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

Permalink
https://escholarship.org/uc/item/08x5k097

Authors
Crummett, LT
Puxty, RJ
Weihe, C
et al.

Publication Date
2016-12-01

DOI
10.1016/j.virol.2016.09.016

License
CC BY 4.0

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

Lisa T. Crummett a,n,1, 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

A R T I C L E   I N F O

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

A B S T R A C T

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.

© 2016 Elsevier Inc. All rights reserved.

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-Espinosa 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; Hellweg, 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
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, petE (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 Amrith, 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 ≥ 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% (x̄=98.8%) across the entire genome and 98.4–100% (x̄=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 (x̄=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 purE, 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 S-RIM32, also possessed a copy of hyp-hyp (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 Synchococcus 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 (Table S1). This difference appeared to be the result of a deletion of gnd in S-CAM9 090855B82. 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

<table>
<thead>
<tr>
<th>Published name</th>
<th>Host</th>
<th>Isolation date</th>
<th>Isolation location</th>
<th>Genome size (kb)</th>
<th>GP % ID</th>
<th>CGP % ID</th>
<th>CDS</th>
<th>rRNAs</th>
<th>AMGs</th>
<th>Accession#</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-CAM1 0208SB26</td>
<td>Syn-</td>
<td>2/15/2008</td>
<td>Southern CA Pacific Ocean</td>
<td>198.01</td>
<td>99.6</td>
<td>99.9</td>
<td>253</td>
<td>43</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM1 0309SB33</td>
<td>Syn-</td>
<td>3/11/2009</td>
<td>Southern CA Pacific Ocean</td>
<td>197.54</td>
<td>99.6</td>
<td>99.9</td>
<td>253</td>
<td>43</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM1 0310NB17</td>
<td>Syn-</td>
<td>3/26/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>197.53</td>
<td>99.6</td>
<td>99.9</td>
<td>253</td>
<td>43</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM1 0809CC03</td>
<td>Syn-</td>
<td>8/27/2009</td>
<td>Southern CA Pacific Ocean</td>
<td>197.26</td>
<td>99.6</td>
<td>99.9</td>
<td>253</td>
<td>43</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM1 0810SB17</td>
<td>Syn-</td>
<td>8/18/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>197.53</td>
<td>99.6</td>
<td>99.9</td>
<td>253</td>
<td>43</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM1 0910CC29</td>
<td>Syn-</td>
<td>9/15/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>197.53</td>
<td>99.6</td>
<td>99.9</td>
<td>253</td>
<td>43</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM3 0808SB25</td>
<td>Syn-</td>
<td>8/15/2008</td>
<td>Southern CA Pacific Ocean</td>
<td>197.84</td>
<td>99.4</td>
<td>99.9</td>
<td>239</td>
<td>41.6</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>S-CAM3 0910TB04</td>
<td>Syn-</td>
<td>9/20/2010</td>
<td>Tijuana, Mexico Pacific Ocean</td>
<td>197.80</td>
<td>99.4</td>
<td>99.9</td>
<td>239</td>
<td>41.6</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>S-CAM3 1010CC42</td>
<td>Syn-</td>
<td>10/13/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>198.19</td>
<td>99.4</td>
<td>99.9</td>
<td>240</td>
<td>41.6</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>S-CAM4 0309CC44</td>
<td>Syn-</td>
<td>3/11/2009</td>
<td>Southern CA Pacific Ocean</td>
<td>191.94</td>
<td>99.3</td>
<td>99.5</td>
<td>238</td>
<td>38.6</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM4 0809SB33</td>
<td>Syn-</td>
<td>8/12/2009</td>
<td>Southern CA Pacific Ocean</td>
<td>191.98</td>
<td>99.3</td>
<td>99.5</td>
<td>238</td>
<td>38.6</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM4 1010NB23</td>
<td>Syn-</td>
<td>10/13/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>191.62</td>
<td>99.3</td>
<td>99.5</td>
<td>237</td>
