Environmental Microbiology (2008) **10**(10), 2810–2823 doi:10.1111/j.1462-2920.2008.01702.x

Portal protein diversity and phage ecology

Matthew B. Sullivan,[1] Maureen L. Coleman,[1] Vanessa Quinlivan,[2] Jessica E. Rosenkrantz,[2] Alicia S. DeFrancesco,[2] G. Tan,[2] Ross Fu,[1] Jessica A. Lee,[2] John B. Waterbury,[3] Joseph P. Bielawski[4] and Sallie W. Chisholm[1,2]*

Departments of [1]Civil and Environmental Engineering, [2]Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[3]Department of Biology, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.

[4]Department of Biology, Dalhousie University, Nova Scotia, Canada.

Summary

Oceanic phages are critical components of the global ecosystem, where they play a role in microbial mortality and evolution. Our understanding of phage diversity is greatly limited by the lack of useful genetic diversity measures. Previous studies, focusing on myophages that infect the marine cyanobacterium *Synechococcus*, have used the coliphage T4 portal-protein-encoding homologue, gene 20 (*g20*), as a diversity marker. These studies revealed 10 sequence clusters, 9 oceanic and 1 freshwater, where only 3 contained cultured representatives. We sequenced g20 from 38 marine myophages isolated using a diversity of *Synechococcus* and *Prochlorococcus* hosts to see if any would fall into the clusters that lacked cultured representatives. On the contrary, all fell into the three clusters that already contained sequences from cultured phages. Further, there was no obvious relationship between host of isolation, or host range, and g20 sequence similarity. We next expanded our analyses to all available g20 sequences (769 sequences), which include PCR amplicons from wild uncultured phages, non-PCR amplified sequences identified in the Global Ocean Survey (GOS) metagenomic database, as well as sequences from cultured phages, to evaluate the relationship between g20 sequence clusters and habitat features from which the phage sequences were isolated. Even in this meta-data set, very few sequences fell into the sequence clusters without cultured representatives, suggesting that the latter are very rare, or sequencing artefacts. In contrast, sequences most similar to the culture-containing clusters, the freshwater cluster and two novel clusters, were more highly represented, with one particular culture-containing cluster representing the dominant g20 genotype in the unamplified GOS sequence data. Finally, while some g20 sequences were non-randomly distributed with respect to habitat, there were always numerous exceptions to general patterns, indicating that phage portal proteins are not good predictors of a phage's host or the habitat in which a particular phage may thrive.

Virus-like particles occur in high abundance (to 10^8 ml^{-1}) in the oceans (Bergh, 1989; Bratbak *et al*., 1990; Proctor and Fuhrman, 1990). One of the most well-studied phage–host systems in this habitat is the phages that infect the marine cyanobacteria *Prochlorococcus* and *Synechococcus*, which are globally important marine primary producers (Waterbury *et al*., 1986; Partensky *et al*., 1999). These 'cyanophages' are abundant (Waterbury and Valois, 1993; Suttle and Chan, 1994; Suttle, 2000; Lu *et al*., 2001; Frederickson *et al*., 2003; Marston and Sallee, 2003; Sullivan *et al*., 2003), contribute to host mortality (Waterbury and Valois, 1993; Suttle and Chan, 1994; Suttle, 2000) and are thought to play a role in maintaining the extensive microdiversity of their hosts (Waterbury and Valois, 1993; Suttle and Chan, 1994; Marston and Sallee, 2003; Sullivan *et al*., 2003) likely through killing the winner (*sensu* Thingstad, 2000) and through the movement of genes throughout the host population (Lindell *et al*., 2004; Coleman *et al*., 2006; Sullivan *et al*., 2006).

Studying the diversity of phages has proven difficult because no universal gene, analogous to the 16S rRNA gene used for microbes, exists throughout all phage families (Paul *et al*., 2002). Thus family-specific genes have been proposed for use as taxonomic tools in phage ecology (Rohwer and Edwards, 2002). One such marker, a homologue to the coliphage T4 portal protein gene 20 (g20), has been developed to study the diversity of

Received 10 February, 2008; accepted 7 June, 2008. *For correspondence. E-mail chisholm@mit.edu; Tel. (+1) 617 253 1771; Fax (+1) 617 324 0336.

