Virology 499 (2016) 219–229

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses

Lisa T. Crummett [a,*,1,2], Richard J. Puxty [a,1], Claudia Weihe [a], Marcia F. Marston [b], Jennifer B.H. Martiny [a]

[a] Dept. of Ecology and Evolutionary Biology, University of California, Irvine, CA 92612, USA
[b] Dept. of Biology and Marine Biology, Roger Williams University, Bristol, RI 02809, USA

ARTICLE INFO

Article history:
Received 24 May 2016
Returned to author for revisions
16 September 2016
Accepted 17 September 2016
Available online 29 September 2016

Keywords:
Cyanophage
Myoviruses
Auxiliary
Metabolic
Genes
Selection
Genome
Evolution
Marine

ABSTRACT

Viruses of marine cyanobacteria frequently contain auxiliary metabolic genes (AMGs) that augment host metabolism during infection, but little is known about their adaptive significance. We analyzed the distribution and genomic context of 33 AMGs across 60 cyanomyovirus genomes. Similarity in AMG content among cyanomyoviruses was only weakly correlated with phylogenetic relatedness; however, AMG content was generally conserved within the same operational taxonomic unit (OTU). A virus' AMG repertoire was also correlated with its isolation host and environment (coastal versus open ocean). A new analytical method based on shared co-linear blocks revealed that variation in the genomic location of an AMG was negatively correlated with its frequency across the genomes. We propose that rare AMGs are more frequently gained or lost as a result of fluctuating selection pressures, whereas common AMGs are associated with stable selection pressures. Finally, we describe a unique cyanomyovirus (S-CAM7) that lacks many AMGs including the photosynthesis gene *psbA*.

1. Introduction

Marine bacteriophages play a key role in ocean carbon and nutrient cycling via lysis of host cells (Breitbart et al., 2007; Fuhrman, 1999; Suttle, 2005a, 2005b), and their sheer abundance means they themselves may be an important reservoir of dissolved organic phosphorous (Bratbak et al., 1994; Jover et al., 2014). Bacteriophages also influence ocean biogeochemistry by modulating host cell metabolism during infection. These metabolic changes include expression of genes that are carried by phages but originated in bacterial cells (Anantharaman et al., 2014; Hagay et al., 2014; Mann et al., 2003; Sharon et al., 2011; Thompson et al., 2011). Such genes are referred to as auxiliary metabolic genes (AMGs) (Breitbart et al., 2007). Bacteriophages that infect the abundant marine cyanobacteria *Synechococcus* and *Prochlorococcus* (cyanophages) carry AMGs that have been acquired from their immediate host as well as more distantly-related bacteria (Ignacio-Espinoza and Sullivan, 2012; Kelly et al., 2013; Lindell et al., 2004; Millard et al., 2009; Sharon et al., 2009; Sullivan et al., 2010). Analysis of metagenomic samples has shown that AMGs are abundant, diverse, and widespread in the oceans (Williamson et al., 2008).

Cyanophage AMGs are associated with a variety of functions including photosynthesis (Mann et al., 2003; Philosof et al., 2011; Sharon et al., 2011), carbon metabolism (Thompson et al., 2011), nucleic acid synthesis and metabolism (Dwivedi et al., 2013; Hagay et al., 2014), and stress tolerance (He et al., 2001; Kelly et al., 2013; Lindell et al., 2005, 2004). AMGs linked to photosynthesis, such as *psbA* (photosystem II D1 protein), *psbD* (photosystem II D2 protein), and *hli* (highlight-inducible protein) genes, are expressed in cyanobacteria during phage infection (Clokie and Mann, 2006; Lindell et al., 2005) and thus, may make more energy available for phage reproduction (Clokie and Mann, 2006; Hellweger, 2009; Lindell et al., 2005; Millard et al., 2009). Generally, however, the adaptive significance of most cyanophage AMGs remains unclear.

Comparative genomics provides several approaches for generating hypotheses about the adaptive significance, if any, of AMGs in natural cyanophage communities. First, the prevalence of

* Corresponding author at: Dept. of Ecology and Evolutionary Biology, University of California, Irvine, CA 92612, USA.
E-mail addresses: lcrummett@soka.edu (L.T. Crummett), rpuxty@uci.edu (R.J. Puxty), cweihe@uci.edu (C. Weihe), mmarston@rwu.edu (M.F. Marston), jmartiny@uci.edu (J.B.H. Martiny).
[1] Contributed equally to the work.
[2] Current address: Soka University of America, 1 University Drive, Aliso Viejo, CA 92656, USA.

http://dx.doi.org/10.1016/j.virol.2016.09.016
0042-6822/

specific AMGs varies widely among isolate genomes, and these patterns may provide insights into their role during viral infection. For instance, *psbA*, *cobS* (cobalamin synthetase), and *mazG* (pyrophosphatase) genes have been found in all known cyanophages in the family Myoviridae (cyanomyoviruses) and are therefore part of the lineage's core genome. In contrast, two genes encoding electron transporters involved in photosynthesis, *petF* (ferredoxin) and *ptox* (plastoquinol terminal oxidase), are sporadically distributed among these viruses (Sullivan et al., 2010) and are therefore part of the flexible genome, analogous to that observed in lineages of bacteria and archaea (Polz et al., 2013; Tettelin et al., 2008). AMGs that are common among cyanophages may encode metabolic functions that are essential under the range of conditions they experience, whereas less common AMGs may be adaptive for only a subset experiencing a particular set of conditions (microhabitat or host-type) (Cordero and Polz, 2014).

Second, environmental correlations between AMG prevalence among genomes and abiotic parameters at the site of isolation provide clues to the adaptive benefit of particular AMGs (Kelly et al., 2013; Williamson et al., 2008). In ocean metagenomes, the relative abundance of some AMGs was positively correlated with temperature (Williamson et al., 2008). Among cyanophage isolates, the relative abundances of *pstS* (phosphate-binding protein) and *phoA* (alkaline phosphatase) genes were higher in those originating from regions with lower phosphate concentrations (Kelly et al., 2013), suggesting that the genes are associated with phosphate stress. Thus, environmental variables like temperature and nutrient availability may select for AMG content (the number and identity of AMGs) in cyanophages.

Finally, the genomic context of AMGs may shed light on their evolution. AMGs that confer an advantage in all circumstances may be located in highly conserved genomic regions. Alternatively, AMGs that are only adaptive under specific conditions, and therefore subject to higher rates of gain and loss, may be found in highly variable genomic regions. Indeed, previous work on cyanomyoviruses revealed a hypothesized mobile gene cassette containing four carbon metabolism AMGs (*ptox*, *petE*, *zwf*, and *gnd*, encoding plastoquinol terminal oxidase, plastocyanin, glucose 6-phosphate dehydrogenase, and 6-phosphogluconate dehydrogenase, respectively). These genes are sporadically distributed across the genomes and located in a hypervariable region between *g16* and *g17* (Millard et al., 2009; Sullivan et al., 2010). Such hypervariable regions have been identified in other T4-like phages. The regions are composed primarily of genes of unknown function, but also contain genes implicated in phage adaptation to the host (Comeau et al., 2007).

To investigate the potential adaptive role and evolution of AMGs in marine cyanomyoviruses, we examined patterns of AMG content and genomic context across 60 genomes isolated from various regions and hosts, 25 of which were newly sequenced. In contrast to previous analyses, we sequenced genomes that were both closely related and genetically diverged, allowing us to examine the degree of AMG conservation at both fine and broad phylogenetic resolution. To date, genomic comparison of AMGs has mainly focused on distantly-related viral isolates from various locations, but slight differences in AMG content between isolates with nearly identical genomes may reveal incremental adaptation to local environmental conditions and/or hosts (Petrov et al., 2010). To target closely-related genomes, we selected several isolates within previously-defined operational taxonomic units or OTUs (Marston and Sallee 2003; Clasen et al., 2013). Previous work indicates that cyanomyovirus OTUs defined by the similarity of *g20* sequences represent discrete natural populations (Deng et al., 2014; Marston and Amrich, 2009) that display seasonal and spatial biogeographic patterns (Marston et al., 2013).

We focus here on three questions: (1) At what phylogenetic scale do we find significant differences in AMG content? For instance, do members of the same operational taxonomic unit (OTU) vary in AMG content? (2) Is AMG content correlated with environmental parameters such as geography or host type, each of which could impose a selective filter on AMG content? (3) Is the genomic context of an AMG related to its frequency among the genomes, perhaps providing insight into its adaptive significance?