<td>38.6</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>S-CAM7 0910CC49</td>
<td>Syn-</td>
<td>9/15/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>216.00</td>
<td>92.6</td>
<td>99.6</td>
<td>268</td>
<td>41.2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>S-CAM7 0910SB42</td>
<td>Syn-</td>
<td>9/15/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>214.03</td>
<td>92.6</td>
<td>99.6</td>
<td>265</td>
<td>41.2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>S-CAM8 0608BB06</td>
<td>Syn-</td>
<td>6/15/2008</td>
<td>Southern CA Pacific Ocean</td>
<td>171.41</td>
<td>*95.6</td>
<td>*98.4</td>
<td>207</td>
<td>39.3</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>S-CAM8 0608SB47</td>
<td>Syn-</td>
<td>6/15/2008</td>
<td>Southern CA Pacific Ocean</td>
<td>171.41</td>
<td>*95.6</td>
<td>*98.4</td>
<td>206</td>
<td>39.3</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>S-CAM8 0810PA29</td>
<td>Syn-</td>
<td>8/20/2010</td>
<td>Padilla Bay, WA Pacific Ocean</td>
<td>172.34</td>
<td>*95.6</td>
<td>*98.4</td>
<td>211</td>
<td>39.2</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>S-CAM9 0808SB05</td>
<td>Syn-</td>
<td>8/15/2008</td>
<td>Southern CA Pacific Ocean</td>
<td>174.81</td>
<td>99.2</td>
<td>99.9</td>
<td>228</td>
<td>39</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>S-CAM9 0908SB82</td>
<td>Syn-</td>
<td>9/1/2008</td>
<td>Southern CA Pacific Ocean</td>
<td>174.66</td>
<td>99.2</td>
<td>99.9</td>
<td>228</td>
<td>39</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>S-CAM22 0210CC35</td>
<td>Syn-</td>
<td>2/17/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>172.23</td>
<td>99.5</td>
<td>99.8</td>
<td>215</td>
<td>39.9</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>S-CAM22 0310NB44</td>
<td>Syn-</td>
<td>3/17/2010</td>
<td>Southern CA Pacific Ocean</td>
<td>172.40</td>
<td>99.5</td>
<td>99.8</td>
<td>214</td>
<td>39.9</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>S-CAM22 1209TA19</td>
<td>Syn-</td>
<td>12/09/2009</td>
<td>Tijuana, Mexico Pacific Ocean</td>
<td>172.35</td>
<td>99.5</td>
<td>99.8</td>
<td>214</td>
<td>39.9</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>S-WAM1 0810PA09</td>
<td>Syn-</td>
<td>8/20/2010</td>
<td>Padilla Bay WA Pacific Ocean</td>
<td>185.10</td>
<td>NA</td>
<td>NA</td>
<td>222</td>
<td>44.7</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>S-WAM2 0810PA29</td>
<td>Syn-</td>
<td>8/10/2010</td>
<td>Padilla Bay WA Pacific Ocean</td>
<td>186.39</td>
<td>NA</td>
<td>NA</td>
<td>227</td>
<td>41.3</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>S-RIM50 RW-29-0704</td>
<td>Syn-</td>
<td>July-2004</td>
<td>Narragansett Bay RI North Atlantic</td>
<td>174.31</td>
<td>NA</td>
<td>NA</td>
<td>227</td>
<td>40.3</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>S-RIM44 W2-07-0710</td>
<td>Syn-</td>
<td>July-2010</td>
<td>Block Island Sound RI North Atlantic</td>
<td>193.00</td>
<td>98.3</td>
<td>99.7</td>
<td>230</td>
<td>40.6</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Syn1</td>
<td>Syn-</td>
<td>10/1/1990</td>
<td>Woods Hole MA North Atlantic</td>
<td>191.20</td>
<td>98.3</td>
<td>99.7</td>
<td>234</td>
<td>40.6</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>S-RIM2 R1-1999</td>
<td>Syn-</td>
<td>Sept-1999</td>
<td>Narragansett Bay RI North Atlantic</td>
<td>173.43</td>
<td>99.6</td>
<td>99.9</td>
<td>212</td>
<td>42.2</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>S-RIM2 R9_2006</td>
<td>Syn-</td>
<td>Sept-2006</td>
<td>Narragansett Bay RI North Atlantic</td>
<td>175.42</td>
<td>99.6</td>
<td>99.9</td>
<td>212</td>
<td>42.2</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>S-RIM2 R21_2007</td>
<td>Syn-</td>
<td>Sept-2007</td>
<td>Narragansett Bay RI North Atlantic</td>
<td>175.43</td>
<td>99.6</td>
<td>99.9</td>
<td>209</td>
<td>42.2</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>Syn9</td>
<td>Syn-</td>
<td>NA</td>
<td>Woods Hole MA North Atlantic</td>
<td>177.30</td>
<td>99.1</td>
<td>99.8</td>
<td>226</td>
<td>40.6</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>Syn10</td>
<td>Syn-</td>
<td>11/1/1986</td>
<td>Gulf Stream</td>
<td>177.30</td>
<td>99.1</td>
<td>99.8</td>
<td>206</td>
<td>40.6</td>
<td>6</td>
<td>19</td>
</tr>
</tbody>
</table>
Table 1 (continued)