Myoviridae – one of the most common phage types observed in metagenomics surveys (Breitbart *et al*., 2002; 2004a; DeLong *et al*., 2006) and among *Synechococcus* cyanophage isolates (Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson *et al*., 1993; Sullivan *et al*., 2003). The g20 homologue is ubiquitous among T4-like myoviruses (see T4-like phages genome website http://phage.bioc.tulane.edu/) with hosts ranging from proteobacteria to cyanobacteria (Fuller *et al*., 1998; Hambly *et al*., 2001; Mann *et al*., 2005; Sullivan *et al*., 2005). The evolution of g20 is likely constrained because its protein product initiates capsid assembly (at least in T4), a process which involves geometric precision (Coombs and Eiserling, 1977; van Driel and Couture, 1978; Hsiao and Black, 1978) through the formation of a proximal vertex (van Driel and Couture, 1978) used for DNA packaging (Hsiao and Black, 1978) and binding the capsid to the tail junction (Coombs and Eiserling, 1977).

The availability of cultured cyanomyophage (Waterbury and Valois, 1993; Wilson *et al*., 1993; Suttle and Chan, 1993; Marston and Sallee, 2003; Sullivan *et al*., 2003) has allowed the design of cyanomyophage-specific g20 sequence PCR primers that have been used to study this component of viral populations in the wild. Early studies using non-degenerate PCR primers and DNA 'fingerprinting' techniques (e.g. denaturing gradient gel electrophoresis and terminal-restriction fragment length polymorphism banding patterns) revealed variability in g20 diversity across gradients in space and time from a variety of different environments (Wilson *et al*., 1999; 2000; Frederickson *et al*., 2003; Dorigo *et al*., 2004; Wang and Chen, 2004; Mühling *et al*., 2005; Sandaa and Larsen, 2006). These studies concluded that g20 diversity was as great within a sample as between oceans (Wilson *et al*., 1999), that phage g20 diversity increased as *Synechococcus* abundance increased (Wilson *et al*., 1999; 2000; Frederickson *et al*., 2003; Wang and Chen, 2004; Sandaa and Larsen, 2006), that some g20 types were ubiquitous in the habitats examined (Wilson *et al*., 1999; 2000; Frederickson *et al*., 2003; Dorigo *et al*., 2004), as well as a temporal study by Muhling and colleagues (2005) that correlated 'cyanophage' diversity (inferred from g20 sequence types) with *Synechococcus* diversity (inferred from *rpoC1* sequence types).

Subsequent cloning and sequencing of g20 PCR amplicons from both cultured isolates and wild populations have allowed phylogenetic analyses of cyanomyophage diversity. Although initial studies (Zhong *et al*., 2002) suggested some correlation between ocean habitat and g20 phylogeny (e.g. phylogenetic cluster II represents 'open ocean' g20 sequences), further sampling revealed that this was not the case, as seven g20 sequences from coastal *Synechococcus* myophages isolated from Rhode Island waters clustered with the putative 'open ocean' sequences (Marston and Sallee, 2003). As more g20 sequence data have accumulated from diverse environments (Zhong *et al*., 2002; Marston and Sallee, 2003; Dorigo *et al*., 2004; Short and Suttle, 2005; Sandaa and Larsen, 2006; Wilhelm *et al*., 2006), it has become clear that marine g20 sequences form nine phylogenetic clusters (first described by Zhong *et al*., 2002), and g20 sequences originating from freshwater environments form a separate, tenth cluster (Dorigo *et al*., 2004; Short and Suttle, 2005; Wilhelm *et al*., 2006). Three of the nine marine clusters (clusters I–III in Zhong *et al*., 2002) contain cultured representatives (hereafter called 'culture-containing clusters'), whereas the remaining six marine clusters (clusters A-F) and the 'freshwater' cluster do not (hereafter called 'environmental-sequence-only clusters'). The cultured representatives were isolated using only *Synechococcus* hosts (7 strains = WH7803, WH7805, WH8007, WH8012, WH8018, WH8101, WH8113), which undoubtedly limits the diversity represented considering the larger diversity of *Synechococcus* strains (Rocap *et al*., 2002; Fuller *et al*., 2003; Ahlgren and Rocap, 2006) and that the sister genus *Prochlorococcus* is also abundant in open ocean waters. This raises the question: could these seven environmental-sequence-only clusters represent novel cyanomyophages that infect this broader diversity of *Synechococcus* host strains, *Prochlorococcus* or other cyanobacteria?