2. Results and discussion

2.1. Phylogenetic conservation of AMG content

We focused on the presence of 33 different AMGs (Table S1) in 60 cyanomyovirus genomes (Table 1). Among the 60 genomes, 36 OTUs (defined as $\geq$99% *g20* nucleotide similarity), were represented, 14 of which included at least two representative genomes. The genomes of isolates within an OTU had an average nucleotide identity (ANI) value of 92.6–99.9% ($\bar{x}$=98.8%) across the entire genome and 98.4–100% ($\bar{x}$=99.8%) across seven core cyanomyoviral genes (Table 1). The 33 AMGs were chosen based on prior studies (Sullivan et al., 2010), and the ability for unambiguous annotation across diverse taxa. The total number of AMGs in a genome ranged from 15 to 26 ($\bar{x}$=17.32) (Table 1), excluding two S-CAM7 isolates that we will discuss below as outliers. The frequency of an AMG across the genomes varied greatly, from only 2 occurrences of *pyrE*, *purN*, and *purH* to 100% of the genomes for *phoH*, *cobS*, *heat shock protein*, *mazG*, and *hli* (including the S-CAM7 genomes).

The similarity in AMG content (presence/absence) between representatives of any two OTUs (n=36) was weakly positively correlated (RELATE test, rho=0.354, p=0.001) with their phylogenetic similarity based on a core gene phylogeny (Fig. S1). However, AMG content among members of the same clade was still quite variable (Fig. 1a). For instance, S-RIM44 and S-RIM32 are representatives of two closely-related OTUs (Fig. S1), and yet these isolates differ in the presence of seven AMGs.

At a finer phylogenetic scale, AMG content was conserved within an OTU for the vast majority of AMGs (30 out of 33). Still, three of the 14 OTUs that had multiple representatives showed variation in AMG content among those representatives. For instance, viral isolate S-RIM44 (W2-07-0710), collected from the coastal waters of Rhode Island in 2010, belongs to the same OTU as viral isolate Syn1, collected from Woods Hole, MA in 1990, and yet they differed in the presence of phosphoribosylaminoimidazole synthetase (*purM*) (Fig. 2a), an enzyme involved in purine biosynthesis. A comparison of the genetic context of the region containing *purM* in viral isolate S-RIM44 with viral isolate Syn1 suggests that there has either been a deletion of *purM* in Syn1 or an insertion into S-RIM44 given high conservation of nucleotide identity and gene order in the upstream and downstream regions (Fig. 2a). The *purM* homolog in S-RIM44 was found on a fragment containing a *hyp-purM-hyp-nrdC* (glutaredoxin) cluster of genes. We searched for this gene cluster in other cyanomyoviruses within this dataset to possibly identify a source for recombination, but were unsuccessful. However, a distantly related *Synechococcus* virus, Syn33, also possessed a copy of *purM* downstream of the *purC-hyp-hyp* gene cluster that was shared in viral isolates S-RIM44 and Syn1, albeit in a disparate location (Fig. 2a). Thus, the presence of *purM* in this context is not unprecedented and suggests that S-RIM44 may have acquired *purM* by homologous recombination with a divergent virus.

A second example of within-OTU variability was S-CAM9, for which two of the three isolates possessed the 6-phosphogluconate dehydrogenase (*gnd*) gene (Fig. 2b). This difference appeared to be the result of a deletion of *gnd* in S-CAM9 0908SB82. The isolate

Table 1
General genomic features of 60 cyanomyoviruses: The isolate name and accession number of newly sequenced genomes is shown in bold font and isolates belonging to the same OTU are grouped in the same row. In the "Host" column, "Syn" refers to *Synechococcus* and "Pro" refers to *Prochlorococcus* and the term directly following the genus is the specific strain. "GP % ID" refers to genomic pairwise percent identity, which is the mean percent identity among isolates of the same OTU based on the entire genome and "CGP % ID" refers to core gene pairwise percent identity, which is the mean percent identity among isolates of the same OTU based on seven conserved core cyanomyovirus genes. "CDS" refers to the total number of annotated coding sequences, "tRNAs" refers to the total number of transfer RNAs, and "AMGs" refers to the number of annotated auxiliary metabolic genes per genome. Within the "Isolation location" column, CA=California (USA), WA=Washington (USA), RI=Rhode Island (USA), MA=Massachusetts (USA), OTC=Ocean Time Series. *Note that the two S-CAM8 genomes from CA have a genomic pairwise % ID of 99.6% (core gene % ID is 99.5%) whereas the S-CAM8 genome from WA has a genomic pairwise % ID of only 93.4% (core gene % ID is 97.5%) with either of the S-CAM8 genomes from CA.

Published name	Host	Isolation date	Isolation location	Genome size (kb)	GP % ID	CGP % ID	CDS	%G+C	tRNAs	AMGs	Accession# (INSDC)
S-CAM1 0208SB26	Syn-WH7803	2/15/2008	Southern CA Pacific Ocean	198.01	99.6	99.9	253	43	8	17	HQ634177.1
S-CAM1 0309SB33	Syn-WH7803	3/11/2009	Southern CA Pacific Ocean	197.54	99.6	99.9	253	43	8	17	**KU686192**
S-CAM1 0310NB17	Syn-WH7803	3/26/2010	Southern CA Pacific Ocean	197.53	99.6	99.9	253	43	8	17	**KU686193**
S-CAM1 0809CC03	Syn-WH7803	8/27/2009	Southern CA Pacific Ocean	197.26	99.6	99.9	253	43	8	17	**KU686194**
S-CAM1 0810SB17	Syn-WH7803	8/18/2010	Southern CA Pacific Ocean	197.53	99.6	99.9	253	43	8	17	**KU686195**
S-CAM1 0910CC29	Syn-WH7803	9/15/2010	Southern CA Pacific Ocean	197.53	99.6	99.9	253	43	8	17	**KU686196**
S-CAM3 0808SB25	Syn-WH7803	8/15/2008	Southern CA Pacific Ocean	197.84	99.4	99.9	239	41.6	10	15	**KU686197**
S-CAM3 0910TB04	Syn-WH7803	9/20/2010	Tijuana, Mexico Pacific Ocean	197.80	99.4	99.9	239	41.6	10	15	**KU686198**
S-CAM3 1010CC42	Syn-WH7803	10/13/2010	Southern CA Pacific Ocean	198.19	99.4	99.9	240	41.6	10	15	**KU686199**
S-CAM4-0309CC44	Syn-WH7803	3/11/2009	Southern CA Pacific Ocean	191.94	99.3	99.5	238	38.6	8	17	**KU686200**
S-CAM4 0809SB33	Syn-WH7803	8/12/2009	Southern CA Pacific Ocean	191.98	99.3	99.5	238	38.6	8	17	**KU686201**
S-CAM4 1010NB23	Syn-WH7803	10/13/2010	Southern CA Pacific Ocean	191.62	99.3	99.5	237	38.6	8	17	**KU686202**
S-CAM7 0910CC49	Syn-WH7803	9/15/2010	Southern CA Pacific Ocean	216.00	92.6	99.6	268	41.2	4	5	**KU686212**
S-CAM7 0910SB42	Syn-WH7803	9/15/2010	Southern CA Pacific Ocean	214.03	92.6	99.6	265	41.2	7	5	**KU686213**
S-CAM8 0608BI06	Syn-WH7803	6/15/2008	Southern CA Pacific Ocean	171.41	*95.6	*98.4	207	39.3	5	21	HQ634178.1
S-CAM8 0608SB47	Syn-WH7803	6/15/2008	Southern CA Pacific Ocean	171.41	*95.6	*98.4	206	39.3	5	21	JF974299
S-CAM8 0810PA29	Syn-WH7803	8/20/2010	Padilla Bay, WA Pacific Ocean	172.34	*95.6	*98.4	211	39.2	5	19	**KU686203**
S-CAM9 0808SB05	Syn-WH7803	8/15/2008	Southern CA Pacific Ocean	174.81	99.2	99.9	228	39	8	16	**KU686204**
S-CAM9 0908SB82	Syn-WH7803	9/1/2008	Southern CA Pacific Ocean	174.66	99.2	99.9	228	39	8	15	**KU686205**
S-CAM9 1109NB16	Syn-WH7803	11/23/2009	Southern CA Pacific Ocean	174.83	99.2	99.9	229	39	8	16	**KU686206**
S-CAM22 0210CC35	Syn-WH7803	2/17/2010	Southern CA Pacific Ocean	172.23	99.5	99.8	215	39.9	5	19	**KU686207**
S-CAM22 0310NB44	Syn-WH7803	3/17/2010	Southern CA Pacific Ocean	172.40	99.5	99.8	214	39.9	5	19	**KU686208**
S-CAM22 1209TA19	Syn-WH7803	12/09/2009	Tijuana, Mexico Pacific Ocean	172.35	99.5	99.8	214	39.9	5	19	**KU686209**
S-WAM1 0810PA09	Syn-WH7803	8/20/2010	Padilla Bay WA Pacific Ocean	185.10	NA	NA	222	44.7	4	15	**KU686210**
S-WAM2 0810PA29	Syn-WH7803	8/10/2010	Padilla Bay WA Pacific Ocean	186.39	NA	NA	227	41.3	12	17	**KU686211**
S-RIM32 RW-108-0702	Syn-WH7803	July-2002	Narragansett Bay RI North Atlantic	194.44	NA	NA	230	39.9	10	15	**KU594606**
S-RIM50 RW-29-0704	Syn-WH7803	July-2004	Narragansett Bay RI North Atlantic	174.31	NA	NA	227	40.3	8	16	**KU594605**
S-RIM44 W2-07-0710	Syn-WH7803	July-2010	Block Island Sound RI North Atlantic	193.00	98.3	99.7	230	40.6	6	16	**KU594607**
Syn1	Syn-WH8102	10/1/1990	Woods Hole MA North Atlantic	191.20	98.3	99.7	234	40.6	6	15	GU071105.1
S-RIM2 R1-1999	Syn-WH7803	Sept-1999	Narragansett Bay RI North Atlantic	175.43	99.6	99.9	211	42.2	6	15	HQ317292.1
S-RIM2 R9_2006	Syn-WH7803	Sept-2006	Narragansett Bay RI North Atlantic	175.42	99.6	99.9	212	42.2	6	15	HQ317291.1
S-RIM2 R21_2007	Syn-WH7803	Sept-2007	Narragansett Bay RI North Atlantic	175.43	99.6	99.9	209	42.2	6	15	HQ317290.1
Syn9	Syn-WH8012	NA	Woods Hole MA North Atlantic	177.30	99.1	99.8	226	40.6	6	19	NC_008296
Syn10	Syn-	11/1/1986	Gulf Stream	177.10	99.1	99.8	206	40.6	6	19	HQ634191