<table>
<thead>
<tr>
<th>Published name</th>
<th>Host</th>
<th>Isolation date</th>
<th>Isolation location</th>
<th>Genome size (kb)</th>
<th>GP % ID</th>
<th>CGP % ID</th>
<th>CDS</th>
<th>SG + C</th>
<th>tRNAs</th>
<th>AMGs</th>
<th>Accession# (INSDC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syn2</td>
<td>WH8107</td>
<td>6/1/1986</td>
<td>North Atlantic</td>
<td>175.60</td>
<td>99.7</td>
<td>99.8</td>
<td>204</td>
<td>41.3</td>
<td>6</td>
<td>17</td>
<td>HQ634190</td>
</tr>
<tr>
<td>Syn19</td>
<td>WH8012</td>
<td>7/1/1990</td>
<td>North Atlantic</td>
<td>175.23</td>
<td>99.7</td>
<td>99.8</td>
<td>216</td>
<td>41.3</td>
<td>6</td>
<td>17</td>
<td>NC_015286</td>
</tr>
<tr>
<td>P-RSM3</td>
<td>Pro-NATL2A</td>
<td>9/13/2000</td>
<td>Red Sea Indian</td>
<td>178.75</td>
<td>99.9</td>
<td>100</td>
<td>210</td>
<td>36.7</td>
<td>0</td>
<td>19</td>
<td>HQ634176</td>
</tr>
<tr>
<td>P-SSM4</td>
<td>Pro-MED4</td>
<td>6/6/2000</td>
<td>North Atlantic</td>
<td>178.25</td>
<td>99.9</td>
<td>100</td>
<td>221</td>
<td>36.7</td>
<td>0</td>
<td>17</td>
<td>NC_006884</td>
</tr>
<tr>
<td>S-SLM2</td>
<td>Syn-WH-8102</td>
<td>9/16/2001</td>
<td>Western Shelf North</td>
<td>179.56</td>
<td>99.6</td>
<td>100</td>
<td>231</td>
<td>41.1</td>
<td>1</td>
<td>16</td>
<td>NC_015281</td>
</tr>
<tr>
<td>S-SSM2</td>
<td>Syn-WH-8102</td>
<td>9/22/2001</td>
<td>North Atlantic</td>
<td>179.98</td>
<td>99.6</td>
<td>100</td>
<td>209</td>
<td>41.1</td>
<td>1</td>
<td>16</td>
<td>JF974292</td>
</tr>
<tr>
<td>P-SSM2</td>
<td>Pro-NATL1A</td>
<td>6/6/2000</td>
<td>North Atlantic</td>
<td>252.40</td>
<td>98.9</td>
<td>99.7</td>
<td>334</td>
<td>35.5</td>
<td>1</td>
<td>26</td>
<td>NC_006883</td>
</tr>
<tr>
<td>P-SSM5</td>
<td>Pro-NATL2A</td>
<td>8/31/1995</td>
<td>North Atlantic</td>
<td>252.01</td>
<td>98.9</td>
<td>99.7</td>
<td>321</td>
<td>35.5</td>
<td>1</td>
<td>26</td>
<td>HQ632825</td>
</tr>
<tr>
<td>S-SSM4</td>
<td>Syn-WH8010</td>
<td>9/22/2001</td>
<td>North Atlantic</td>
<td>182.80</td>
<td>NA</td>
<td>NA</td>
<td>224</td>
<td>39.4</td>
<td>3</td>
<td>18</td>
<td>NC_020875</td>
</tr>
<tr>
<td>S-SSM5</td>
<td>Syn-WH8012</td>
<td>9/22/2001</td>
<td>North Atlantic</td>
<td>176.18</td>
<td>NA</td>
<td>NA</td>
<td>226</td>
<td>40.4</td>
<td>4</td>
<td>21</td>
<td>NC_015289</td>
</tr>
<tr>
<td>S-SSM7</td>
<td>Syn-WH8109</td>
<td>9/22/2001</td>
<td>North Atlantic</td>
<td>232.88</td>
<td>NA</td>
<td>NA</td>
<td>319</td>
<td>37.9</td>
<td>5</td>
<td>21</td>
<td>NC_015287</td>
</tr>
<tr>
<td>S-PM2</td>
<td>Syn-WH7803</td>
<td>6/14/1995</td>
<td>English Channel</td>
<td>196.28</td>
<td>NA</td>
<td>NA</td>
<td>245</td>
<td>37.8</td>
<td>25</td>