To address this question, we isolated phages on a broad diversity of *Prochlorococcus* and *Synechococcus* hosts (Table 1), sequenced their g20 homologues and analysed their diversity in the context of published PCR-generated sequences from natural populations. We then combined the g20 sequences from these new cultured isolates with all environmental g20 sequences available [including all PCR-generated environmental sequences, as well as primer-independent sequences available in the Global Ocean Survey (GOS) metagenomic data set], to examine the broad diversity of g20 observed in the wild. This allowed us to ask: do any of the new environmental sequences cluster with the previously observed environmental-sequence-only clusters? Furthermore, are g20 sequence clustering patterns ecologically meaningful? Do they reflect the habitat – and by inference the microbial community – of the site from which they were isolated?

Results and discussion

Analysis of g20 diversity captured by several g20 primer sets

As our understanding of marine myoviruses has grown over the years, multiple primer sets have been developed

Table 1. Efficacy of three different primer sets at amplifying the g20 gene from cultured cyanophage.

Phage strain	Original host strain isolated on	Site of Isolation	Depth (m)	Date isolated	Family[a]	g20 primer set			Refs[b]
						CPS4GC/5	CPS1/8	CPS1.1/8.1	
Prochlorococcus cyanophage									
P-SSP1	MIT 9215	BATS/31°48′N, 64°16′W	100	6 June 2000	P	–	–	–	1
P-RSP1	MIT 9215	Red Sea/29°28′N, 34°53′E	0	15 July 2000	P	–	–	–	1
P-RSP2	MIT 9302	Red Sea/29°28′N, 34°53′E	0	15 July 2000	P	–	–	–	1
P-SSP2	MIT 9312	BATS/31°48′N, 64°16′W	120	29 September 1999	P	–	–	–	1
P-SSP3	MIT 9312	BATS/31°48′N, 64°16′W	100	29 September 1999	P	–	–	–	1
P-SSP4	MIT 9312	BATS/31°48′N, 64°16′W	70	26 September 1999	P	–	–	–	1
P-SSP5	MIT 9515	BATS/31°48′N, 64°16′W	120	29 September 1999	P	–	–	–	1
P-SSP6	MIT 9515	BATS/31°48′N, 64°16′W	100	26 September 1999	P	–	–	–	1
P-SSP7	MED4	BATS/31°48′N, 64°16′W	100	26 September 1999	P	–	–	–	1
P-GSP1	MED4	Gulf Stream/38°21′N, 66°49′W	40	6 October 1999	P	–	–	–	1
P-SSP8	NATL2A	BATS/31°48′N, 64°16′W	100	26 September 1999	P	–	–	–	1
P-RSP3	NATL2A	Red Sea/29°28′N, 34°55′E	50	13 September 2000	P	–	–	–	1
P-SP1	SS120	Slope/38°10′N, 73°09′W	83	17 September 2001	P	–	–	–	1
P-SSM8	**MIT 9211**	**W Sargasso Sea/34°24′N, 72°03′W**	**30**	**22 September 2001**	**M**	+	+	+	**2**