Table 1 (*continued*)

Published name	Host	Isolation date	Isolation location	Genome size (kb)	GP % ID	CGP % ID	CDS	%G+C	tRNAs	AMGs	Accession# (INSDC)
	WH8107		North Atlantic								
Syn2	Syn-WH8012	6/1/1986	Sargasso Sea North Atlantic	175.60	99.7	99.8	204	41.3	6	17	HQ634190
Syn19	Syn-WH8109	7/1/1990	Sargasso Sea North Atlantic	175.23	99.7	99.8	216	41.3	6	17	NC_015286
P-RSM3	Pro-NATL2A	9/13/2000	Red Sea Indian Ocean	178.75	99.9	100	210	36.7	0	19	HQ634176
P-SSM4	Pro-MED4	6/6/2000	Sargasso Sea North Atlantic	178.25	99.9	100	221	36.7	0	19	NC_006884
S-ShM2	Syn-WH-8102	9/16/2001	Western Shelf North Atlantic	179.56	99.6	100	231	41.1	1	16	NC_015281
S-SSM2	Syn-WH8102	9/22/2001	Sargasso Sea North Atlantic	179.98	99.6	100	209	41.1	1	16	JF974292
P-SSM2	Pro-NATL1A	6/6/2000	Sargasso Sea North Atlantic	252.40	98.9	99.7	334	35.5	1	26	NC_006883
P-SSM5	Pro-NATL2A	8/31/1995	Sargasso Sea North Atlantic	252.01	98.9	99.7	321	35.5	1	26	HQ632825
S-SSM4	Syn-WH8010	9/22/2001	Sargasso Sea North Atlantic	182.80	NA	NA	224	39.4	3	18	NC_020875
S-SSM5	Syn-WH8102	9/22/2001	Sargasso Sea North Atlantic	176.18	NA	NA	226	40	4	21	NC_015289
S-SSM7	Syn-WH8109	9/22/2001	Sargasso Sea North Atlantic	232.88	NA	NA	319	37.9	5	21	NC_015287
S-PM2	Syn-WH7803	6/14/1995	English Channel	196.28	NA	NA	245	37.8	25	12	NC_006820
S-RSM4	Syn-WH8103	4/27/1999	Red Sea Indian Ocean	194.45	NA	NA	237	41.1	12	19	NC_013085
S-SM1	Syn-WH6501	9/17/2001	Off Western Shelf North Atlantic	174.08	NA	NA	234	41.1	6	21	NC_015282
S-SM2	Syn-WH8107	9/17/2001	Off Western Shelf North Atlantic	190.80	NA	NA	269	40.4	11	22	NC_015279
Syn33	Syn-WH7803	1/1/1995	Gulf Stream, North Atlantic	174.28	NA	NA	227	39.6	5	17	NC_015285
Syn30	Syn-WH7803	12/1/1987	NE Providence Channel North Atlantic	178.81	NA	NA	213	39.9	6	20	NC_021072
S-RIM8 A.HR1	Syn-WH7803	9/23/1999	Narragansett Bay RI North Atlantic	171.21	NA	NA	211	40.6	8	16	NC_020486
P-HM1	Pro-MED4	3/9/2006	Hawaii OTC North Pacific North Pacific	181.04	NA	NA	241	37.8	0	17	NC_015280
P-HM2	Pro-MED4	3/10/2006	Hawaii OTC North Pacific	183.81	NA	NA	242	38.1	0	16	NC_015284
MED4-213	Pro-MED4	3/9/2006	Hawaii OTC North Pacific	180.98	NA	NA	220	37.8	0	17	NC_020845
P-RSM1	Pro-MIT9303	9/8/2000	Red Sea Indian Ocean	177.21	NA	NA	213	40.2	2	18	NC_021071
P-RSM4	Pro-MIT9303	9/13/2000	Red Sea Indian Ocean	176.43	NA	NA	240	37.6	3	19	NC_015283
P-RSM6	Pro-NAT2LA	9/13/2000	Red Sea Indian Ocean	192.50	NA	NA	221	39.3	3	18	NC_020855
P-SSM3	Pro-NATL2A	5/6/1996	Sargasso Sea North Atlantic	179.06	NA	NA	219	36.7	0	18	NC_021559
P-SSM7	Pro-NATL1A	1/9/1999	Sargasso Sea North Atlantic	182.18	NA	NA	237	37.1	4	23	NC_015290

shared high nucleotide identity (99.4%) with the other OTU members in the surrounding genes and a 3′ *gnd* fragment remained in the S-CAM9 0908SB82 sequence, indicating that AMG losses can be gene specific (Fig. 2b). Further, this variation did not appear to be restricted to a geographic location; the isolate missing *gnd* was collected from the same location as S-CAM9 0808SB05 only one month later, and *gnd* was present in the third isolate collected nearby (22 km away) more than a year later. In host cyanobacteria, *gnd* functions to regenerate ribulose-5-phosphate as part of the pentose phosphate pathway with concomitant production of NADPH (Thompson et al., 2011). This would benefit phages during infection by fueling deoxynucleotide biosynthesis for phage replication (Thompson et al., 2011). However, the amino acid sequences of phage-encoded *gnd* clustered separately from those of any microbial group (Ignacio-Espinoza and Sullivan, 2012) and therefore suggest modified activity of this enzyme compared with the host.

The third example of variable AMG content within an OTU was S-CAM8, whose isolates varied for two different AMGs. Generally, the California isolates were more similar to one another (99.6% ANI-genome and 99.5% ANI-core genes) than either was to the Washington isolate (average of the two isolates was 93.4% ANI-genome and 97.5% ANI-core genes). The isolate from Washington State was missing ferredoxin (*petF*), an electron transfer protein, and a carboxylesterase family protein, both of which were present in two isolates from California. As with the previous two examples, we observed high nucleotide identity in the regions surrounding the AMG re-arrangements (Fig. 2c). However, unlike the case for *gnd* in S-CAM9, there was no evidence of remnants of the *petF* sequence, nor the accompanying hypothetical protein in the Washington isolate (S-CAM8 0810PA29). There were, however, remnants of the intergenic region upstream and downstream of *psbD* in the Washington isolate (Fig. 2c). The genetic differences surrounding *psbD* suggest two distinct recombination events rather than exchange of the whole module, although this cannot be ruled out. Further, a potential promoter early stem loop (PeSLs) structure was identified in the S-CAM8 sequences upstream of *psbD* (Fig. 2c). Promoter early stem loop (PeSL) sequence motifs present in