<td>12</td>
<td>NC_006820</td>
</tr>
<tr>
<td>S-RSM4</td>
<td>Syn-WH8013</td>
<td>4/27/1999</td>
<td>Red Sea Indian</td>
<td>194.45</td>
<td>NA</td>
<td>NA</td>
<td>237</td>
<td>41.1</td>
<td>12</td>
<td>19</td>
<td>NC_013085</td>
</tr>
<tr>
<td>S-SM1</td>
<td>Syn-WH6501</td>
<td>9/17/2001</td>
<td>Off Western Shelf North</td>
<td>174.08</td>
<td>NA</td>
<td>NA</td>
<td>234</td>
<td>41.1</td>
<td>6</td>
<td>21</td>
<td>NC_015282</td>
</tr>
<tr>
<td>S-SM2</td>
<td>Syn-WH8017</td>
<td>9/17/2001</td>
<td>Off Western Shelf North</td>
<td>190.80</td>
<td>NA</td>
<td>NA</td>
<td>269</td>
<td>40.4</td>
<td>11</td>
<td>22</td>
<td>NC_015279</td>
</tr>
<tr>
<td>Syn33</td>
<td>Syn-WH7803</td>
<td>1/1/1995</td>
<td>Gulf Stream</td>
<td>174.28</td>
<td>NA</td>
<td>NA</td>
<td>227</td>
<td>39.6</td>
<td>5</td>
<td>17</td>
<td>NC_015285</td>
</tr>
<tr>
<td>Syn30</td>
<td>Syn-WH7803</td>
<td>12/1/1987</td>
<td>NE Providence Channel</td>
<td>178.81</td>
<td>NA</td>
<td>NA</td>
<td>213</td>
<td>39.9</td>
<td>6</td>
<td>20</td>
<td>NC_021072</td>
</tr>
<tr>
<td>S-RIM8 A.HR1</td>
<td>Syn-WH7803</td>
<td>9/23/1999</td>
<td>Narragansett Bay RI</td>
<td>171.21</td>
<td>NA</td>
<td>NA</td>
<td>211</td>
<td>40.6</td>
<td>8</td>
<td>16</td>
<td>NC_020486</td>
</tr>
<tr>
<td>S-PM1</td>
<td>Pro-MED4</td>
<td>3/9/2006</td>
<td>North Atlantic</td>
<td>181.04</td>
<td>NA</td>
<td>NA</td>
<td>241</td>
<td>37.8</td>
<td>0</td>
<td>17</td>
<td>NC_015280</td>
</tr>
<tr>
<td>P-HM2</td>
<td>Pro-MED4</td>
<td>3/10/2006</td>
<td>North Atlantic</td>
<td>183.81</td>
<td>NA</td>
<td>NA</td>
<td>242</td>
<td>38.1</td>
<td>0</td>
<td>16</td>
<td>NC_015284</td>
</tr>
<tr>
<td>P-RSM1</td>
<td>Pro-MED4</td>
<td>3/9/2006</td>
<td>North Atlantic</td>
<td>180.98</td>
<td>NA</td>
<td>NA</td>
<td>220</td>
<td>37.8</td>
<td>0</td>
<td>17</td>
<td>NC_020845</td>
</tr>
<tr>
<td>P-RSM4</td>
<td>Pro-MIT9303</td>
<td>9/13/2000</td>
<td>Red Sea Indian</td>
<td>177.21</td>
<td>NA</td>
<td>NA</td>
<td>213</td>
<td>40.2</td>
<td>2</td>
<td>18</td>
<td>NC_021071</td>
</tr>
<tr>
<td>P-RSM6</td>
<td>Pro-NATL2A</td>
<td>9/13/2000</td>
<td>Red Sea Indian</td>
<td>176.43</td>
<td>NA</td>
<td>NA</td>
<td>240</td>
<td>37.6</td>
<td>3</td>
<td>19</td>
<td>NC_015283</td>
</tr>
<tr>
<td>P-SSM3</td>
<td>Pro-NATL2A</td>
<td>5/6/1996</td>
<td>North Atlantic</td>
<td>192.50</td>
<td>NA</td>
<td>NA</td>
<td>221</td>
<td>39.3</td>
<td>3</td>
<td>18</td>
<td>NC_020855</td>
</tr>
<tr>
<td>P-SSM7</td>
<td>Pro-NATL1A</td>
<td>1/9/1999</td>
<td>North Atlantic</td>
<td>182.18</td>
<td>NA</td>
<td>NA</td>
<td>237</td>
<td>37.1</td>
<td>4</td>
<td>23</td>
<td>NC_015290</td>
</tr>
</tbody>
</table>

