologous groups using OrthoMCL program (van Dongen and Abreu-Goodger, 2012) (see
506 below) with the available cyanophage genomes on Proportal (http://proportal.mit.edu/). The
507 protein functional annotations were updated based on the information available on ProPortal.
508
509 *Comparative genomics*

For Figure 3 and Figure 4, all marine cyanopodovirus proteins were compared using the program BLASTP (NCBI). The genomes in Figure 3 were extracted from the GenBank file using the software BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) and imported in Adobe Illustrator. The comparison of P-SSP2 and P-SSP3 was done using BLASTP and the genome maps were generated in R using the package GenoplotR (Guy et al., 2010).

Core genome analysis

The method used for clustering cyanopodovirus proteins into homologous groups was similar to that described previously(Kettler et al., 2007; Sullivan et al., 2010). All marine cyanopodovirus proteins were paired using a reciprocal best BLASTP hit analysis where the sequence alignment covered at least 75% of the protein length of the longest protein and where the percentage of identity was at least 35%. The clusters were then built by transiently grouping these pairs. To increase the sensitivity of the method, HMM profiles (Sonnhammer et al., 1998) were built for each cluster from an alignment of proteins made with Muscle (version 3.7 (Edgar, 2004; Edgar, 2004). The protein database was then searched *de novo* using the HMM models to group proteins with significant homology (E-value ≤ 1e-5). HMMBUILD and HMMSEARCH from HMMER were used to build and search for motifs in the sequence database, respectively.

Phylogeny of the core genome and of the DNA polymerase

All marine cyanopodoviruses were included for this analysis while the freshwater cyanopodoviruses were excluded because they lack most of the core genes. For each phage, the core protein sequences were concatenated in the same order, from the single strand binding protein to the terminase. The concatenated protein sequences were then aligned with MUSCLE (Edgar, 2004; Edgar, 2004) using the default parameters. The alignment was converted to phylip format using the BioPython package (Cock et al., 2009). Phylogenetic analysis of the concatenated proteins was performed using PhyML 3.0 (Guindon et al., 2010). The trees were

built from the command line with the following options: -d aa -b -4 -m JTT -v e -c 4 -a e -o tlr. Both trees are unrooted. The approach NNIs was used to search the tree topology. The initial tree was based on the BioNJ algorithm using the substitution model JTT (Jones et al., 1992). A discrete gamma model was estimated by the software with 4 categories and a gamma shape of 1.384 with a proportion of invariant a.a. of 0.042. The maximum likelihood was estimated using the Shimodaira–Hasegawa–like procedure (Shimodaira, 2002). Finally, the trees were visualized with the online tool iTOL (Letunic and Bork, 2007; Letunic and Bork, 2011). The sequences of the DNA polymerase were retrieved from ACLAME database (ACLAME MGEs. Version 0.4 - family_vir_proph_26 (Leplae et al., 2009)) and were aligned as described above; the tree was built using the same approach as the core genome phylogeny analysis.

Phage/host shared genes and hypervariable genetic islands in cyanopodoviruses

Clustering cyanopodovirus/host and cyanopodovirus/cyanomyovirus shared genes was performed using the OrthoMCL program (van Dongen and Abreu-Goodger, 2012). The clustering was done with a conservative value of 35% for the percent identity and an E-value of 1E-05. To avoid clustering proteins solely on the basis of conserved domains, we pre-filtered our BLASTP results to accept the orthologous pairs only if the sequence alignment covered at least 75% of the length of the longer of the two sequences. The cyanophage and picocyanobacterial genomes used in the clustering analysis are listed in supplemental Table 1. Figure 5 was generated using the python matplotlib module (Hunter, 2007).

P-RSP2 promoter analysis and transcriptional factor searches

The P-RSP2 genome was screened for promoters as previously described (Vogel et al., 2003; Lindell et al., 2007). Briefly, a position-specific weight matrix was built from the -10 box of *Prochlorococcus* MED4 (Vogel et al., 2003) with the Motif module from the BioPython package (Cock et al., 2009). The phage genomes were searched for this motif. The threshold was set at 7.2

562 based on the distribution of scores for the established motif for the -10 promoter box sequences.
563 P-RSP2 coding sequences were analyzed to detect transcription factors using InterProScan
564 (Zdobnov and Apweiler, 2001), Pfam (Punta et al., 2012), and CDD (Marchler-Bauer and Bryant,
565 2004). We were specifically looking for conserved protein domains related to transcription
566 factors or DNA binding domain. Except for the phage proteins known to be involved in DNA
567 metabolism (DNA polymerase, endo/exonuclease, DNA primase, single strand binding protein),
568 no DNA binding motifs could be detected nor conserved domains related to transcription factors.
569
570 *Metagenomics*
571 Six metagenomic datasets were used in this study: four from the bacterial fraction, (The Global
572 Ocean Survey dataset (GOS (Rusch et al., 2007)), the deep chlorophyll max Mediterranean
573 dataset (Ghai et al., 2010), the Pacific Ocean datasets (Station Hawaii Ocean Time-Series –
574 HOT179 and HOT186 (Frias-Lopez et al., 2008; Coleman and Chisholm, 2010)) and two viral
575 fraction datasets (the MarineVirome (Angly et al., 2006) and the Pacific Ocean dataset (HOT212
576 (this study – NCBI accession: SRA059090)). All datasets, except HOT212, were obtained from
577 the CAMERA website (http://camera.calit2.net/index.shtm). Only the sites with more than 10,000
578 reads were used from the GOS database. The methods used were similar to those described by
579 Malmstrom *et al* (2012) , and the reference genomes used for recruitment are listed in
580 supplemental Table 2. Briefly, metagenomic reads were matched to reference genomes using
581 BLASTN (Table S1), and those with a bit score of at least 40 were compared against the NCBI nt
582 database to assess if there were other best hits. The number of recruited reads at a GOS site was
583 normalized against the number of reads in the GOS database from that site. Finally, to compare
584 the relative abundance of cyanopodo- and cyanomyoviruses, the normalized read counts for each
585 GOS site were normalized to the average genome size of each phage family – 188780 bp and
586 46320 bp for the cyanomyo- and cyanopodoviruses respectively. The bar graphs were generated
587 in R using ggplot2 package (Wickham, 2009) and the map was generated in R using ggplot2

(Wickham, 2009), maps (http://CRAN.R-project.org/package=maps), gpclib (http://CRAN.R-project.org/package=gpclib), and maptools (http://CRAN.R-project.org/package=maptools) packages. The shapefile used to create the Galapagos Islands inset was downloaded from © OpenStreetMap contributors (http://downloads.cloudmade.com).

Note added to proof

After this manuscript was accepted, we learned that a new version of P60 genome has been generated (Feng Chen, pers. comm.), which contains significant changes from the published version (Chen and Lu, 2002). We re-examined our data in the context of this revised P60 genome and found that some of our statements need to be modified, but the main conclusions of the paper remain the same.

First, the revised P60 genome organization now makes it more similar to the other cyanopodoviruses, and all the genes are coded on the same DNA strand. Further, this genome makes P60 fall squarely in the P60-like genus as defined by Lavigne et al. (2008). The revised sequence also affects our core gene analysis such that marine cyanopodoviruses and P60 now share 15 core genes instead of 12.

Acknowledgments

We are grateful to Jessie W. Thompson and Qinglu Zeng for comments and edits on the manuscript, and Katherine Huang for her advice and analyses in the early stages of the genome sequencing. This work was supported by grants from the Gordon and Betty Moore Foundation (SWC and MRH), the US National Science Foundation (NSF) Biological Oceanography Section, the NSF Center for Microbial Oceanography Research and Education (C-MORE) (grant numbers OCE-0425602 and EF 0424599), and the US Department of Energy-GTL. SJL was supported by a postdoctoral fellowship from the “Fonds Québecois de la recherche sur la nature et les technologies”.

614 **References**

615 Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., et al. (2006) The
616 marine viromes of four oceanic regions. *PLoS Biol.* **4**: e368.
617 Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011) Genomic island variability
618 facilitates *Prochlorococcus*-virus coexistence. *Nature.* **474**: 604-608.
619 Bergh, O., Børsheim, K.Y., Bratbak, G., and Heldal, M. (1989) High abundance of viruses found
620 in aquatic environments. *Nature.* **340**: 467-468.
621 Béjà, O., Fridman, S., and Glaser, F. (2012) Viral clones from the GOS expedition with an
622 unusual photosystem-I gene cassette organization. *ISME. J.* **6**: 1617-1620.
623 Bhaya, D., Dufresne, A., Vaulot, D., and Grossman, A. (2002) Analysis of the hli gene family in
624 marine and freshwater cyanobacteria. *FEMS Microbiol. Lett.* **215**: 209-219.
625 Botstein, D. (1980) A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.* **354**:
626 484-490.
627 Bragg, J.G., and Chisholm, S.W. (2008) Modeling the fitness consequences of a cyanophage-
628 encoded photosynthesis gene. *PLoS One.* **3**: e3550.
629 Breitbart, M., Thompson, L.R., Suttle, C.A., and Sullivan, M.B. (2007) Exploring the Vast
630 Diversity of Marine Viruses. *Oceanography.* **20**: 135-139.
631 Brody, N., Kassavetis, A., Ouhammouch, M., Sanders, M., Tinker, L., and Geiduschek, P. (1995)
632 Old phage, new insights: Two recently recognized mechanisms of transcriptional regulation
633 in bacteriophage T4 development. *FEMS Microbiol. Lett.* **128**: 1-8.
634 Chen, F., and Lu, J. (2002) Genomic Sequence and Evolution of Marine Cyanophage P60: a New
635 Insight on Lytic and Lysogenic Phages. *Appl. Environ. Microbiol.* **68**: 2589-2594.
636 Chenard, C., and Suttle, C.A. (2008) Phylogenetic Diversity of Cyanophage Photosynthetic
637 Genes (*psbA*) in Marine and Fresh Waters. *Appl. Environ. Microbiol.* **74**: 5317-5324.
638 Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., et al. (2009) Biopython:
639 freely available Python tools for computational molecular biology and bioinformatics.
640 *Bioinformatics.* **25**: 1422-1423.
641 Coleman, M.L., and Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed
642 through comparative population genomics. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 18634-
643 18639.
644 Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., DeLong, E.F., and
645 Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*.
646 *Science.* **311**: 1768-1770.
647 Comeau, A.M., Bertrand, C., Letarov, A., Tétart, F., and Krisch, H.M. (2007) Modular
648 architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery.
649 *Virology.* **362**: 384-396.
650 Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., et al. (2004) Tropism
651 switching in *Bordetella* bacteriophage defines a family of diversity-generating
652 retroelements. *Nature.* **431**: 476-481.
653 Dunn, J.J., Studier, F.W., and Gottesman, M. (1983) Complete nucleotide sequence of
654 bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166**: 477-535.
655 Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and
656 space complexity. *BMC. Bioinformatics.* **5**: 113.
657 Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high
658 throughput. *Nucleic Acids Res.* **32**: 1792-1797.
659 Enav, H., Béjà, O., and Mandel-Gutfreund, Y. (2012) Cyanophage tRNAs may have a role in
660 cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. *ISME J.* **6**: 619-628.
661 Filée, J., Bapteste, E., Susko, E., and Krisch, H.M. (2006) A selective barrier to horizontal gene
662 transfer in the T4-type bacteriophages that has preserved a core genome with the viral
663 replication and structural genes. *Mol. Biol. Evol.* **23**: 1688-1696.