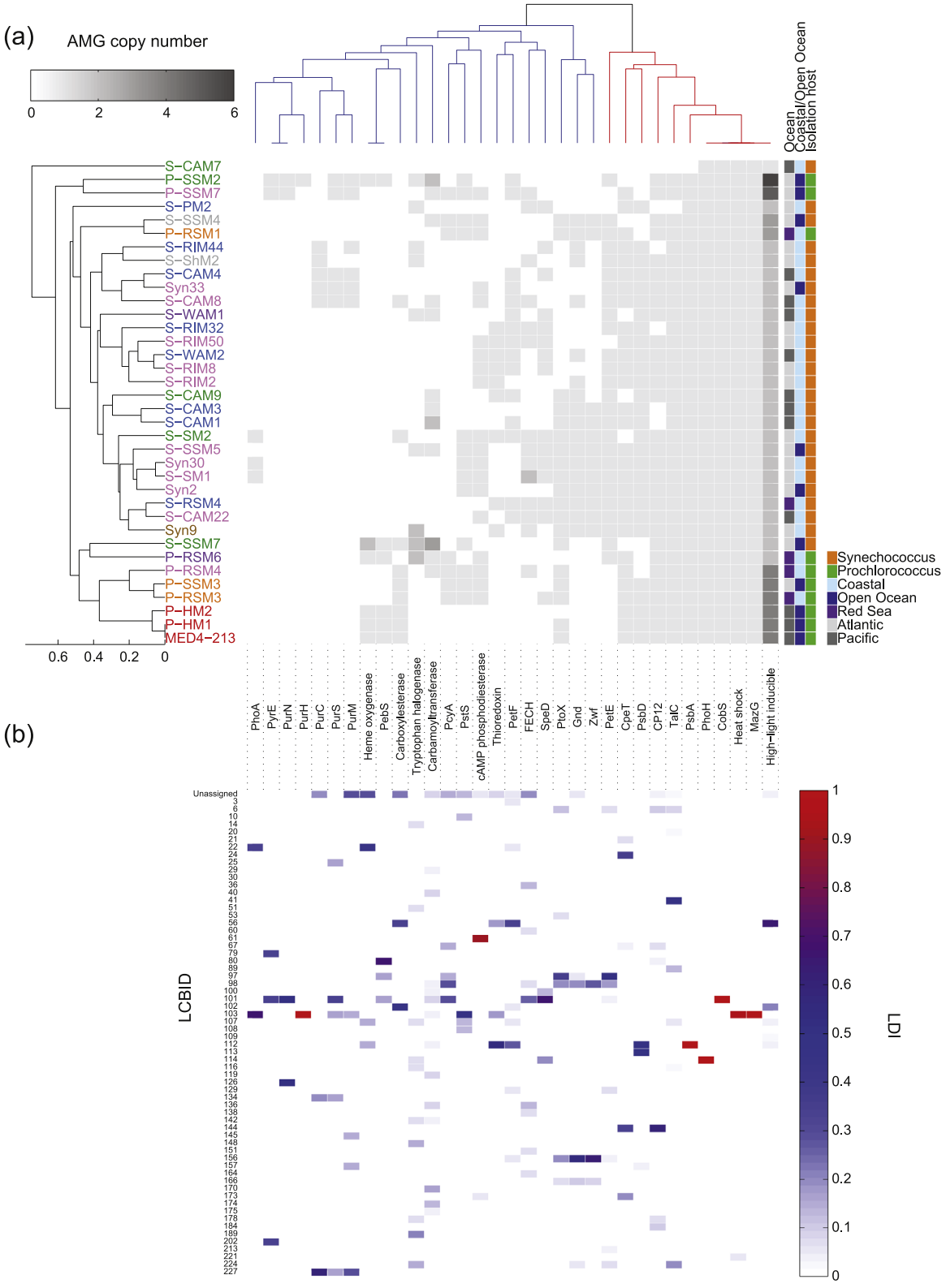

Fig. 1. (a) Heat map showing gene copy number matrix for 33 axillary metabolic genes (AMGs) across 36 OTU representatives. On the left, the OTU representatives are clustered based on similarity in AMG content. At the top, the AMGs are clustered based on their presence/absence among the phages (see methods). Virus names are colored by their phylogenetic clade associations (Fig. S1). Colored boxes on the right signify the ocean region (coastal vs. open ocean) and genus (*Synechococcus*/*Prochlorococcus*) of isolation. (b) Heat map showing assignment of AMGs to LCBs. LDI refers to LCB distribution index, which is calculated from the number of discrete LCBs an AMG is found in divided by the frequency of the AMG across 60 genomes; the more variable the genomic location, the closer the LDI is to one. The LCB ID given on the y axis is arbitrary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

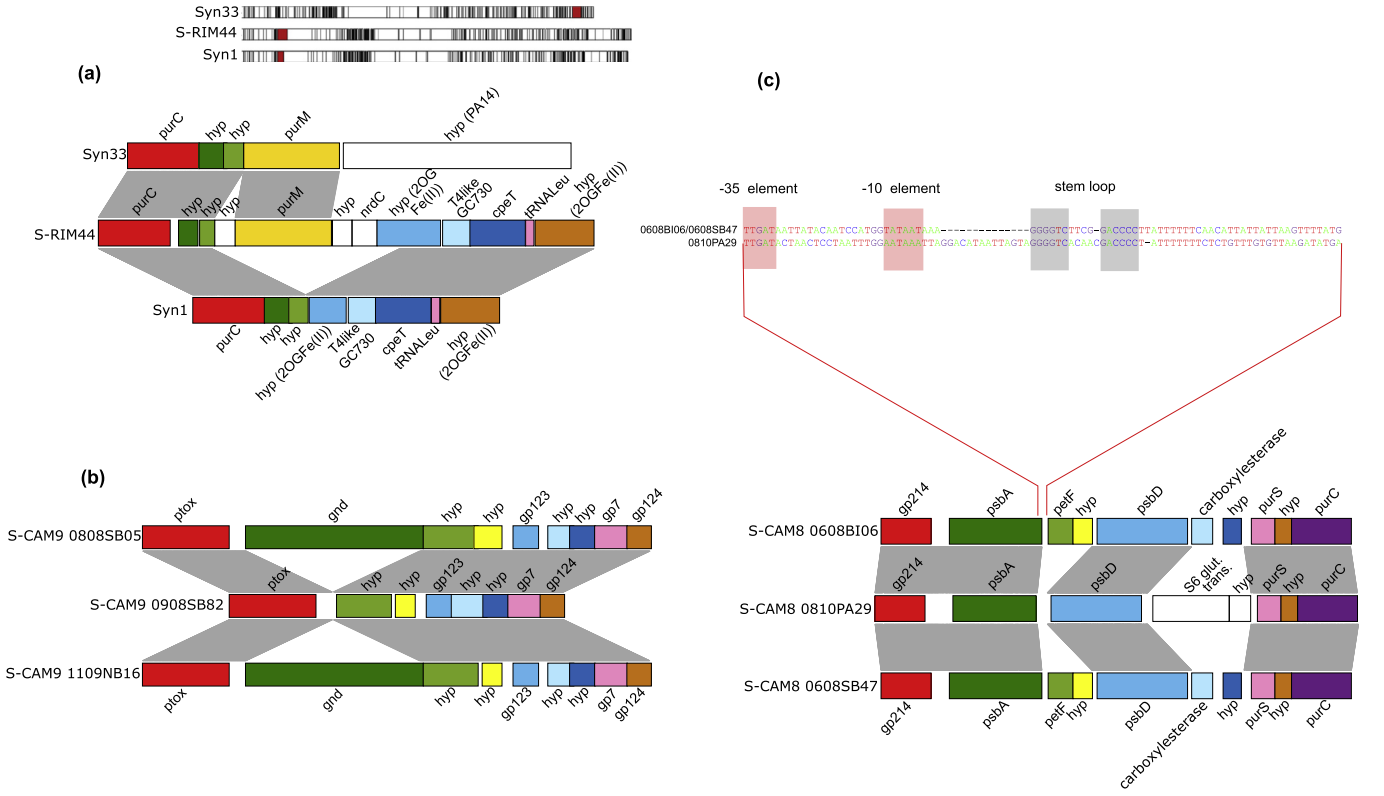

Fig. 2. Differences in AMG content between members of the same OTU. Homologous proteins as determined by protein clustering by CD-HIT are the same color. Grey lines indicate homologous nucleotides resulting from whole genome alignments by Mauve. Gene annotations are shown next to the predicted proteins, and white boxes are genes with no homologs. *Hyp* indicates hypothetical proteins, *nrdC* indicates ribonucleotide reductase C gene, tRNALeu indicates the tRNA Leucine gene and *S6 glut. trans*. indicates a ribosomal S6 glutamyl transferase gene. Protein domains are shown in parentheses where PA14 indicates a PA14 adhesin domain (PF07691), 2OGFe(II) indicates a 2 oxoglutarate and iron dependent oxygenase domain (PF03171). (a) Re-arrangement of the *purM* gene between S-RIM44 and Syn1. The top panel shows the genomic organization of *purM* in a phylogenetically distant Syn33. (b) Re-arrangement of the *gnd* gene in members of the S-CAM9 OTU. (c) Re-arrangement of the carboxylesterase gene in members of the S-CAM8 OTU. The upper panel indicates the presence of the PeSL-like motif upstream of *psbD;* S-CAM8_0608BI06 and S-CAM8_0608SB47 exhibit identical sequences whereas S-CAM8_0810PA29 is divergent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

T4-like cyanomyovirus genomes may allow for homologous recombination between closely related genomes while maintaining transcriptional autonomy of the invading DNA fragment (Arbiol et al., 2010). Thus, the identified PeSL may have contributed to the loss or gain of this fragment.