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 deoxyxynucleotide 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.)
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.
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, thioreredoxin, 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 $pcaA$, $psbD$ and $psbS$ (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 $psbS$ (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 naive 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 $carboxomyltransferase$ (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
the best polynomial thus it was assigned to two distinct LCBs. The grey lines indicate 2nd order polynomial greater than one because a breakpoint between two LCBs intersected the gene and still found near the homing endonuclease.

While the invading AMG (i.e. insertion of \textit{psbA}) in S-CAM7 is not unique to the T4-like myovirus clade (Fig. 3), 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 \textit{psbA} is not unique to the T4-like myovirus group. Podoviruses (Dekel-Bird et al., 2013; Sullivan et al., 2006) and other divergent myoviruses (Sabelh et al., 2012) also possess the gene. Cyanophage encoded psbA is thought to overcome a metabolic bottleneck during infection at high light, where photosynthesis is reduced in the photosystem II reaction center inhibits \textit{psbA} (Table 1). S-CAM7 belongs to a divergent clade (Fig. 3), but other members of this clade are replete in AMG content.

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 \textit{psbA} in S-SSM1 and S-CAM22, insertion of \textit{FECH} in Syn30 and P-SSM7, and insertion of \textit{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 cyanomyoviruses (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. carbonoyltransferase, \textit{pmA}, \textit{purM}, \textit{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 \textit{psbA}, \textit{maZG}, \textit{phoH}, \textit{cobS} and \textit{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 \textit{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 \textit{psbA} is not unique to the T4-like myovirus group. Podoviruses (Dekel-Bird et al., 2013; Sullivan et al., 2006) and other divergent myoviruses (Sabelh et al., 2012) also possess the gene. Cyanophage encoded \textit{psbA} is thought to overcome a metabolic bottleneck during infection at high light, where photosynthesis is reduced in the photosystem II reaction center inhibits

\textit{psbA} (Bragg and Chisholm, 2008; Hellweguer, 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 \textit{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-SSM2, 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 \textit{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
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 μE 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 x 10⁶ phage/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 purMs 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 ≥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 superfam topology (Ignacio-Espinoza and Sullivan, 2012). The genes include those encoding gp6 baseplate wedge, gp22 head 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 +Γ4 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 co-linear 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 cyanomvirus 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 re-

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


Arbiol, C., Comeau, A., Kutateladze, M., Adamia, R., Krisch, H., 2010. Mobile reg-


Breitbart, M., Thompson, L., Stute, C., Sullivan, M., 2007. Exploring the vast di-

versity of marine viruses. Oceanography 20, 135–139.


Comeau, A., Bertrand, C., Letarov, A., Tetart, F., Krisch, H., 2007. Modular archi-
tecture of the T4 phage superfamily: a conserved core genome and a plastic peripheral. Virology 362, 384–396.


Hagay, E., Mandel-Gutfreund, Y., Béja, O., 2014. Comparative metagenomics ana-
lyses reveal viral-induced shifts of host metabolism towards nucleotide bio-
synthesis. Microbiota 2, 5.


Hendrix, R., Smith, M., Burns, R., Ford, M., Hatfull, G., 1999. Evolutionary relation-


Ignacio-Espinoza, J.C., Sullivan, M.B., 2012. Phylogenomics of T4 cyanophages:

OECD-1029684 and OECD-1332782 to MM; OECD-1031783 and OECD-1332740 to JM) and the RWU Foundation to Promote Scholarship to MM.