664 Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W., and
665 DeLong, E.F. (2008) Microbial community gene expression in ocean surface waters. *Proc.*
666 *Natl. Acad. Sci. U. S. A.* **105**: 3805-3810.
667 Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature.*
668 **399**: 541-548.
669 Funk, C., and Vermaas, W. (1999) A cyanobacterial gene family coding for single-helix proteins
670 resembling part of the light-harvesting proteins from higher plants. *Biochemistry.* **38**: 9397-
671 9404.
672 Ghai, R., Martin-Cuadrado, A.B., Molto, A.G., Heredia, I.G., Cabrera, R., Martin, J., et al. (2010)
673 Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid
674 library 454 pyrosequencing. *ISME J.* **4**: 1154-1166.
675 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010)
676 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
677 performance of PhyML 3.0. *Syst. Biol.* **59**: 307-321.
678 Guy, L., Kultima, J.R., and Andersson, S.G. (2010) genoPlotR: comparative gene and genome
679 visualization in R. *Bioinformatics.* **26**: 2334-2335.
680 He, Q., Dolganov, N., Bjorkman, O., and Grossman, A.R. (2001) The high light-inducible
681 polypeptides in *Synechocystis* PCC6803. Expression and function in high light. *J. Biol.*
682 *Chem.* **276**: 306-314.
683 Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999) Evolutionary
684 relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc.*
685 *Natl. Acad. Sci. U. S. A.* **96**: 2192-2197.
686 Henn, M.R., Sullivan, M.B., Stange-Thomann, N., Osburne, M.S., Berlin, A.M., Kelly, L., et al.
687 (2010) Analysis of High-Throughput Sequencing and Annotation Strategies for Phage
688 Genomes. *PLoS One.* **5**: e9083.
689 Huang, S., Wang, K., Jiao, N., and Chen, F. (2011) Genome sequences of siphoviruses infecting
690 marine *Synechococcus* unveil a diverse cyanophage group and extensive phage-host genetic
691 exchanges. *Environ. Microbiol.* **14**: 540-558.
692 Hunter, J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**: 90-95.
693 Ignacio-Espinoza, J.C., and Sullivan, M.B. (2012) Phylogenomics of T4 cyanophages: lateral
694 gene transfer in the 'core' and origins of host genes. *Environ. Microbiol.* **14**: 2113-2126.
695 Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) The rapid generation of mutation data
696 matrices from protein sequences. *Comput. Appl. Biosci.* **8**: 275-282.
697 Kelly L, Ding H, Huang KH, Osburne M, Chisholm SW. (submitted) Features of cyanomyophage
698 gene content and evolution in wild populations and cultured genomes. *ISME J.*
699 Kelly, L., Huang, K.H., Ding, H., and Chisholm, S.W. (2012) ProPortal: a resource for integrated
700 systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res.* **40**: D632-D640.
701 Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., et al. (2007)
702 Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS*
703 *Genet.* **3**: e231.
704 Krakauer, D.C., and Jansen, V.A. (2002) Red queen dynamics of protein translation. *J. Theor.*
705 *Biol.* **218**: 97-109.
706 Labonté, J.M., Reid, K.E., and Suttle, C.A. (2009) Phylogenetic analysis indicates evolutionary
707 diversity and environmental segregation of marine podovirus DNA polymerase gene
708 sequences. *Appl. Environ. Microbiol.* **75**: 3634-3640.
709 Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.W., and Kropinski, A.M. (2008) Unifying
710 classical and molecular taxonomic classification: analysis of the *Podoviridae* using
711 BLASTP-based tools. *Res. Microbiol.* **159**: 406-414.
712 Leiman, P.G., Kanamaru, S., Mesyanzhinov, V.V., Arisaka, F., and Rossmann, M.G. (2003)
713 Structure and morphogenesis of bacteriophage T4. *Cell. Mol. Life Sci.* **60**: 2356-2370.

714 Leplae, R., Lima-Mendez, G., and Toussaint, A. (2009) ACLAME: A CLAssification of Mobile
715 genetic Elements, update 2010. *Nucleic Acids Res.* **38**: D57-D61.
716 Letunic, I., and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic
717 tree display and annotation. *Bioinformatics.* **23**: 127-128.
718 Letunic, I., and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of
719 phylogenetic trees made easy. *Nucleic Acids Res.* **39**: W475-W478.
720 Limor-Waisberg, K., Carmi, A., Scherz, A., Pilpel, Y., and Furman, I. (2011) Specialization
721 versus adaptation: two strategies employed by cyanophages to enhance their translation
722 efficiencies. *Nucleic Acids Res.* **39**: 6016-6028.
723 Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., et al. (2007)
724 Genome-wide expression dynamics of a marine virus and host reveal features of co-
725 evolution. *Nature.* **449**: 83-86.
726 Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis
727 genes in marine viruses yield proteins during host infection. *Nature.* **438**: 86-89.
728 Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004)
729 Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad.*
730 *Sci. U. S. A.* **101**: 11013-11018.
731 Liu, H., Campbell, L., Landry, M.R., Nolla, H.A., Brown, S.L., and Constantinou, J. (1998)
732 *Prochlorococcus* and *Synechococcus* growth rates and contributions to production in the
733 Arabian Sea during the 1995 Southwest and Northeast Monsoons. *Deep-Sea Res. Part II.* **45**:
734 2327-2352.
735 Liu, H., Nolla, H.A., and Campbell, L. (1997) *Prochlorococcus* growth rate and contribution to
736 primary production in the equatorial and subtropical North Pacific Ocean. *Aquat. Microb.*
737 *Ecol.* **12**: 39-47.
738 Liu, M., Deora, R., Doulatov, S.R., Gingery, M., Eiserling, F.A., Preston, A., et al. (2002)
739 Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science.* **295**:
740 2091-2094.
741 Liu, M., Gingery, M., Doulatov, S.R., Liu, Y., Hodes, A., Baker, S., et al. (2004) Genomic and
742 genetic analysis of *Bordetella* bacteriophages encoding reverse transcriptase-mediated
743 tropism-switching cassettes. *J. Bacteriol.* **186**: 1503-1517.
744 Liu, X., Kong, S., Shi, M., Fu, L., Gao, Y., and An, C. (2008) Genomic analysis of freshwater
745 cyanophage Pf-WMP3 Infecting cyanobacterium *Phormidium foveolarum*: the conserved
746 elements for a phage. *Microb. Ecol.* **56**: 671-680.
747 Liu, X., Shi, M., Kong, S., Gao, Y., and An, C. (2007) Cyanophage Pf-WMP4, a T7-like phage
748 infecting the freshwater cyanobacterium *Phormidium foveolarum*: complete genome
749 sequence and DNA translocation. *Virology.* **366**: 28-39.
750 Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., et al. (2012)
751 Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and
752 biogeographic analysis. *ISME J.* 10.1038/ismej.2012.89.
753 Mann, N.H. (2003) Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol.*
754 *Rev.* **27**: 17-34.
755 Marchler-Bauer, A., and Bryant, S.H. (2004) CD-Search: protein domain annotations on the fly.
756 *Nucleic Acids Res.* **32**: W327-W331.
757 McDaniel, L., Houchin, L.A., Williamson, S.J., and Paul, J.H. (2002) Lysogeny in marine
758 *Synechococcus*. *Nature.* **415**: 496.
759 Millard, A., Clokie, M.R., Shub, D.A., and Mann, N.H. (2004) Genetic organization of the
760 psbAD region in phages infecting marine *Synechococcus* strains. *Proc. Natl. Acad. Sci. U. S.*
761 *A.* **101**: 11007-11012.
762 Millard, A.D., Zwirglmaier, K., Downey, M.J., Mann, N.H., and Scanlan, D.J. (2009)
763 Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of

764 *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of
765 cyanophage evolution. *Environ. Microbiol.* **11**: 2370-2387.
766 Molineux, I. (2006) The T7 group, In *The bacteriophages*. Calendar, R. (eds). New York: Oxford,
767 UK: Oxford University Press, pp. 277-301.
768 Ortmann, A.C., Lawrence, J.E., and Suttle, C.A. (2002) Lysogeny and Lytic Viral Production
769 during a Bloom of the Cyanobacterium *Synechococcus* spp. *Microb. Ecol.* **43**: 225-231.
770 Partensky, F., Hess, W.R., and Vaulot, D. (1999) *Prochlorococcus*, a marine photosynthetic
771 prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**: 106-127.
772 Pavlova, O., Lavysh, D., Klimuk, E., Djordjevic, M., Ravcheev, D.A., Gelfand, M.S., et al.
773 (2012) Temporal Regulation of Gene Expression of the *Escherichia coli* Bacteriophage
774 phiEco32. *J. Mol. Biol.* **416**: 389-399.
775 Pope, W.H., Weigele, P.R., Chang, J., Pedulla, M.L., Ford, M.E., Houtz, J.M., et al. (2007)
776 Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a
777 "horned" bacteriophage of marine *Synechococcus*. *J. Mol. Biol.* **368**: 966-981.
778 Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012) The
779 Pfam protein families database. *Nucleic Acids Res.* **40**: D290-D301.
780 Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., et al.
781 (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through
782 Eastern Tropical Pacific. *PLoS Biology.* **5**: 398-431.
783 Sabehi, , G., and Lindell D (2012) The P-SSP7 Cyanophage Has a Linear Genome with Direct
784 Terminal Repeats. *PLoS One.* **7**: e36710.
785 Sabehi, G., Shaulov, L., Silver, D.H., Yanai, I., Harel, A., and Lindell, D. (2012) A novel lineage
786 of myoviruses infecting cyanobacteria is widespread in the oceans. Proc. Natl. Acad. Sci. U.
787 S. A. **109**: 2037-2042.Scanlan, D.J., and West, N.J. (2002) Molecular ecology of the marine
788 cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol. Ecol.* **40**: 1-
789 12.
790 Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., et al. (2009)
791 Photosystem I gene cassettes are present in marine virus genomes. *Nature.* **461**: 258-262.
792 Shimodaira, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.*
793 **51**: 492-508.
794 Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998) Pfam: multiple
795 sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research.* **26**:
796 320-322.
797 Studier, F.W. (1972) Bacteriophage T7. *Science.* **176**: 367-376.
798 Studier, F.W., and Maizel, J.V. (1969) T7-directed protein synthesis. *Virology.* **39**: 575-586.
799 Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three
800 *Prochlorococcus* Cyanophage Genomes: Signature Features and Ecological Interpretations.
801 *PLoS Biol.* **3**: e144.
802 Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R., et al.
803 (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like
804 myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**: 3035-3056.
805 Sullivan, M.B., Krastins, B., Hughes, J.L., Kelly, L., Chase, M., Sarracino, D., and Chisholm,
806 S.W. (2009) The genome and structural proteome of an ocean siphovirus: a new window
807 into the cyanobacterial 'mobilome'. *Environ. Microbiol.* **11**: 2935-2951.
808 Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W.
809 (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial
810 viruses and their hosts. *PLoS Biology.* **4**: 1344-1357.
811 Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic
812 cyanobacterium *Prochlorococcus*. *Nature.* **424**: 1047-1051.
813 Summers, W.C., and Szybalski, W. (1968) Totally asymmetric transcription of coliphage T7 in
814 vivo: correlation with poly G binding sites. *Virology.* **34**: 9-16.