These three examples of intra-OTU variability in AMG content demonstrate that AMG gain or loss may be relatively frequent. AMG rearrangements within the same OTU presumably have occurred since a recent common ancestor. The variability among very closely-related genomes could be selectively neutral or, alternatively, suggest an adaptive role for AMGs at this fine genetic scale. We speculate that this variation is a result of adaptation, because we always detected the full length of an AMG in a genome. Although recombination events might initially be random, selection appears to have favored the insertion or deletion of AMGs at the gene's boundaries, similar to that observed for moron genes in dsDNA coliphages (Juhala et al., 2000). Finally, the weak correlation between phylogenetic topology and AMG content indicates that AMGs have little influence in structuring the main branches of the cyanomyovirus phylogeny.

2.2. Environmental associations with AMG content

The AMG content of a cyanomyovirus was related to the host genus (*Prochlorococcus* versus *Synechococcus*) on which it was originally isolated. Genomes isolated from the same host genus were more similar in AMG content than those isolated on different genera (ANOSIM test, global $R=0.424$, $p=0.001$). For example, cyanomyoviruses isolated on *Prochlorococcus* tended to cluster together by AMG content (Fig. 1a). The specific strain of *Prochlorococcus* appeared to matter as well. P-SSM2 and P-SSM7 were the only ones isolated on *Prochlorococcus sp*. NATL1A (Table 1), and these genomes were more similar to one another in AMG content than either of them were to any other isolate (Fig. 1a). Similar strain-specific clustering occurred for P-SSM3 and P-RSM3 (isolated on *Prochlorococcus* sp. NATL2A; Table 1) and MED4-213, P-HM1, and P-HM2 (isolated on *Prochlorococcus sp*. MED4; Table 1).

While AMG content may be correlated with isolation host, there was little evidence that any AMGs were genus-specific. Only one gene, *thioredoxin*, was restricted to viruses isolated on *Synechococcus* (Fig. 1a). Given that the gene only occurred in nine genomes, this pattern may be overturned with additional sampling. Indeed, our study detected several AMGs isolated from both *Synechococcus* and *Prochlorococcus* that had previously only been observed in one of the genera (Sullivan et al., 2010). For instance, *gnd* and *zwf* was found in P-RSM1, isolated on *Prochlorococcus*, and *pcyA* was found in S-SSM4, isolated on *Synechococcus*.

AMG content was also related to whether a virus was isolated from coastal versus open ocean. Viruses from coastal waters had more similar AMG content than those from open ocean waters and visa versa (ANOSIM test, global $R=0.239$, $p=0.002$; Fig. 1a). This pattern also held for AMG copy number. For instance, six out of the eight isolates with a high (4–6) copy number of high-light inducible genes (*hli*) were from open waters such as the Hawaiian Island region in the Pacific Ocean and Sargasso Sea in the Atlantic

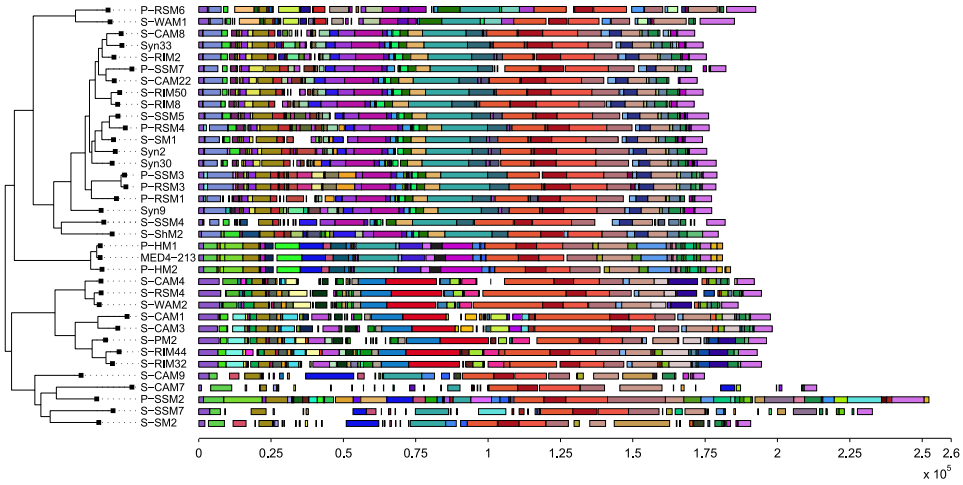

Fig. 3. Positions of the 227 shared LCBs from the progressiveMauve analysis. Shared LCBs are colored the same across genomes. Uncolored regions of the genomes indicate loci that are not shared among other viruses in the dataset. The tree represents the core phylogenetic tree (Fig. S1). The x-axis signifies the genomic coordinate of the LCB in each genome. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Ocean, where host cells experience intense solar radiation (Fig. 1a). Hli proteins are thought to protect the host's photosynthetic apparatus by dissipating excess light energy (He et al., 2001), and thus their copy number is expected to be under strong environmental selection (Lindell et al., 2004).

Finally, AMG content did not vary by ocean region of isolation (Pacific, Atlantic, or Red Sea) (ANOSIM test, global $R=0.048$, $p=0.207$; Fig. 1a). For example, these isolate pairs exhibited AMG content more similar to one another than to any other OTU representative (only one difference in each pair) and yet the members of each pair are from different ocean regions: P-SSM3/P-RSM3 (Atlantic/Red Sea), S-WAM2/S-RIM8 (Pacific/Atlantic). Furthermore, the closely-related S-RIM44 and S-RIM32 isolates (Fig. S1) were collected from the same location eight years apart (Table 1), and yet differed by seven AMGs (Fig. 1a, Table S1; *purC, purM, thioredoxin, gnd, petE, petF, and tryptophan halogenase*). Similarly, the closely-related P-RSM1 and P-RSM3 isolates (Fig. S1) were collected from the same location at the same time (Table 1) and differed by seven AMGs (Fig. 1a, Table S1; *gnd, zwf, psbD, speD, petE, carboxylesterase,* and one *hli*).

The broad patterns of AMG content associated with isolation host genus and coastal versus open ocean did not appear to be driven by a particular functional category of AMG (Table S1). The top three AMGs contributing to differences in AMG content by host genus were *pcyA*, *psbD* and *pstS* (contributing 8.2%, 6.4%, and 6.2% of the variation between the two groups; SIMPER analysis), with the former two genes associated with photosynthesis and the latter associated with phosphate stress. The top three AMGs contributing to differences in AMG content between coastal versus open ocean isolates were *pstS* (7.38%), cAMP phosphodiesterase (6.19%), and carbamoyltransferase (5.94%), genes involved in phosphate acquisition, regulation of signal transduction ("other metabolism"), and nucleotide synthesis, respectively. Thus, no particular AMG or functional type of AMG drove the significant differences in AMG content between isolation host genus and between coastal versus open ocean types.

2.3. Genomic context of AMGs

Previous studies have sought to classify the genomic context of AMGs with reference to nearby predicted ORFs (Millard et al., 2009; Sullivan et al., 2010). These studies have highlighted 'hypervariable/hyperplastic' regions of the genome, such as that flanked by the *g16-g17* genes, which frequently contains the AMGs *ptox*, *petE, gnd,* and *zwf* (Millard et al., 2009; Sullivan et al., 2010). Similarly, in cyanopodoviruses, AMGs appear to be constrained to particular 'island' regions (Labrie et al., 2013).

Identifying such patterns by eye quickly becomes overwhelming with additional genomes and genes of interest. Therefore, we developed a quantitative method to compare the genomic context of a set of target genes (here, the 33 AMGs). We first identified locally co-linear blocks (LCBs) between single representative isolates of each OTU using the progressiveMauve algorithm (Darling et al., 2010). LCBs are regions of nucleotide sequence conservation between two or more taxa (Darling et al., 2010) (Fig. S2). They are insensitive to deletions resulting from small-scale gene gain and loss and are naïve to predicted ORF boundaries. LCB size ranged from 15–28,767 bp with a median of 1856 bp and standard deviation of 3747 bp. LCB size followed a lognormal distribution ($\mu=7.29$, $\sigma=1.50$). We then quantified the number of discrete genomic contexts (the number of LCBs within which it is contained) of each AMG across all genomes. We hypothesize that the presence of an AMG in each LCB is related to at least one acquisition event.

As noted in previous studies, AMGs were distributed non-randomly across the genomes. We identified 227 LCBs among the genomes encompassing 6.03 Mbp (89.7%) of shared sequence (Fig. 3). Most (96.3%) of the AMGs (634 AMGs across all the genomes) were located in just 63 discrete LCBs, which encompassed 31.4% of the total LCB sequence space (Fig. 2b). There were only 25 instances of AMGs not assigned to any LCB (3.7%), meaning that the AMG fell in a stretch of sequence that was not conserved in any other genome.