P-SSM1	**MIT 9303**	**BATS/31°48′N, 64°16′W**	**100**	**6 June 2000**	**M**	+	–	+	**1**
P-RSM1	**MIT 9303**	**Red Sea/29°28′N, 34°53′E**	**0**	**15 July 2000**	**M**	+	–	+	**1**
P-RSM4	**MIT 9303**	**Red Sea/29°28′N, 34°55′E**	**130**	**13 September 2000**	**M**	+	+	+	**2**
P-ShM1	**MIT 9313**	**Shelf/39°60′N, 71°48′W**	**40**	**16 September 2001**	**M**	+	–	+	**1**
P-ShM2	**MIT 9313**	**Shelf/39°60′N, 71°48′W**	**0**	**16 September 2001**	**M**	–	–	+	**1**
P-SSM2	**NATL1A**	**BATS/31°48′N, 64°16′W**	**100**	**6 June 2000**	**M**	+	+	+	**1**
P-RSM5	**NATL1A**	**Red Sea/29°28′N, 34°55′E**	**130**	**13 September 2000**	**M**	+	+	+	**2**
P-SSM7	**NATL1A**	**BATS/31°48′N, 64°16′W**	**120**	**29 September 1999**	**M**	–	–	+	**2**
P-SSM3	**NATL2A**	**BATS/31°48′N, 64°16′W**	**100**	**6 June 2000**	**M**	–	–	+	**1**
P-SSM4	**NATL2A**	**BATS/31°48′N, 64°16′W**	**10**	**6 June 2000**	**M**	–	–	+	**1**
P-SSM5	**NATL2A**	**BATS/31°48′N, 64°16′W**	**15**	**26 September 1999**	**M**	+	–	+	**1**
P-SSM6	**NATL2A**	**BATS/31°48′N, 64°16′W**	**40**	**29 September 1999**	**M**	–	–	+	**1**
P-RSM2	**NATL2A**	**Red Sea/29°28′N, 34°55′E**	**50**	**13 September 2000**	**M**	+	–	+	**1**
P-RSM3	**NATL2A**	**Red Sea/29°28′N, 34°55′E**	**50**	**13 September 2000**	**M**	–	–	+	**1**
P-SSM9	**NATL2A**	**W Sargasso Sea/34°24′N, 72°03′W**	**0**	**22 September 2001**	**M?**	+	–	+	**2**
P-SSM10	**NATL2A**	**W Sargasso Sea/34°24′N, 72°03′W**	**0**	**22 September 2001**	**M?**	+	–	+	**2**
P-SSM11	**NATL2A**	**W Sargasso Sea/34°24′N, 72°03′W**	**0**	**22 September 2001**	**M?**	+	–	+	**2**
P-SSM12	**NATL2A**	**W Sargasso Sea/34°24′N, 72°03′W**	**95**	**22 September 2001**	**M?**	+	–	+	**2**
Synechococcus cyanophage									
Syn5	WH 8109	Sargasso Sea/36°58′N, 73°42′W	0	December 1990	P	–	–	–	1
Syn12	WH 8017	Gulf Stream/34°06′N, 61°01′W	0	July 1990	P	–	–	–	1
S-SM1	**WH 6501**	**Slope/38°10′N, 73°09′W**	**0**	**17 September 2001**	**M**	–	–	+	**1**
S-ShM1	**WH 6501**	**Shelf/39°60′N, 71°48′W**	**0**	**16 September 2001**	**M**	+	+	+	**1**
S-SSM1	**WH 6501**	**W Sargasso Sea/34°24′N, 72°03′W**	**70**	**22 September 2001**	**M**	+	+	+	**1**
Syn 2	**WH 8012**	**Sargasso Sea/34°06′N, 61°01′W**	**0**	**July 1990**	**M**	–	+	+	**3**
Syn 9	**WH 8012**	**Woods Hole/41°31′N, 71°40′W**	**0**	**October 1990**	**M**	+	+	+	**3**