Suttle, C.A., and Chan, A.M. (1994) Dynamics and Distribution of Cyanophages and Their Effect on Marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **60**: 3167-3174.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* **102**: 13950-13955.

Thingstad, T.F. (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**: 1320-1328.

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A.U., Stubbe, J., and Chisholm, S.W. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **108**: E757-E764.

Uhl, M.A., and Miller, J.F. (1996) Integration of multiple domains in a two-component sensor protein: the *Bordetella pertussis* BvgAS phosphorelay. *EMBO Journal.* **15**: 1028-1036.

van Dongen, S., and Abreu-Goodger, C. (2012) Using MCL to extract clusters from networks. *Methods. Mol. Biol.* **804**: 281-295.

Vogel, J., Axmann, I.M., Herzel, H., and Hess, W.R. (2003) Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. *Nucleic Acids Res.* **31**: 2890-2899.

Waterbury, J.B., and Valois, F.W. (1993) Resistance to cooccurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl. Environ. Microbiol.* **59**: 3393-3399.

Waterbury, J.B., Watson, S.W., Valois, F.W., and Franks, D.G. (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat.. Sci.* **214**: 71-120.

Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F., et al. (2007) Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus. Environ. Microbiol.* **9**: 1675-1695.

Weinbauer, M.G., and Rassoulzadegan, F. (2004) Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6**: 1-11.

Wickham, H. (2009) *Ggplot2 : elegant graphics for data analysis.* New York: Springer.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples. *PLoS ONE.* **3**: e1456.

Wolfson, J., Dressler, D., and Magazin, M. (1972) Bacteriophage T7 DNA replication: a linear replicating intermediate (gradient centrifugation-electron microscopy-*E. coli*-DNA partial denaturation). *Proc. Natl. Acad. Sci. U. S. A.* **69**: 499-504.

Wommack, K.E., and Colwell, R.R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**: 69-114.

Yerrapragada, S., Siefert, J.L., and Fox, G.E. (2009) Horizontal gene transfer in cyanobacterial signature genes. *Methods Mol. Biol.* **532**: 339-366.

Zdobnov, E.M., and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* **17**: 847-848.

Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., and Béjà, O. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ. Microbiol.* **7**: 1505-1513.

Zeng, Q., and Chisholm, S.W. (2012) Marine Viruses Exploit Their Host's Two-Component Regulatory System in Response to Resource Limitation. *Curr. Biol.* **22**: 124-128.

Tables

Table 1. General features of the cyanopodoviruses from this study, and of those whose genomes have been previously published.

MPP[1]	Phage	Original host	Host Clade[2]	Genome size (kb)	# ORFs	Host %GC content	Phage %GC content	Site of origin	Depth	Lat.	Long.	Date water sampled	Accession #	Reference
MPP-B1	P-SSP11	*Prochlorococcus* MIT9515	HL(I)	47039	54	30.8	39.2	BATS	100	31°48'N	64°16'W	1-Sep-99	HQ634152	This study
	P-SSP10	*Prochlorococcus* NATL2A	LL(I)	47325	52	35	39.2	BATS	100	31°48'N	64°16'W	5-Jun-96	HQ337022	This study
	P-HP1	*Prochlorococcus* NATL2A	LL(I)	47536	65	35	39.9	HOTS[4]	25m	22° 45'N	158°00'W	8-Mar-06	GU071104	This study
MPP-B2	P-GSP1	*Prochlorococcus* MED4	HL(I)	44945	53	30.8	39.6	Gulf Stream	80	38°21'N	66°49'W	Aug-95	HQ332140	This study
	P-SSP7	*Prochlorococcus* MED4	HL(I)	44970	54	30.8	38.8	BATS[5]	100	31°48'N	64°16'W	1-Sep-99	NC_006882	(Sullivan et al., 2005)
	P-SSP3	*Prochlorococcus* MIT9312	HL(II)	46198	56	31.2	37.9	BATS	100	31°48'N	64°16'W	31-Aug-95	HQ332137	This study
	P-SSP2	*Prochlorococcus* MIT9312	HL(II)	45890	59	31.2	37.9	BATS	120	31°48'N	64°16'W	31-Aug-95	GU071107	This study
	P-RSP5	*Prochlorococcus* NATL1A	LL(I)	47741	68	35.1	38.7	Red Sea	130	29°28'N	34°55'E	13-Sep-00	GU071102	This study
MPP-A	P-SSP9	*Prochlorococcus* SS-120	LL(II)	46997	53	36.4	40.5	BATS	100	31°48'N	64°16'W	31-Aug-95	HQ316584	This study
	SYN5	*Synechococcus* WH8109	Syn.	46214	61	60.1	55	Sargasso Sea	Surface	36°58'N	73°42'W	30-Nov-86	NC_009531	(Pope et al., 2007)
	P60	*Synechococcus* WH7803	Syn.	47872	80	60.2	53.2	Satilla River[6]	Surface	-	-	12-Jul-88	AF338467	(Chen and Lu, 2002)
-	P-RSP2	*Prochlorococcus* MIT9302	HL(II)	42257	48	-	34	Red Sea	Surface	29°28'N	34°53'E	14-Jul-96	HQ332139	This study
-	Pf-WMP3	*Leptolyngbya foveolarum*	FC[3]	43249	41	-	46.5	Lake Weiming	nd	-	-	22-Jul-03	EF537008.1	(Liu et al., 2008)
-	Pf-WMP4	*Leptolyngbya foveolarum*	FC	40938	55	-	51.8	Lake Weiming	nd	-	-	22-Jul-03	DQ875742.1	(Liu et al., 2007)

[1] Classification of phage genomes based on the concatenated core genes phylogeny. "–" indicates a phage that is not classified in one of the three groups (Fig. 1).
[2] Clade names for *Prochlorococcus* as defined in Rocap et al., (2002)
[3] FC = Freshwater cyanophage
[4] HOTS = Hawaii Ocean Time Series Station
[5] BATS = Bermuda Atlantic Time Series Station
[6] Satilla River: estuary - salinity = 30‰ – See note added in proofs.

Table 2. Host range of some of the cyanopodoviruses reported here. + indicates successful infection; - indicates no infection. Clade designations for Prochlorococcus refer to light adaptation properties of host cells as defined in Rocap and colleagues (2002). [Correction added on 29 January 2013 after first online publication: P-SSP3 and P-SSP10 were removed from Table 2 as irregularities were detected in the lysates after publication. This does not affect the genomic data or any of the conclusions of the pape

		Phage					
Host strains tested	Host clade	P-SSP7	P-GSP1	P-HP1	P-RSP5	P-RSP2	P-SSP11
Prochlorococcus MIT9302	HL(II)	-	-	-	-	+	-
*Prochlorococcus MIT*9312	HL(II)	-	-	-	-	-	-
Prochlorococcus MIT9215	HL(II)	-	-	-	-	-	-
Prochlorococcus GP2	HL(II)	-	-	-	-	-	-
Prochlorococcus MIT9202	HL(II)	-	+	-	-	-	-
Prochlorococcus AS9601	HL(II)	-	-	-	-	-	-
Prochlorococcus MIT9301	HL(II)	-	-	-	-	-	-
Prochlorococcus MED4	HL(I)	+	+	-	-	-	-
Prochlorococcus MIT9515	HL(I)	-	-	-	-	-	+
Prochlorococcus NATL2A	LL(I)	-	-	+	+	-	-
Prochlorococcus NATL1A	LL(I)	-	-	-	+	-	-
Prochlorococcus MIT9313	LL(IV)	-	-	-	-	-	-

Table 3. Relatively conserved genes in cyanopodoviruses. Core genes of marine cyanopodoviruses are shown in bold. Classes of genes are as defined for P-SSP7 by Lindell et al. (2007), depicting the order of the timing of their transcription (see Fig. 3). Class II-b genes, which include *talC,* are transcribed with Class II genes, even though they are positioned at the end of the genome (Lindell et al., 2007)