The AMGs varied greatly in the degree to which their genomic context was conserved. Some AMGs were always found in one particular LCB (for instance, *psbA*, *phoH*, *cobS*, *hsp*, *mazG*). In contrast, others were found in diverse contexts, up to 14 LCBs in the case of *carbomoyltransferase* (Fig. 1b). Further, the AMGs that were found in a variety of LCBs appear to be non-syntenic when plotted across the genomes, whereas those restricted to one LCB are found in the same relative genomic location (Fig. S3).

Our analysis also revealed that the genomic context of AMGs is more variable than previously thought (Millard et al., 2009; Sullivan et al., 2010). As previously noted, the *g16-g17* region contains a cassette that includes AMGs of intermediate frequency (*petE*, *ptoX*, *gnd* and *zwf*) (Fig. S3). However, the addition of new genomes indicates that this cassette is associated with a homing endonuclease (green squares in Fig. S3). Thus, when the homing

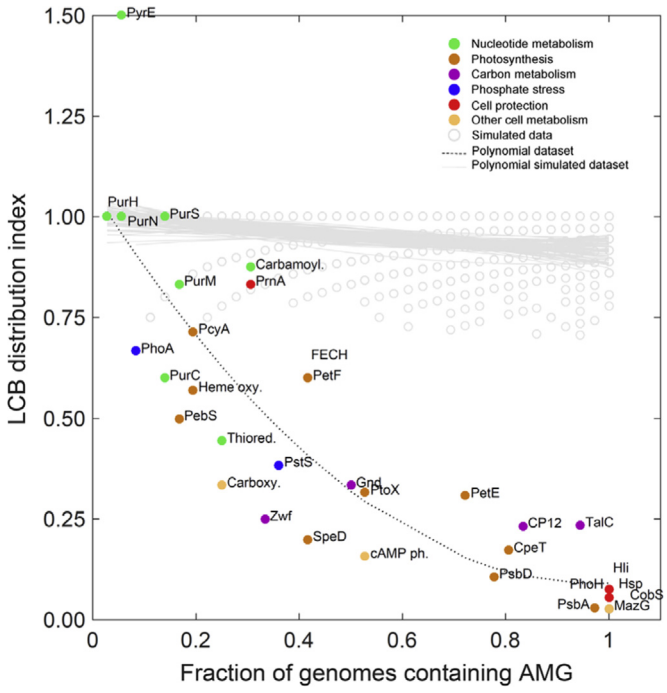

Fig. 4. Relationship between the fraction of genomes containing an AMG and the LCB-distribution index (LDI). The LDI is calculated as the number of discrete LCBs in which an AMG is found divided by the frequency of the AMG across the 60 genomes; the more variable the genomic location, the closer the LDI is to one. *PyrE* is greater than one because a breakpoint between two LCBs intersected the gene and thus it was assigned to two distinct LCBs. The grey lines indicate 2nd order polynomial fits to each of 100 simulated datasets (open circles). The dashed black line is the best polynomial fit to the actual data.

endonuclease is not in the *g16-g17* region (e.g., in the clade containing P-HM1, MED4-213, and P-HM2), the AMGs (if present) are still found near the homing endonuclease.

In addition, we identified several new AMG-variable regions, that is, LCBs that contain a high frequency of AMGs (Fig. 1b). While we could not identify any obvious reason why these regions should be variable, many of the AMG rearrangements appear to be the result of interruptions to homologous stretches of sequences in their nearest relatives that are specific to the gene boundaries of the invading AMG (i.e. insertion of *speD* in S-SM1 and S-CAM22, insertion of *FECH* in Syn30 and P-SSM7, and insertion of *pebS* in S-SSM7; Fig. S2). Such distinct gene insertions have been observed previously in other dsDNA bacteriophages (Hatfull et al., 2010; Hendrix, 2002; Hendrix et al., 1999) and suggest high levels of selection for a functional AMG (Hatfull et al., 2010).

The variability in genomic context of an AMG was negatively related to its frequency among cyanomyovirues (Fig. 4), and significantly more negative than expected from a simulated distribution ($p < 0.001$). Only a handful of AMGs occurring at relatively low frequencies among the genomes (e.g. carbomoyltransferase, *prnA*, *purM*, *purS*) fall within the randomized data (grey circles in Fig. 4), suggesting that the distribution of these AMGs could be random. However, the vast majority of AMGs were restricted to many fewer LCBs than expected by chance (Fig. 4). In particular, the previously defined cyanomyovirus 'core' genes *psbA, mazG, phoH, cobS and hsp* are only found in one LCB across all genomes. The negative pattern does not appear to be driven by AMGs from a particular functional category, and the trend holds for AMGs within the larger categories of nucleotide metabolism and photosynthesis (Fig. 4).

The observation that an AMG is found in a high diversity of genomic locations suggests that the gene has undergone multiple acquisitions and losses or intra-isolate rearrangements. We hypothesize that lower frequency AMGs found in diverse genomic contexts have experienced selection pressures that vary over season, by habitat, and so forth, leading to higher levels of gene gain and loss. Conversely, high-frequency AMGs found in a conserved genomic context may have been acquired early during evolution of this group, and consistent positive selection has maintained their presence in cyanomyovirus genomes. Previous studies have sought to understand the origins of AMGs using phylogenetic inference (Ignacio-Espinoza and Sullivan, 2012). The genomic context information adds to these analyses by providing a separate line of evidence of gene loss and gain, even in cases where exchange may have been phylogenetically congruent with the core genome phylogeny. Of course, the true rate of AMG exchange is likely much higher than either approach portrays because recombination events that decrease phage fitness are unlikely to be detected.

2.4. Absence of many AMGs in S-CAM7

Finally, we note the paucity of AMGs (only five) in the two S-CAM7 isolates. Unlike all other known cyanomyoviruses, their repertoire did not include *psbA* (Table 1). S-CAM7 belongs to a divergent clade (Fig. 3), but other members of this clade are replete in AMG content. The LCB comparisons reveal large genomic regions that are not shared amongst other clades of cyanomyoviruses, nor members of the same clade (Fig. 3) despite comparable branch lengths with other clades in the core gene phylogeny (Fig. S1). Moreover, members of this clade tend to have unusually large genomes (Fig. 3) and appear to be enriched in genes involved in carbohydrate biosynthesis (data not shown). These unique features may indicate a divergent life strategy by members of this clade. The possession of *psbA* is not unique to the T4-like myovirus group. Podoviruses (Dekel-Bird et al., 2013; Sullivan et al., 2006) and other divergent myoviruses (Sabehi et al., 2012) also possess the gene. Cyanophage encoded *psbA* is thought to overcome a metabolic bottleneck during infection at high light, where photodamage to the photosystem II (PSII) reaction center inhibits photochemical ATP production (Bragg and Chisholm, 2008; Hellweger, 2009). The continual synthesis of D1 polypeptides by the infecting cyanophage ensures energy production for phage morphogenesis (Lindell et al., 2005; Mann et al., 2003). The absence of *psbA* in S-CAM7 may reveal a distinct ecology for this virus suggesting, perhaps, that energy for morphogenesis is not strictly dependent on the maintenance of photochemistry. Future experiments might compare life history traits such as burst size and latency period in S-CAM7 isolates versus other isolates from the same clade (P-SSM2, S-SSM7, S-SM2, S-CAM9) that are not deficient in common AMGs.

Whole genome sequencing of single cells have also detected long contiguous sequences with high sequence similarity to the S-CAM7 clade associated with cells belonging to the marine *Roseobacter* clade (Labonte et al., 2015). Therefore, we speculate that another ecological difference of viruses within this clade may be an expanded host range, perhaps in some cases accompanied by a reduction in AMG content.

3. Conclusion

Host-like metabolic genes give clues to the forces shaping viral evolution. Although direct tests of the fitness consequences of cyanomyovirus AMGs are needed, the patterns observed in this study suggest that AMGs are subject to variable selection regimes, In particular, AMGs appear to be subject to variable levels of vertical and horizontal evolution. Similar to a model suggested for bacteria (Cordero and Polz, 2014), we suggest that high-frequency

(common across cyanomyovirus genomes) AMGs are consistently adaptive to all cyanomyoviruses and are therefore maintained through vertical inheritance. Their relatively stable genomic context supports this interpretation. In contrast, sporadic AMGs are highly variable in their genomic context and therefore appear to be subject to extensive horizontal evolution. Indeed, we found that even closely-related cyanomyovirus genomes (within the same OTU) can vary in their AMG content. These less common AMGs may likely confer traits that are subject to varying selection pressures. These pressures may include a coarse level of host preference and habitat type, as suggested by correlations between AMG genome content and these variables. Even with further sampling, however, such factors will be difficult to disentangle with comparative genomic studies because they often co-vary. Ultimately, future studies will require creative ways to uncover the drivers underlying the diversity in AMG content in marine cyanophages.