		Marine cyanopodoviruses												**Freshwater Cyano T7-like phage**	
Gene Class	**Putative Function**	**P-SSP7**	**P-SSP2**	**P-SSP3**	**P-GSP1**	**P-HP1**	**P-RSP5**	**P-SSP11**	**P-SSP10**	**Syn5**	**P-SSP9**	**P-RSP2**	**P60 ***	**Pf-WMP3**	**Pf-WMP4**
Class II	RNA polymerase	gp13	gp29	gp42	gp11	gp51	gp28	gp54	gp29	gp15	gp6	-	gp6	-	-
	SSB	gp14	gp30	gp41	gp10	gp50	gp26	gp53	gp28	gp21	gp1	gp47	-	-	-
	Endonuclease	**gp15**	**gp31**	**gp40**	**gp9**	**gp49**	**gp25**	**gp52**	**gp26**	**gp22**	**gp52**	**gp46**	**gp16-17**	-	**gp17**
	Primase/Helicase	**gp16**	**gp32**	**gp39**	**gp8**	**gp48**	**gp24**	**gp51**	**gp25**	**gp24**	**gp50**	**gp45**	**gp18**	**gp9**	**gp12**
	DNA polymerase	**gp17**	**gp34**	**gp38**	**gp7**	**gp46**	**gp23**	**gp50**	**gp24**	**gp27**	**gp49**	**gp44**	**gp20**	**gp12-14**	**gp19**
	Exonuclease	**gp19**	**gp35**	**gp37**	**gp6**	**gp44**	**gp22**	**gp49**	**gp23**	**gp29**	**gp47**	**gp42**	**gp21**	-	-
	Rnr	gp20	gp38	gp35	gp4	gp41	gp19	gp46	gp20	gp33	gp44	gp40	-	-	-
	gp34	gp21	gp39	gp34	gp3	gp40	gp18	gp45	gp19	gp34	gp43	gp39	-	-	-
	-	**gp22**	**gp40**	**gp33**	**gp52**	**gp39**	**gp17**	**gp44**	**gp18**	**gp35**	**gp42**	**gp37**	**gp28-43**	-	-
	Portal	**gp24**	**gp42**	**gp31**	**gp50**	**gp37**	**gp13**	**gp42**	**gp16**	**gp37**	**gp40**	**gp35**	**gp41**	-	-
	Scaffolding protein	**gp25**	**gp43**	**gp30**	**gp49**	**gp36**	**gp11**	**gp41**	**gp15**	**gp38**	**gp38**	**gp33**	**gp38-39**	-	-
	Hli	gp26	gp44	gp29	gp48	gp35	gp9	gp40.5	gp14	-	gp38.5	-	-	-	-
	PsbA	gp27	gp46	gp27	gp47	gp34	gp8	gp40	gp13	-	-	-	-	-	-
Class III	**MCP**	**gp29**	**gp48**	**gp25**	**gp46**	**gp29**	**gp5**	**gp36**	**gp8**	**gp39**	**gp37**	**gp32**	**gp37**	**gp32**	-
	Tail tube A	**gp30**	**gp50**	**gp23**	**gp45**	**gp28**	**gp2**	**gp35**	**gp7**	**gp40**	**gp36**	**gp31**	**gp35-36**	-	-
	Tail tube B	**gp31**	**gp51**	**gp22**	**gp44**	**gp27**	**gp1**	**gp33-34**	**gp6**	**gp41**	**gp35**	**gp29**	**gp33-34**	-	-
	-	gp32	gp53	gp20	gp43	gp26	gp68	gp32	gp5	gp42	gp34	-	-	-	-
	Internal core protein	gp35	gp56	gp17	gp39	gp19	gp65	gp27-26	gp2	gp45	gp31	gp26	-	-	-
	Tail fiber	gp36	gp57	gp16	gp38-35	gp16	gp64	gp25	gp1	gp46	gp30-28	gp25	-	-	-
	-	gp43	gp2	gp10	gp32	gp9	gp60	gp20	gp49	-	gp23	gp16	-	-	-
	-	gp45	gp4	gp8	gp30	gp07	gp59	gp19	gp48	-	gp21	gp18	-	-	-
	gp49	**gp47**	**gp6**	**gp7**	**gp26**	**gp5**	**gp56**	**gp17**	**gp46**	**gp49**	**gp18**	**gp11**	**gp70**	-	-
	Terminase	**gp51**	**gp10**	**gp3**	**gp21**	**gp1**	**gp51**	**gp13**	**gp48**	**gp60**	**gp14**	**gp9**	**gp54-55**	**gp36**	**gp40**

Table 4. Genes found in Island II (Fig 3, 5) – an island found in all but 3 of the cyanopodoviruses in Fig. 3 – showing whether they have orthologs in host genomes (*Prochlorococcus* and *Synechococcus*), and/or those of cyanomyoviruses.

		Phage									
Cluster name[1]	**Putative function**	**P-GSP1**	**P-HP1**	**P-RSP5**	**P-SSP10**	**P-SSP2**	**P-SSP3**	**P-SSP11**	**P-SSP7**	**Orthologs present in cyanomyoviruses**	**Orthologs present in hosts**
PsbA_COG	PsbA	gp47	gp34	gp9	gp13	gp46	gp27	gp40	gp27	+	+
Hli_COG	Hli	gp48	gp35	gp8	gp14	gp44	gp29	gp40.5	gp26	+	+
gp222_COG	gp222[2]		gp33	gp7	gp12	gp45	gp28	gp39		+	-
gp30_COG	hypothetical protein		gp30	gp33	gp9			gp37		+	-
gp32_COG	hypothetical protein		gp32		gp11			gp38		-	+
gp47_COG	hypothetical protein					gp47	gp26			-	-
orphan	hypothetical protein			gp10						-	-
orphan	hypothetical protein			gp6						-	-
orphan	hypothetical protein				gp10					-	-
orphan	hypothetical protein								gp28	-	-
orphan	hypothetical protein		gp31							-	-

[1] Cluster names refer to the putative function or a phage gene representing the cluster

[2] gp222: conserved hypothetical protein

Table 5. The only genome differences between the most closely related cyanopodoviruses, P-SSP2 and P-SSP3, which were isolated from the same site, on the same host. The remainder of the proteins share ≥95% identity (see also Fig. 4).

Orthologous proteins			
P-SSP2	P-SSP3	% id (aa)	Putative function
gp14	gp55	76.4	Hypothetical protein
gp17	absent	-	Hypothetical protein
gp18	gp52	66.3	Hypothetical protein
gp32	gp39	†	Primase/helicase
gp57	gp16	77.7	Tail fiber
absent	gp15	-	Hypothetical protein
absent	gp14	-	Hypothetical protein

† Frameshifts -- High similarity between the nucleotide sequences

Table 6. Cyanopodovirus genes shared with (A) picocyanobacterial hosts, *Synechococcus* and *Prochlorococcus* ("phage/host share genes"), (B) cyanomyoviruses ("podo/myo shared genes"), or (C) both (phage/host and podo/myo shared genes"). Single-strand binding protein (SSB, bolded) is the only core gene in this set.

			Phage										
	Class[1]	**Putative function**	**P-SSP7**	**P-GSP1**	**P-HP1**	**P-RSP2**	**P-RSP5**	**P-SSP10**	**P-SSP2**	**P-SSP3**	**P-SSP11**	**P-SSP9**	**Syn5**
A	Class I	DNA primase	-	-	-	-	-	-	-	-	-	gp10	-
		RNA polymerase	gp13	gp11	gp51	-	gp28	gp29	gp29	gp42	gp54	gp6	gp15
		gp 0.7[2]	gp11	-	-	-	-	-	gp26	gp44	-	gp4	
		SSB[3]	**gp14**	**gp10**	**gp50**	**gp47**	**gp26**	**gp28**	**gp30**	**gp41**	**gp53**	**gp1**	**gp21**
	Class II	Unknown	-	-	gp32	-	-	gp11	-	-	gp38	-	-
	Class III	Unknown	-	-	-	-	gp41	-	-	-	-	-	-
		Unknown	-	-	gp65	-	gp48	gp40	-	-	-	-	-
		Thymidylate synthase	-	-	-	-	-	-	-	-	-	-	gp61
B	Class I	Endonuclease	-	-	-	-	-	-	-	-	-	gp48	-
		Unknown	-	-	-	-	-	-	-	-	-	-	gp25
		Unknown	-	-	-	-	gp46	-	-	-	-	-	-
	Class II	gp222[4]	-	-	gp33	-	gp7	gp12	gp45	gp28	gp39	-	-
		Unknown	-	-	gp30	-	gp33	gp9	-	-	gp37	-	-
	Class III	Unknown	gp43	gp32	gp9	-	gp60	gp49	gp2	gp20	-	gp23	-
		Unknown	-	gp30	-	-	-	-	-	gp8	-	-	-
		Unknown	-	gp29	-	-	-	-	-	-	-	-	-
		Endonuclease	-	-	-	gp43	-	-	-	-	-	-	-
		Unknown	-	-	-	-	gp43	-	-	-	-	-	-
		Unknown	-	-	-	-	gp49	-	-	-	-	-	-
		Unknown	-	-	-	-	-	-	-	-	-	-	gp61
C	Class II	Hli	gp26	gp48	gp35	-	gp8	gp14	gp44	gp29	gp40.5	gp38.5	-
		PsbA	gp27	gp47	gp34	-	gp9	gp13	gp46	gp27	gp40	-	-
	Class III	HNH endonuclease	gp49	gp25	gp3	gp10	-	gp44	gp8	gp5	gp15	-	-
	Class IIb	TalC	gp54	gp19	gp2	-	gp50	gp43	gp12	gp1	gp14	-	-

[1] Class of genes as defined for P-SSP7 by Lindell *et al.* (2007), according to the timing of their transcription
[2] gp 0.7: transcriptional regulator
[3] Core gene, SSB: Single Strand Binding protein.
[4] gp222: conserved hypothetical protein.

Supplemental table 1. Cyanophage and picocyanobacterial genomes used for the protein clustering analysis.

Bacterial or viral strain	Genome size (kb)	NCBI accession number	CAMERA accession number
Cyanophage			
S-TIM5	152.3	JQ245707.1	-
Syn33	174.3	GU071108.1	BROADPHAGEGENOMES_SMPL_SYN33_G1163
syn9	177.3	DQ149023.2	-
MED4-213	181.0		CAM_SMPL_001226
P-HM1	181.0	GU071101.1	BROADPHAGEGENOMES_SMPL_P-HM1_G1154
P-HM2	183.8	GU075905.1	BROADPHAGEGENOMES_SMPL_P-HM2_G1155
P-RSM1	177.2		CAM_SMPL_001227
P-RSM3	178.8		CAM_SMPL_001229
P-RSM4	176.4	GU071099.1	BROADPHAGEGENOMES_SMPL_P-RSM4_G1161
P-SSM2	252.4	GU071092.1	-
P-SSM3	179.1		CAM_SMPL_000950
P-SSM4	178.2	AY940168.2	CAM_SMPL_000897
P-SSM5	252.0		CAM_SMPL_000949
P-SSM7	182.2	GU071103.1	BROADPHAGEGENOMES_SMPL_P-SSM7_G1169
S-PM2	196.3	AJ630128.1	-
P-RSM6	192.5		CAM_SMPL_001230
S-RSM4	194.5	FM207411.1	-
S-SM1	174.1	GU071094.1	BROADPHAGEGENOMES_SMPL_S-SM1_G1061
S-SM2	190.8	GU071095.1	BROADPHAGEGENOMES_SMPL_S-SM2_G1159
S-SSM4	182.8		CAM_SMPL_000897
S-SSM5	176.2	GU071097.1	BROADPHAGEGENOMES_SMPL_S-SSM5_G1166
S-SSM7	232.9	GU071098.1	BROADPHAGEGENOMES_SMPL_S-SSM7_G1167
S-ShM2	179.6	GU071096.1	BROADPHAGEGENOMES_SMPL_S-SHM2_G1164
Syn1	191.2	GU071105.1	BROADPHAGEGENOMES_SMPL_SYN1_G1160
Syn10	177.1		CAM_SMPL_001202
Syn19	175.2	GU071106.1	BROADPHAGEGENOMES_SMPL_SYN19_G1165
Syn2	175.6		CAM_SMPL_001201
Syn30	178.8		CAM_SMPL_001200
P60	47.9	AF338467.1	-
P-SSP7	45.0	AY939843.2	-
P-RSP5	47.7	GU071102.1	BROADPHAGEGENOMES_SMPL_NATL1A-7_G1172
P-SSP2	45.9	GU071107.1	-
P-SSP9	47.0	GU071104.1	CAM_SMPL_000899
P-SSP10	47.3		
P-GSP1	44.9		CAM_SMPL_000948
P-HP1	47.5	GU071104.1	BROADPHAGEGENOMES_SMPL_NATL2A-133_G1171
P-SSP11	47.0		CAM_SMPL_000947
Syn5	46.2	EF372997.1	-
P-RSP2	42.3		CAM_SMPL_000945

P-SSP3	47.1		CAM_SMPL_000946
MED4-184	38.3		CAM_SMPL_001191
MED4-117	38.8		CAM_SMPL_001190
P-SS2	107.5	GQ334450.1	-
Cyanobacteria			
Prochlo. MED4	1657	BX548174	-
Prochlo. MIT9313	2410	BX548175	-
Prochlo. MIT9303	2682	CP000554	-
Prochlo. NATL1A	1864	CP000553	-
Prochlo. NATL2A	1842	CP000095	-
Prochlo. AS9601	1669	CP000551	-
Prochlo. MIT9515	1704	CP000552	-
Prochlo. MIT9215	1738	CP000825	-
Prochlo. MIT9211	1688	CP000878	-
Prochlo. MIT9312	1709	CP000111	-
Prochlo. SS120	1751	AE017126	-
Prochlo. MIT9301	1641	CP000576	-
Synecho. CC9311	2606	CP000435	-
Synecho. CC9605	2510	CP000110	-
Synecho. CC9902	2234	CP000097	-
Synecho. WH8102	2434	BX548020	-
Synecho. WH7803	2366	CT971583	-
Synecho. RCC307	2224	CT978603	-
Synecho. WH7805	2620	AAOK00000000	-
Prochlo. MIT9202	1691		MF_SMPL_P9202
Synecho. BL107	2283	AATZ00000000	-
Synecho. RS9917	2579	AANP00000000	-
Synecho. RS9916	2664	AAUA00000000	-
Synecho. WH5701	3043		MF_SMPL_WH5701

Supplemental table 2. Cyanophage reference genomes used for the metagenomic read recruitment.

Phage	Phage family	Genome size (kb)	NCBI accession number	CAMERA sample accession number
S-TIM5	*Myo.*	152.3	JQ245707.1	-
Syn33	*Myo.*	174.3	GU071108.1	BROADPHAGEGENOMES_SMPL_SYN33_G1163
syn9	*Myo.*	177.3	DQ149023.2	-
MED4-213	*Myo.*	181.0		CAM_SMPL_001226
P-HM1	*Myo.*	181.0	GU071101.1	BROADPHAGEGENOMES_SMPL_P-HM1_G1154
P-HM2	*Myo.*	183.8	GU075905.1	BROADPHAGEGENOMES_SMPL_P-HM2_G1155
P-RSM1	*Myo.*	177.2		CAM_SMPL_001227
P-RSM3	*Myo.*	178.8		CAM_SMPL_001229
P-RSM4	*Myo.*	176.4	GU071099.1	BROADPHAGEGENOMES_SMPL_P-RSM4_G1161
P-SSM2	*Myo.*	252.4	GU071092.1	-
P-SSM3	*Myo.*	179.1		CAM_SMPL_000950
P-SSM4	*Myo.*	178.2	AY940168.2	CAM_SMPL_000897
P-SSM5	*Myo.*	252.0		CAM_SMPL_000949
P-SSM7	*Myo.*	182.2	GU071103.1	BROADPHAGEGENOMES_SMPL_P-SSM7_G1169
S-PM2	*Myo.*	196.3	AJ630128.1	-
P-RSM6	*Myo.*	192.5		CAM_SMPL_001230
S-RSM4	*Myo.*	194.5	FM207411.1	-
S-SM1	*Myo.*	174.1	GU071094.1	BROADPHAGEGENOMES_SMPL_S-SM1_G1061
S-SM2	*Myo.*	190.8	GU071095.1	BROADPHAGEGENOMES_SMPL_S-SM2_G1159
S-SSM4	*Myo.*	182.8		CAM_SMPL_000897
S-SSM5	*Myo.*	176.2	GU071097.1	BROADPHAGEGENOMES_SMPL_S-SSM5_G1166
S-SSM7	*Myo.*	232.9	GU071098.1	BROADPHAGEGENOMES_SMPL_S-SSM7_G1167
S-ShM2	*Myo.*	179.6	GU071096.1	BROADPHAGEGENOMES_SMPL_S-SHM2_G1164
Syn1	*Myo.*	191.2	GU071105.1	BROADPHAGEGENOMES_SMPL_SYN1_G1160
Syn10	*Myo.*	177.1		CAM_SMPL_001202
Syn19	*Myo.*	175.2	GU071106.1	BROADPHAGEGENOMES_SMPL_SYN19_G1165
Syn2	*Myo.*	175.6		CAM_SMPL_001201
Syn30	*Myo.*	178.8		CAM_SMPL_001200
P60	*Podo.*	47.9	AF338467.1	-
P-SSP7	*Podo.*	45.0	AY939843.2	-
P-RSP5	*Podo.*	47.7	GU071102.1	BROADPHAGEGENOMES_SMPL_NATL1A-7_G1172
P-SSP2	*Podo.*	45.9	GU071107.1	-
P-SSP9	*Podo.*	47.0	GU071104.1	CAM_SMPL_000899
P-SSP10	*Podo.*	47.3		-
P-GSP1	*Podo.*	44.9		CAM_SMPL_000948
P-HP1	*Podo.*	47.5	GU071104.1	BROADPHAGEGENOMES_SMPL_NATL2A-133_G1171
P-SSP11	*Podo.*	47.0		CAM_SMPL_000947
Syn5	*Podo.*	46.2	EF372997.1	-

P-RSP2	*Podo.*	42.3		CAM_SMPL_000945
P-SSP3	*Podo.*	47.1		CAM_SMPL_000946
MED4-184	*Sipho.*	38.3		CAM_SMPL_001191
MED4-117	*Sipho.*	38.8		CAM_SMPL_001190
P-SS2	*Sipho.*	107.5	GQ334450.1	-
S-CBS1	*Sipho.*	53.7	HM480106.1	-
S-CBS2	*Sipho.*	73.5	GU936714.1	-
S-CBS3	*Sipho.*	28.0	GU936715.1	-
S-CBS4	*Sipho.*	62.9	HQ698895.1	-

FIGURES

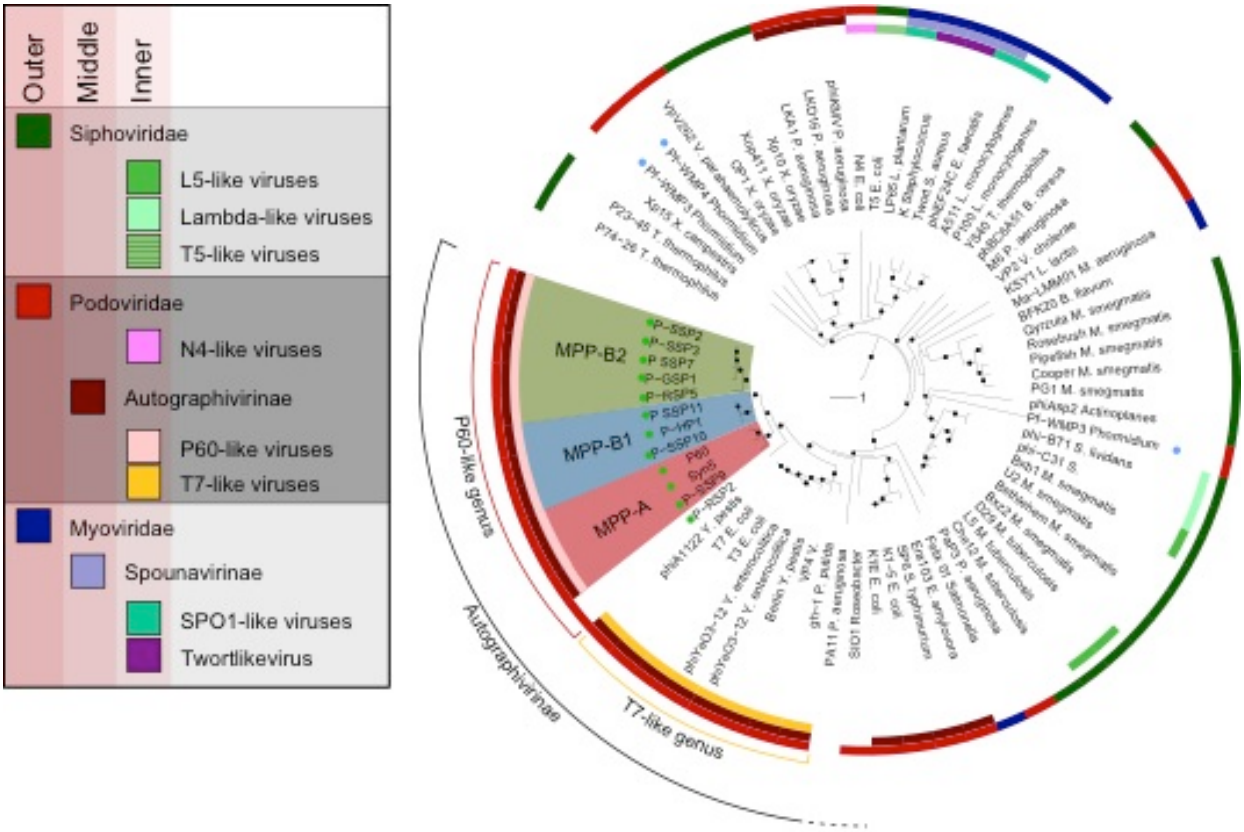

Figure 1. Maximum likelihood, circular phylogenetic tree of phage DNA polymerase sequences retrieved from ACLAME database (ACLAME MGEs. Version 0.4 - family_vir_ 14 (Leplae et al., 2009)). The bar represents 1 amino acid substitution per site and branches with a bootstrap value greater than 80% are indicated by a black dot. Green dots indicate marine cyanopodoviruses while the three blue dots mark DNA polymerase genes from the two freshwater cyanopodoviruses, one of which encodes DNA polymerase with two genes. The outer, middle and inner rings respectively indicate the phage families, subfamilies and genus when available in NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy).