4. Methods

4.1. Cyanomyovirus genome collection

Thirty-five genomes were downloaded from NCBI, which encompassed all available cyanomyovirus genomes at the time of analysis, and 25 additional isolates were sequenced and annotated for the first time here (Table 1). The new cyanomyoviruses were isolated on *Synechococcus* sp. WH7803 from surface seawater samples collected from Southern California (n=19) between 2008 and 2010; Padilla Bay, Washington (n=3) in August 2010; and Narragansett Bay, Rhode Island (n=3) from 2004 to 2010 as described in Clasen et al. (2013) and Marston et al. (2013).

4.2. Genome sequencing

Plaque-purified isolates were removed from 4 °C storage and regrown on *Synechococcus* sp. WH7803 liquid culture (100 ml) in 250 ml Erlenmeyer flasks in a light incubator at 10 $\mu E\ m^{-2}\ s^{-1}$ on a 14:10 light: dark cycle. Phage titer was determined for each lysate via SYBR Green I staining and fluorescent microscopy. Phage isolates that yielded lysates with low phage-titer (less than 8×10^8 phages/ml) were grown multiple times (DNA products were combined) to provide sufficient quantities (1 μg) of genomic DNA. Phage genomic DNA was extracted as described in Henn et al. (2010). A genomic DNA library was prepared for Illumina sequencing as described in the "Low Sample (LS) Protocol" of the Illumina TruSeq DNA sample prep kit. Samples were sequenced on an Illumina HiSeq2000 sequencer (single read, paired-end with 100 cycles) at the UCI Genomics High-throughput Facility.

4.3. Genome assembly and annotation

Genome assembly of paired-end reads was performed with CLC Genomics Workbench 6.0.2 software using the default software parameters. Genome coverage (nucleotide redundancy) ranged from 2000x–8000x. ORF calling and primary annotation of protein coding sequences (CDS) and transfer RNAs were performed using RAST (Aziz et al., 2008). Secondary manual annotation of AMGs was performed with CLC Genomics Workbench as follows. An AMG database was created, which contained protein sequences from 33 AMGs that were extracted from various cyanomyovirus genomes. Thirty-two of the AMGs were chosen based on prior recognition (Sullivan et al., 2010) and one additional AMG was included (cAMP phosphodiesterase) based on its potentially significant role in host metabolism and its sporadic distribution amongst the 60 cyanomyovirus genomes. Each of the 60 genomes was blasted against this AMG database using tblastx (amino acid query to amino acid database). Protein sequences from CDS-hits that had a tblastx E-value less than 10^{-2} were reciprocal-blasted against the NCBI non-redundant protein database. Gene identity was assigned to a given hypothetical CDS if the blastp E-value was less than or equal to 10^{-5}, sequence identity was at least 35%, and the query cover was at least 60% (similar criteria reported in: (Kelly et al., 2013; Millard et al., 2009; Sullivan et al., 2010)). This same technique was used to annotate structural genes and conserved T4-like myovirus core genes in the 25 genomes presented here for the first time as well as previously annotated genomes that required further annotation. To ensure AMGs formed single homologous clusters we used CD-HIT (Li and Godzik, 2006), with a 40% identity threshold and a word size of 2. This revealed some erroneous previous annotations. In particular, genes previously annotated as *purM* in some of the genomes did not form one homologous cluster with other *purM*s and we could find no evidence to suggest these were bona fide copies of this gene. Therefore, they were omitted from the analysis. To compare all 60 genomes, we designated *g32*, encoding single-stranded DNA binding protein, as the first gene in each linearized genome.

4.4. Phylogenetic analysis

Seven conserved structural protein genes common to cyanomyoviruses were used to construct a core gene phylogeny for 36 viral isolates representing 36 OTUs from nine locations (Fig. S1). Members of an OTU have $\geq$99% nucleotide similarity in the *g20* portal protein gene (Clasen et al., 2013; Marston and Amrich, 2009; Marston et al., 2013). These seven conserved structural proteins were selected from previous studies (Ignacio-Espinoza and Sullivan, 2012) and were chosen based on the fact that their single gene topology was reported to be congruent with a reference T4 core supertree topology (Ignacio-Espinoza and Sullivan, 2012). The genes include those encoding gp6 baseplate wedge, gp22 prohead core scaffold, gp26 baseplate hub subunit, gp32 single-stranded DNA binding, gp55 sigma factor transcription, RegA translational regulator, and NrdA ribonuclease reductase alpha subunit. The amino acid sequences of the genes were aligned independently using MUSCLE (Edgar, 2004), using a –gap-open score of −2.9, a –gapextend of 0, and the BLOSUM62 substitution matrix. The independent alignments were concatenated using Geneious v8.0.4 (Kearse et al., 2012). Low information regions of the alignment were trimmed using GBLOCKS (Talavera and Castresana, 2007), where the minimum sequences was set to n/2, maximum number of non-conserved contiguous sites was set to 50 and the minimum block length was set to 5. Phylogenetic inference was based on the resulting alignment and conducted in RAxML (Stamatakis, 2014) using a WAG $+I+\Gamma_4$ model. Bootstrap support was given by RAxML rapid bootstrap option (-f) with 100 replicates.

4.5. Whole genome alignment

The genomes of the 36 OTU representatives were aligned using progressive Mauve v2.3.1 (Darling et al., 2010). The match seed weight was set to 15 and the minimum weight was set as default (3). The default HOXD scoring matrix was used and gap open and extend scores were set to −400 and −30 respectively. Progressive Mauve allows for prediction of genome regions, termed local colinear blocks (LCBs), that are shared between representatives of the alignment and are tolerant to genome re-arrangement as well as within block minor gene gain and loss. LCB coordinates (Table S2) were retrieved from the .xmfa file using a custom script and the distribution of AMGs within these LCBs was determined (Table S3). We further used progressiveMauve to compare whole genome

alignments of members of the same OTU (Fig. 2), using the same settings as above, to identify genetic rearrangements and remnants of deleted gene regions.

4.6. Statistical analyses

The following statistical analysis were conducted in PRIMER v6 (Clarke and Warwick, 2001). Pairwise distances between cyanomyovirus genomes were calculated using a Euclidean distance metric of a presence/absence matrix of the 33 AMGs. An agglomerative hierarchical cluster tree (dendrogram) was constructed from the pairwise distances using the unweighted average distance method. Likewise, AMGs were clustered using the same methods based on their presence/absence across the genomes of representatives of the 36 OTUs. An analysis of similarity (ANOSIM) was performed to determine the significance of the relationship between AMG content (ranked Euclidean distance of the presence/absence matrix) across the genomes and the following factors: host genus, coastal vs. open ocean, and ocean region (Pacific, Atlantic, Red Sea). The global R value was compared to 999 permutations of the matrix to infer significance. A similarity percentage analysis (SIMPER) was then performed to determine the relative contribution of each AMG to the observed differences among factors. A RELATE test was used to assess the correlation between a phylogenetic similarity matrix (generated from the above core gene phylogenetic analyses) and an AMG content similarity matrix for the 36 genomes using the spearman rank correlation method. The correlation coefficient (Rho) was compared to 999 permutations of the AMG content matrix to test for significance.

The LCB distribution index (LDI) was calculated as the number of distinct LCBs an AMG was found in divided by the frequency of that AMG across the 36 genomes used in the analysis (including one representative from each OTU). We used a second order polynomial fit to quantify the relationship between the fraction of genomes containing an AMG and the LDI. To test for the significance of this relationship, we simulated the expected relationship assuming that AMGs are gained and lost randomly, without consideration of the genetic relatedness of the genomes. We followed the following procedure: 1. For each of 33 AMGs, we assigned a random distribution across genomes (i.e., each AMG could be found in n genomes, where n is between 1 and 36 genomes, based on a uniform distribution). 2. We then assigned each AMG a length (in bp) for each of the n genomes based on its actual length distribution across the genomes (assuming a normal distribution and estimating its mean and standard deviation from the data). 3. For each AMG in each of its n genomes, the AMG was inserted into each genome at a random location (where each genome corresponds to one of the 36 actual genomes that vary in length and LCB locations). 4. We calculated the number of LCBs that each AMG intersected and then the LDI for each of the 33 AMGs. 6. We repeated steps 1–5 100 times, producing 100 polynomial relationships that could be compared to the observed relationship. This approach recaptures any length biases in AMGs and LCBs that are in the dataset. The simulations were conducted in MATLAB vR2015b and the code is available upon request.

Acknowledgments

We thank Nada Awad, Stephanie Chen, Susan Cho, Jessica Clasen, Aayah Fatayerji, China Hanson, Andrew Ho, Yazeed Ibrahim, Steven Nguyen, Sanae Ogura, Jeniffer Peña, and Kelli Shiroma for assistance with virus isolation and characterization. Funding for this research was provided by the National Science Foundation (OCE-1029684 and OCE-1332782 to MM; OCE-1031783 and OCE-1332740 to JM) and the RWU Foundation to Promote Scholarship to MM.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2016.09.016.