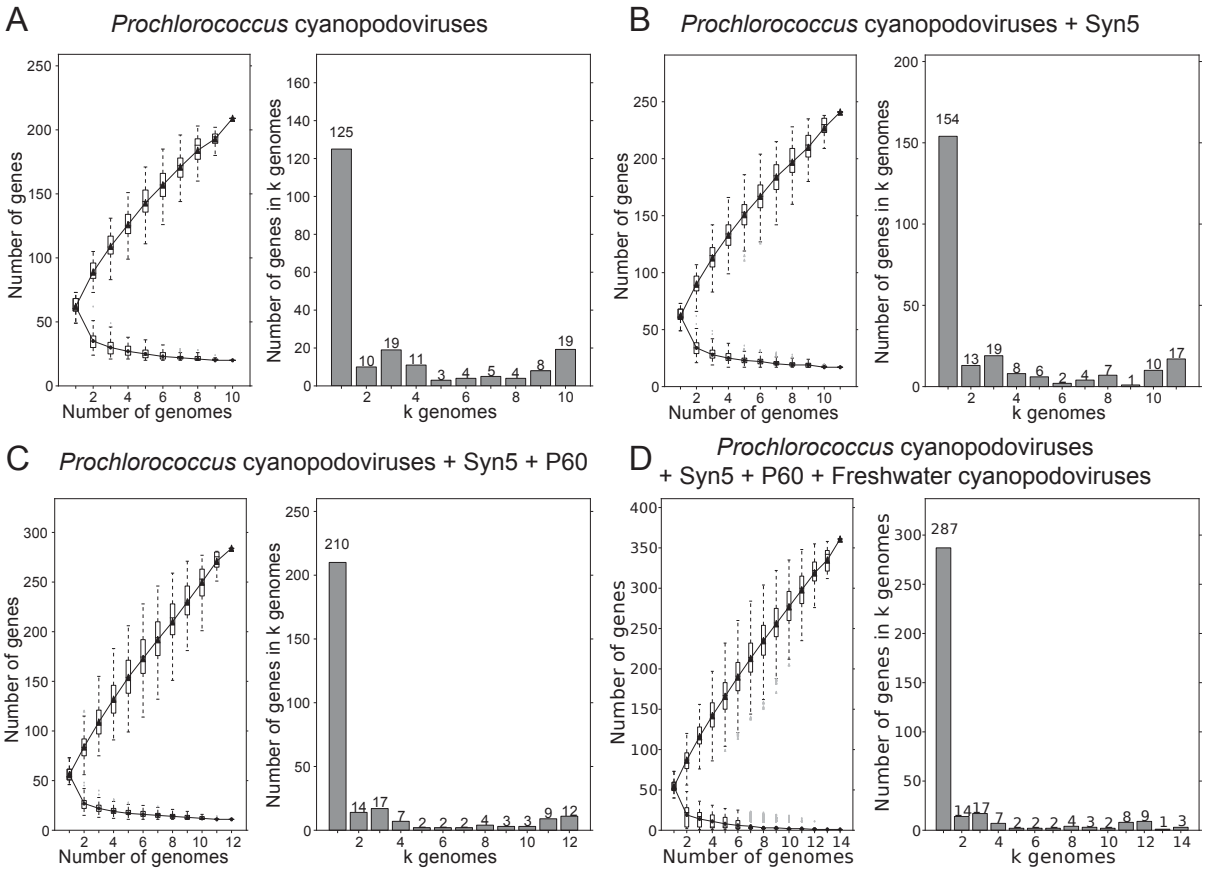

Figure 2. Core and pan-genome analysis using different sets of phage genomes in the analysis, as indicated by the headers in A-D. Left panel in each pair: Number of total genes in the core- (circles) and pan- (triangles) genomes as a function of the number of genomes included in the analysis. The core genome is the set of genes shared by all the genomes included in the analyzed subset, while the pan-genome is the total number of unique genes found in the same subset. All possible combinations of genomes were analyzed; the line is drawn through the average. Right panel in each pair: The frequency distribution of genes among the genomes, showing that genes found in only one (k=1) of the genomes are the most common (See note added in proofs for panel C)

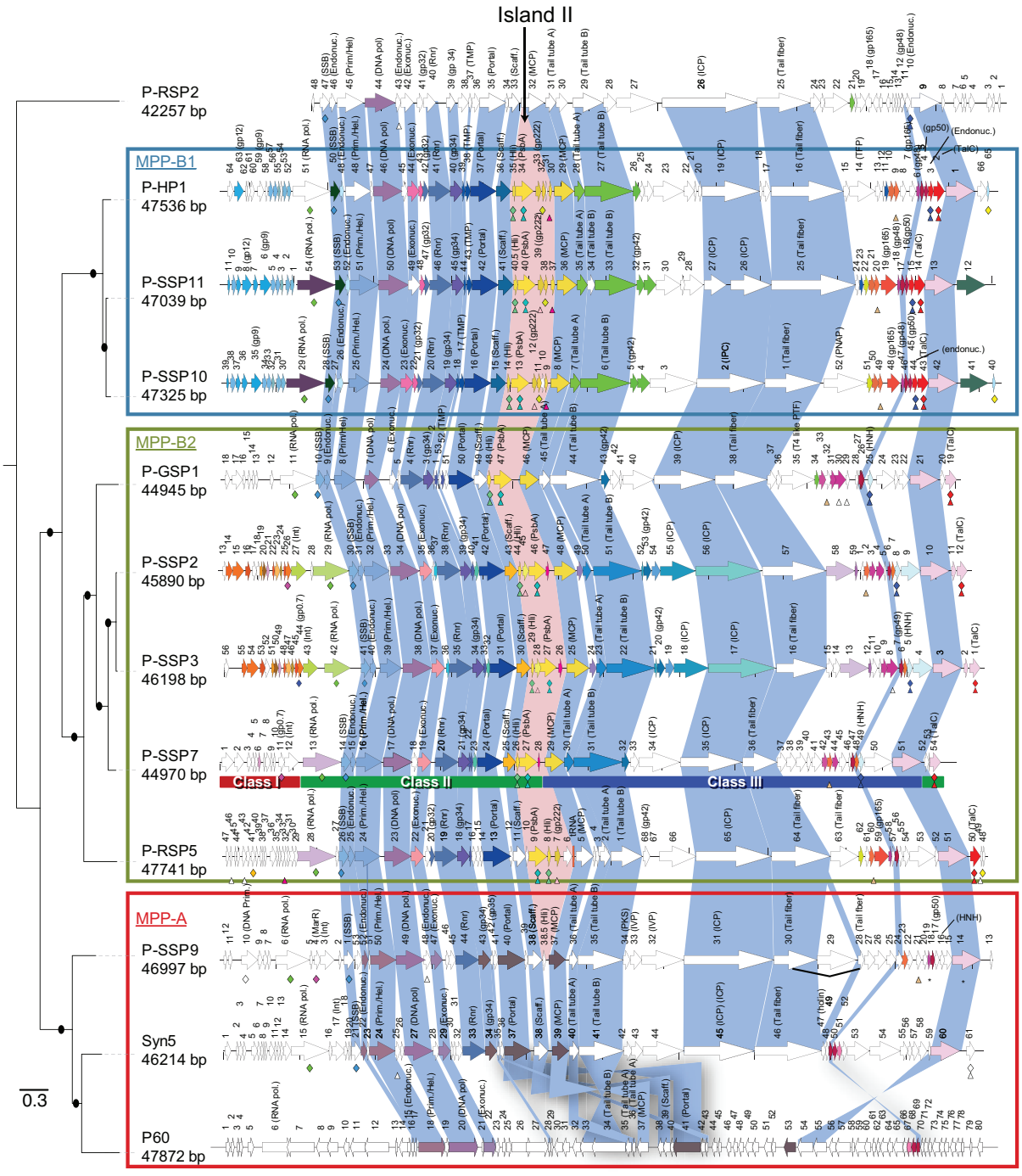

Figure 3. Alignment of the genomes of 12 cyanopodoviruses. Orthologous proteins represented in color other than white share 60% amino acid identity or more, while those shown in white do not. The core proteins shared by all cyanopodoviruses are linked by blue shading and genomic Island II (see Fig. 5) is highlighted by pink shading. Cyanopodovirus/host shared proteins and cyanopodovirus/cyanomyovirus shared proteins are designated by small diamonds and triangles, respectively (see also Fig. 5 & Table 6), and each different cluster is represented by a different color except for singletons that are represented in white. The phylogenetic tree on the left was generated from an alignment of the concatenated core protein sequences using a maximum likelihood method. Branches with a bootstrap value greater than 80% are indicated by a black dot. The phage genomes were classified into three groups based on the concatenated core gene phylogeny of the 12 cyanopodoviruses (Boxes – MPP-A, MPP-B1 and MPP-B2 (MPP: Marine picocyanopodovirus)); P-RSP2 is an outlier based on this analysis. The bar represents 0.3 amino acid substitutions per site. (P60 genomes – See note added in proofs)

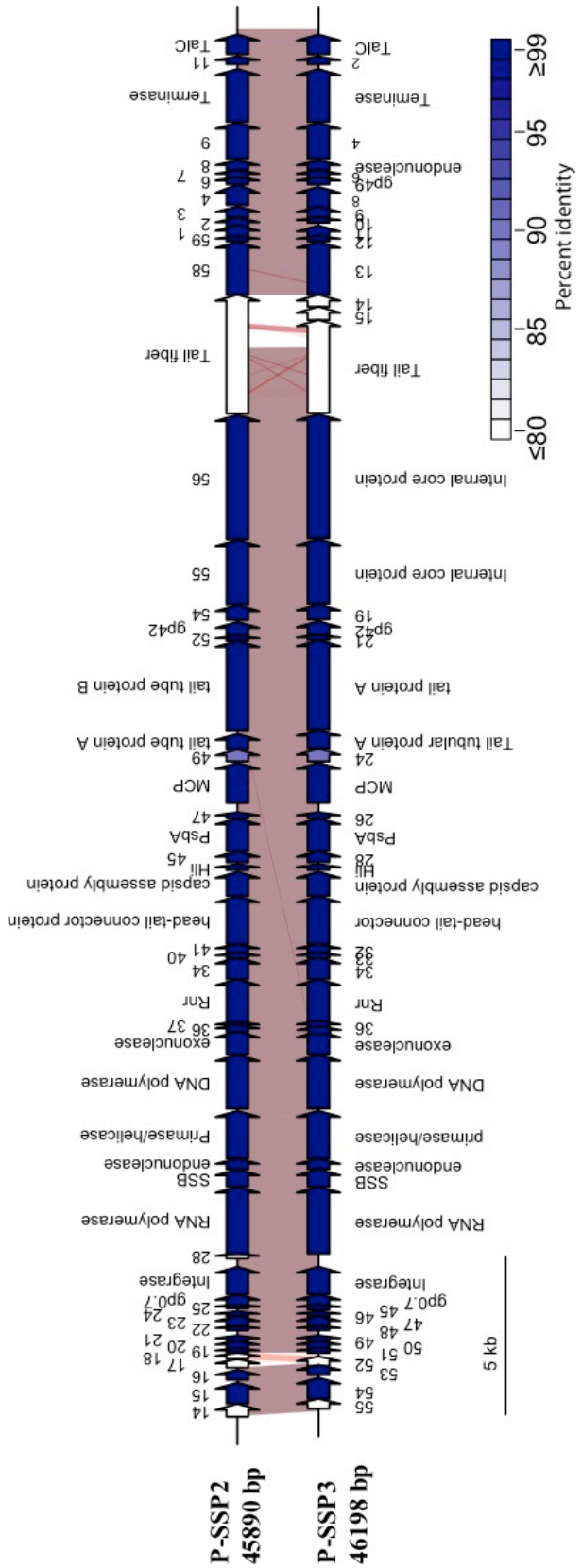

Figure 4. Alignment of the genomes of phage P-SSP2 and P-SSP3. Each gene product was aligned with its homolog and the percent identity was calculated using the length of the longest protein as the denominator. The colors indicate the percent identity between proteins (from 80% to 100%) while the red shading and thin red lines indicate the zone of homology between the DNA sequences where bit score is higher than 40. Proteins in white share less than 80% identity and are reported in Table 5.

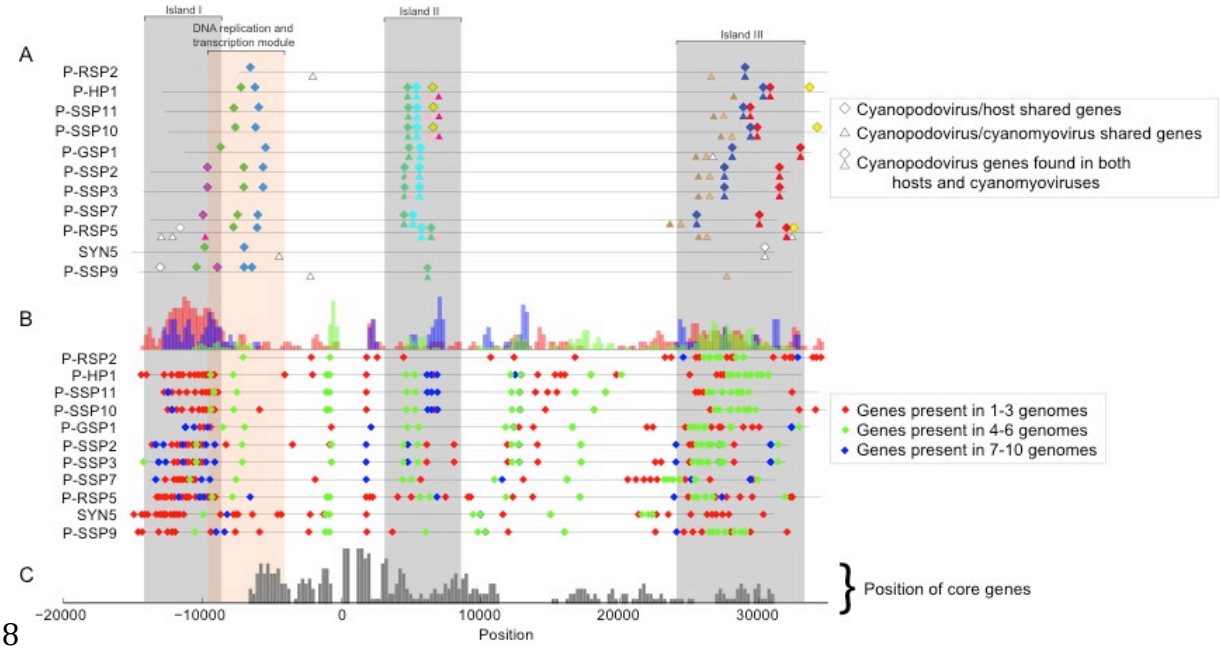

Figure 5. A) Position of cyanopodovirus/host shared genes (diamonds) and cyanopodovirus/cyanomyovirus shared genes (triangles) in cyanopodovirus genomes (symbols are positioned in the middle of the genes). The position of the genes is relative to the position (marked as 0) of the ribonucleotide reductase genes (*rnr*). When a diamond and a triangle co-localize, the cyanopodovirus gene is shared by both host and cyanomyovirus genomes. Orthologs determined using OrthoMCL are represented in the same color. Singletons are shown in white. B) Position of flexible genes (Fig. 2B) in the genomes, according to their frequency distribution (see Fig. 2) Red diamonds indicate genes shared by 1-3 genomes; green diamonds shared by 4-6 genomes; and blue diamonds shared by 7-10 genomes. The histogram on top indicates the relative counts of genes in the various categories present in overlapping sliding windows of 500 bp. The grey shading indicates apparent genome islands. Island I is identified primarily by the set of the most hypervariable genes, occurring in only a few genomes (red diamonds, panel B), while the other two islands are evident in both panels A and B. The orange shading marks the region of the genome involved in DNA replication and transcription, which is not considered a genomic island as these genes are shared by all branches of the tree of life. C) Relative counts of core genes present in overlapping sliding windows of 500 bp.

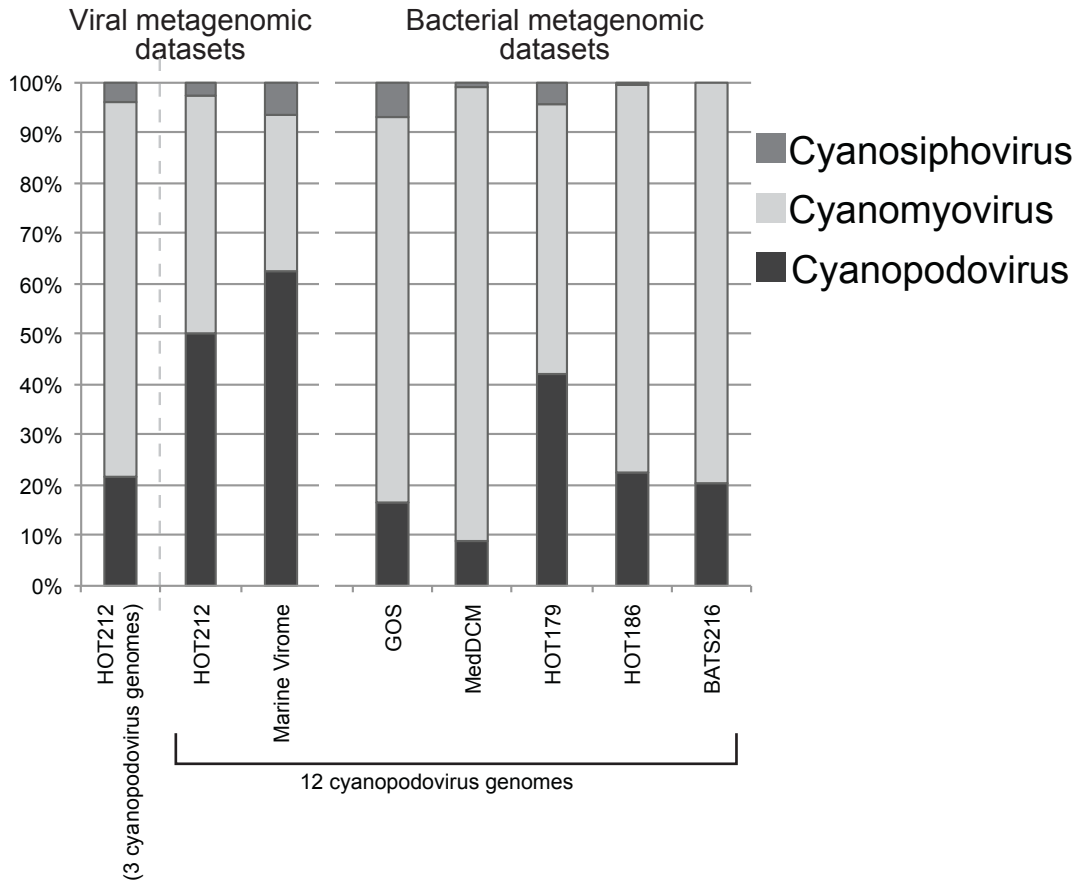

Figure 6. Proportion of reads recruited from different metagenomic datasets by different families of cyanophage. The number of recruited reads was normalized to the average size of the genome of each phage family. “Bacterial metagenomes” refers to viral sequences found in samples that were designed to collect the bacterial fraction; viruses are by-catch. “Viral metagenomes” refers to samples that were collected specifically to capture the viral fraction. For the HOT212 sample, we compare the recruitment proportions obtainedusing the cyanopodovirus genomes extant before this study (3 phage: P-SSP7, Syn5 and P60), and those obtainedusing all marine cyanopodoviruses.