References

Anantharaman, K., Duhaime, M., Breier, J., Wendt, K., Toner, B., Dick, G., 2014. Sulfur oxidation genes in diverse deep-sea viruses. Science 344, 757–760.
Arbiol, C., Comeau, A., Kutateladze, M., Adamia, R., Krisch, H., 2010. Mobile regulatory cassettes mediate modular shuffling in T4-type phage genomes. Genome Biol. Evol. 2, 140–152.
Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G., Olson, R., Osterman, A., Overbeek, R., McNeil, L., Paarmann, D., Paczian, T., Parrello, B., Pusch, G., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., 2008. The RAST server: rapid annotations using subsystems technology. BMC Genom. 9, 75.
Bragg, J., Chisholm, S., 2008. Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. PLoS One 3, e3550.
Bratbak, G., Thingstad, F., Heldal, M., 1994. Viruses and the microbial loop. Microb. Ecol. 28, 209–221.
Breitbart, M., Thompson, L., Suttle, C., Sullivan, M., 2007. Exploring the vast diversity of marine viruses. Oceanography 20, 135–139.
Clarke, K., Warwick, R., 2001. Change in marine communities: an approach to statistical analysis and interpretation. Primer v6 User Manual. PRIMER-E Ltd., Plymouth.
Clasen, J., Hanson, C., Ibrahim, Y., Weihe, C., Marston, M., Martiny, J., 2013. Diversity and temporal dynamics of Southern California coastal marine cyanophage isolates. Aquat. Microb. Ecol. 69, 17–31.
Clokie, M., Mann, N., 2006. Marine cyanophages and light. Environ. Microbiol. 8, 2074–2082.
Comeau, A., Bertrand, C., Letarov, A., Tetart, F., Krisch, H., 2007. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. Virology 362, 384–396.
Cordero, O., Polz, M., 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nat. Rev. Microbiol. 12, 263–273.
Darling, A., Mau, B., Perna, N., 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147.
Dekel-Bird, N., Avrani, S., Sabehi, G., Pekarsky, I., Marston, M., Kirzner, S., Lindell, D., 2013. Diversity and evolutionary relationships of T7-like podoviruses infecting marine cyanobacteria. Environ. Microbiol. 15, 1476–1491.
Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P., Sullivan, M.B., 2014. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. Nature 513, 242–245.
Dwivedi, B., Xue, B., Lundin, D., Edwards, R., Breitbart, M., 2013. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. BMC Evol. Biol., 13.
Edgar, R., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Fuhrman, J., 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399, 541–548.
Hagay, E., Mandel-Gutfreund, Y., Béjà, O., 2014. Comparative metagenomics analyses reveal viral-induced shifts of host metabolism towards nucleotide biosysnthesis. Microbiome 2, 9.
Hatfull, G., Jacobs-Sera, D., Lawrence, J., Pope, W., Russell, D., Ko, C., Weber, R., Patel, M., Germane, K., Edgar, R., Hoyte, N., Bowman, C., Tantoco, A., Paladin, E., Myers, M., Smith, A., Grace, M., Pham, T., O'Brien, M., Vogelsberger, A., Hryckowian, A., Wynalek, J., Donis-Keller, H., Bogel, M., Peebles, C., Cresawn, S., Hendrix, R., 2010. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J. Mol. Biol. 397, 119–143.
He, Q., Dolganov, N., Bjorkman, O., Grossman, A., 2001. The high light-inducible polypeptides in *Synechocystis* PCC6803 – expression and function in high light. J. Biol. Chem. 276, 306–314.
Hellweger, F., 2009. Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. Environ. Microbiol. 11, 1386–1394.
Hendrix, R., 2002. Bacteriophages: evolution of the majority. Theor. Popul. Biol. 61, 471–480.
Hendrix, R., Smith, M., Burns, R., Ford, M., Hatfull, G., 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. USA 96, 2192–2197.
Henn, M.R., Sullivan, M.B., Stange-Thomann, N., Osburne, M.S., Berlin, A.M., Kelly, L., Yandava, C., Kodira, C., Zeng, Q., Weiand, M., Sparrow, T., Saif, S., Giannoukos, G., Young, S.K., Nusbaum, C., Birren, B.W., Chisholm, S.W., 2010. Analysis of high-throughput sequencing and annotation strategies for phage genomes. PLoS One 5, e9083.
Ignacio-Espinoza, J.C., Sullivan, M.B., 2012. Phylogenomics of T4 cyanophages:

lateral gene transfer in the 'core' and origins of host genes. Environ. Microbiol. 14, 2113–2126.
Jover, L., Effler, T., Buchan, A., Wilhelm, S., Weitz, J., 2014. The elemental composition of virus particles: implications for marine biogeochemical cycles. Nat. Rev. Microbiol. 12, 519–528.
Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., Hendrix, R.W., 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages[1]. J. Mol. Biol. 299, 27–51.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.
Kelly, L., Ding, H., Huang, K., Osburne, M., Chisholm, S., 2013. Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. ISME J. 7, 1827–1841.
Labonte, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., Sullivan, M.B., Woyke, T., Eric Wommack, K., Stepanauskas, R., 2015. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. ISME J. 9, 2386–2399.
Labrie, S.J., Frois-Moniz, K., Osburne, M.S., Kelly, L., Roggensack, S.E., Sullivan, M.B., Gearin, G., Zeng, Q., Fitzgerald, M., Henn, M.R., Chisholm, S.W., 2013. Genomes of marine cyanopodoviruses reveal multiple origins of diversity. Environ. Microbiol. 15, 1356–1376.
Li, W., Godzik, A., 2006. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
Lindell, D., Jaffe, J., Johnson, Z., Church, G., Chisholm, S., 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438, 86–89.
Lindell, D., Sullivan, M., Johnson, Z., Tolonen, A., Rohwer, F., Chisholm, S., 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc. Natl. Acad. Sci. USA 101, 11013–11018.
Mann, N., Cook, A., Millard, A., Bailey, S., Clokie, M., 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 424, 741–741.
Marston, M.F., Sallee, J.L., 2003. Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. Appl. Environ. Microbiol. 69, 4639–4647.
Marston, M., Amrich, C., 2009. Recombination and microdiversity in coastal marine cyanophages. Environ. Microbiol. 11, 2893–2903.
Marston, M.F., Taylor, S., Sme, N., Parsons, R.J., Noyes, T.J., Martiny, J.B., 2013. Marine cyanophages exhibit local and regional biogeography. Environ. Microbiol. 15, 1452–1463.
Millard, A., Zwirglmaier, K., Downey, M., Mann, N., Scanlan, D., 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environ. Microbiol. 11, 2370–2387.
Petrov, V., Ratnayaka, S., Nolan, J., Miller, E., Karam, J., 2010. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. Virol. J. 7, 292–311.
Philosof, A., Battchikova, N., Aro, E., Beja, O., 2011. Marine cyanophages: tinkering with the electron transport chain. ISME J. 5, 1568–1570.
Polz, M., Alm, E., Hanage, W., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 29, 170–175.
Sabehi, G., Shaulov, L., Silver, D., Yanai, I., Harel, A., Lindell, D., 2012. A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. Proc. Natl. Acad. Sci. USA 109, 2037–2042.
Sharon, I., Battchikova, N., Aro, E., Giglione, C., Meinnel, T., Glaser, F., Pinter, R., Breitbart, M., Rohwer, F., Beja, O., 2011. Comparative metagenomics of microbial traits within oceanic viral communities. ISME J. 5, 1178–1190.
Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R., Partensky, F., Koonin, E., Wolf, Y., Nelson, N., Beja, O., 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature 461, 258–262.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Sullivan, M., Lindell, D., Lee, J., Thompson, L., Bielawski, J., Chisholm, S., 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. 4, 1344–1357.
Sullivan, M., Huang, K., Ignacio-Espinoza, J., Berlin, A., Kelly, L., Weigele, P., DeFrancesco, A., Kern, S., Thompson, L., Young, S., Yandava, C., Fu, R., Krastins, B., Chase, M., Sarracino, D., Osburne, M., Henn, M., Chisholm, S., 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ. Microbiol. 12, 3035–3056.
Suttle, C., 2005a. The viriosphere: the greatest biological diversity on Earth and driver of global processes. Environ. Microbiol. 7, 481–482.
Suttle, C., 2005b. Viruses in the sea. Nature 437, 356–361.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.
Tettelin, H., Riley, D., Cattuto, C., Medini, D., 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. 11, 472–477.
Thompson, L., Zeng, Q., Kelly, L., Huang, K., Singer, A., Stubbe, J., Chisholm, S., 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc. Natl. Acad. Sci. USA 108, E757–E764.
Williamson, S., Rusch, D., Yooseph, S., Halpern, A., Heidelberg, K., Glass, J., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C., Sutton, G., Frazier, M., Venter, J., 2008. The sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3, e1456.