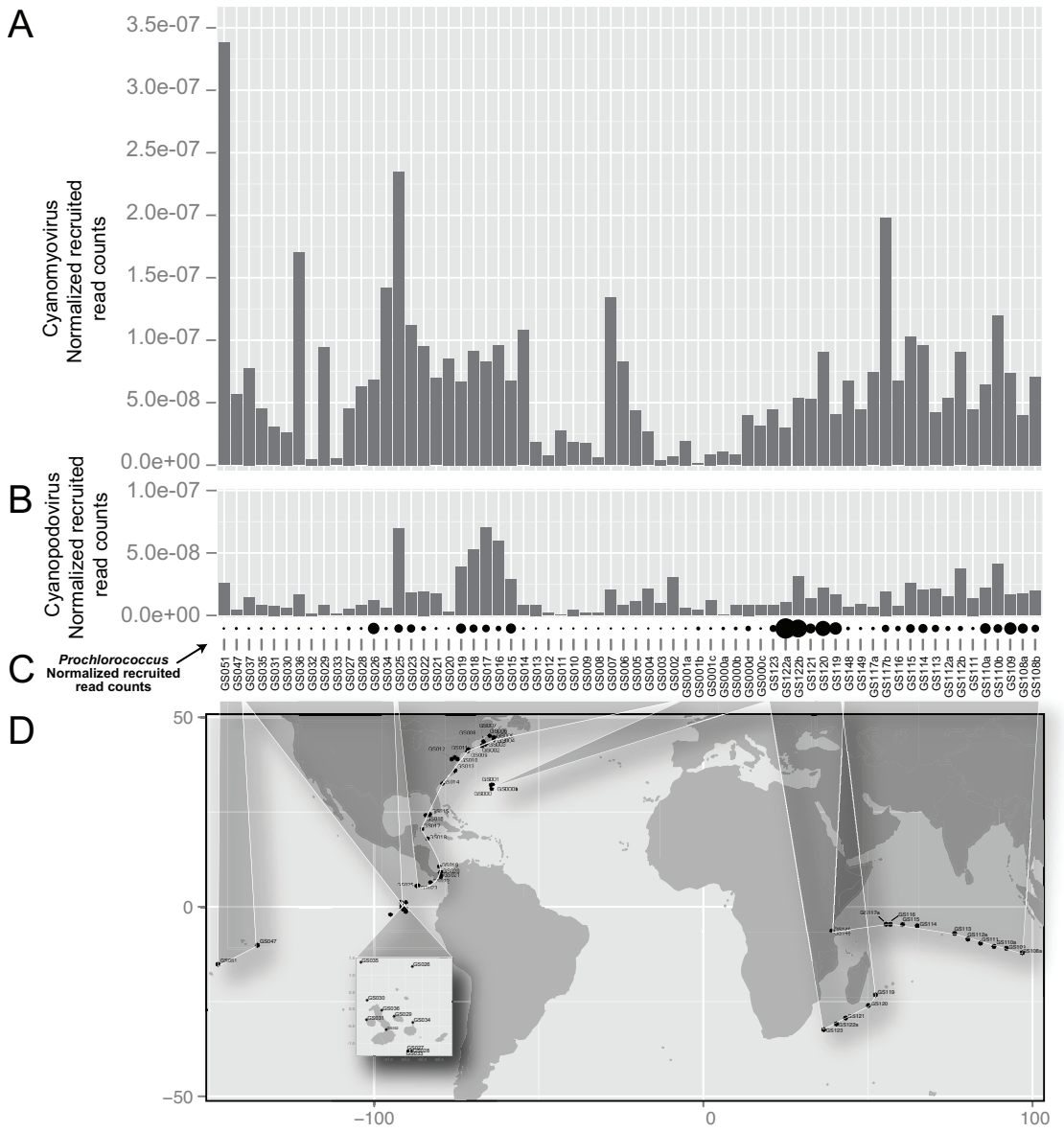

Figure 7. Normalized recruited read counts corresponding to (A) cyanomyoviruses and (B) cyanopodoviruses in the GOS database. Each bar represents a sampling site. The number of reads was normalized to the average size of the genome of each phage family and to the total number of sequencing reads at each of the GOS sites. (C) The relative abundance of *Prochlorococcus* is shown as a series of dots for which the size is proportional to the counts of normalized recruited reads. (D) Map illustrating the position of the GOS sites.

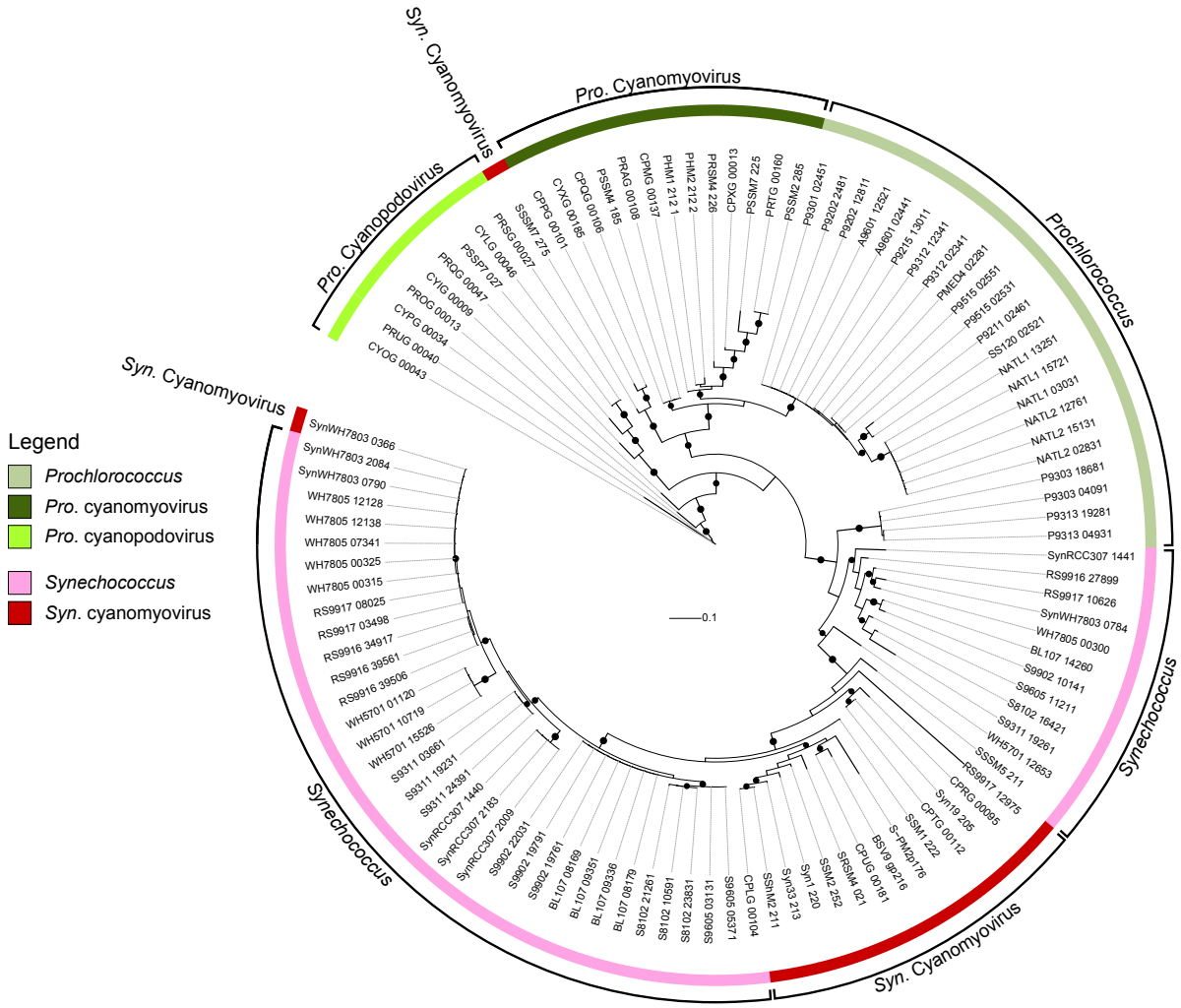

Figure S1. Maximum likelihood, circular phylogenetic tree of PsbA from cyanophage and marine picocyanobacteria ((Kelly et al., 2012) - http://proportal.mit.edu/). The bar represents 0.1 amino acid substitutions per site and branches with a bootstrap value greater than 80% are indicated by a black dot. The ring indicates the origin of PsbA sequences.

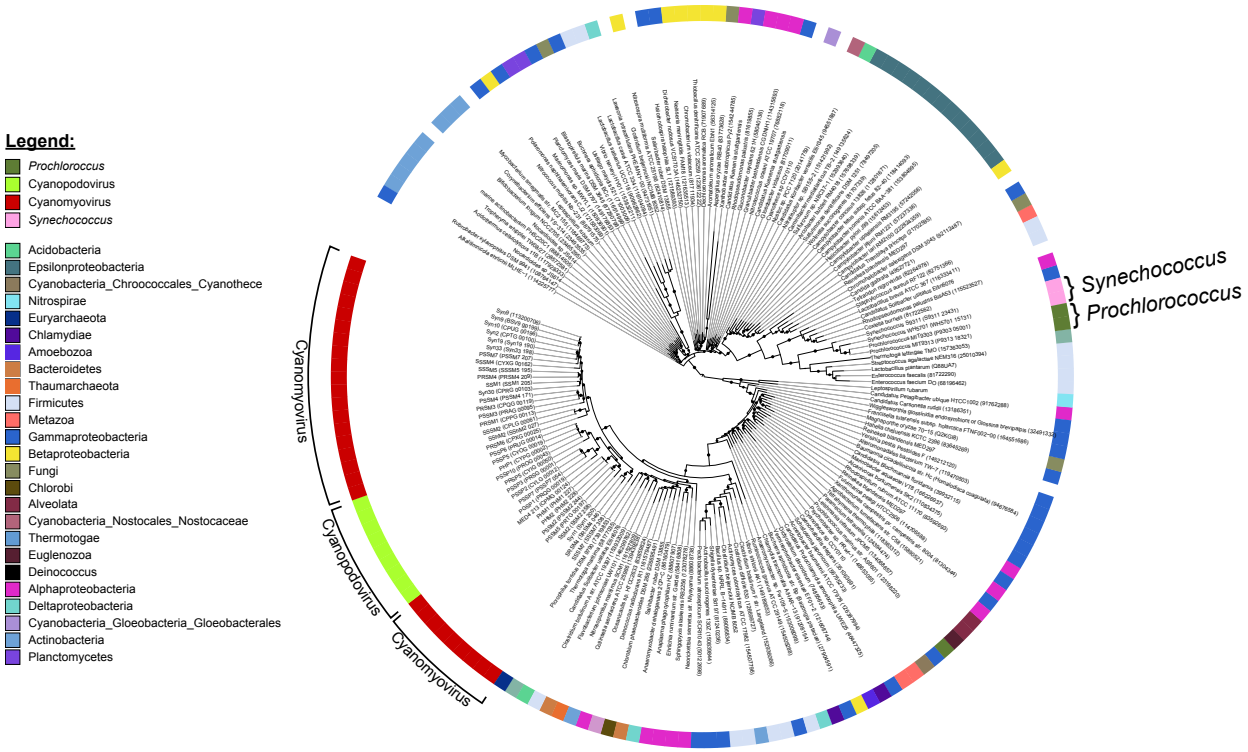

Figure S2. Maximum likelihood, circular phylogenetic tree of cyanopodovirus TalC sequences and orthologous sequences extracted from Pfam family PF00923 (http://pfam.sanger.ac.uk/family/PF00923). The bar represents 0.1 amino acid substitutions per site and branches with a bootstrap value greater than 80% are indicated by a black dot. The ring indicates the origin of TalC sequences.