Syn 10	**WH 8017**	**Gulf Stream/36°58′N, 73°42′W**	**0**	**December 1990**	**M**	+	+	+	**3**
Syn 26	**WH 8017**	**NE Providence Channel/25°53′N, 77°34′W**	**0**	**January 1992**	**M**	+	+	+	**3**
S-SM2	**WH 8017**	**Slope/38°10′N, 73°09′W**	**15**	**17 September 2001**	**M**	+	–	+	**2**
Syn30	**WH 8018**	**NE Providence Channel/25°53′N, 77°34′W**	**0**	**January 1992**	**M**	+	–	+	**3**
S-SSM3	**WH 8018**	**W Sargasso Sea/34°24′N, 72°03′W**	**0**	**22 September 2001**	**M**	+	+	+	**2**
S-SSM4	**WH 8018**	**W Sargasso Sea/34°24′N, 72°03′W**	**110**	**22 September 2001**	**M**	+	+	+	**2**
S-RIM3	WH 8018	Mt. Hope Bay, RI/41°39′N, 71°15′W	0	September 1999	M?	+	–	+	4
Syn 33	**WH 7803**	**Gulf Stream/25°51′N, 79°26′W**	**0**	**January 1995**	**M**	+	+	+	**3**
S-PM2	WH 7803	English Channel/50°18′N, 4°12′W	0	23 September 1992	M	+	+	+	5
S-WHM1	WH 7803	Woods Hole/41°31′N, 71°40′W	0	11 August 1992	M	+	+	+	5
S-RIM9	WH 7803	Mt. Hope Bay, RI/41°39′N, 71°15′W	0	May 2000	M?	+	–	+	4
S-RIM17	WH 7803	Mt. Hope Bay, RI/41°39′N, 71°15′W	0	July 2001	M?	+	–	+	4
S-RIM24	WH 7803	Mt. Hope Bay, RI/41°39′N, 71°15′W	0	December 2001	M?	+	–	+	4
S-RIM30	WH 7803	Mt. Hope Bay, RI/41°39′N, 71°15′W	0	June 2002	M?	+	–	+	4
Syn 1	**WH 8101**	**Woods Hole/41°31′N, 71°40′W**	**0**	**August 1990**	**M**	+	–	+	**3**
S-ShM2	**WH 8102**	**Shelf/39°60′N, 71°48′W**	**0**	**16 September 2001**	**M**	+	+	+	**1**
S-SSM2	**WH 8102**	**W Sargasso Sea/34°24′N, 72°03′W**	**0**	**22 September 2001**	**M**	+	+	+	**1**
S-SSM5	**WH 8102**	**W Sargasso Sea/34°24′N, 72°03′W**	**95**	**22 September 2001**	**M**	+	+	+	**2**
Syn 19	**WH 8109**	**Sargasso Sea/34°06′N, 61°01′W**	**0**	**July 1990**	**M**	–	–	+	**3**
S-SSM6	**WH 8109**	**W Sargasso Sea/34°24′N, 72°03′W**	**70**	**22 September 2001**	**M**	+	+	+	**2**
S-SSM7	**WH 8109**	**W Sargasso Sea/34°24′N, 72°03′W**	**95**	**22 September 2001**	**M**	+	+	+	**2**
Other phages									
IH6-ϕ1	IH6	Inner Harbor, Baltimore, MD	0	17 November 2000	M	–	–	–	6
IH6-ϕ7	IH6	Inner Harbor, Baltimore, MD	0	17 November 2000	P	–	–	–	6
IH11-ϕ2	*Alteromonas*	Inner Harbor, Baltimore, MD	0	17 November 2000	M	–	–	–	6
IH11-ϕ5	*Alteromonas*	Inner Harbor, Baltimore, MD	0	17 November 2000	P	–	–	–	6
CB8-ϕ2	CB8	Chesapeake Bay, MD	0	17 November 2000	M	–	–	–	6
CB8-ϕ6	CB8	Chesapeake Bay, MD	0	17 November 2000	M	–	–	–	6
CB-ϕ8	*Vibrio alginolyticus*	Chesapeake Bay, MD	0	17 November 2000	M	–	–	–	6
HER320	H7	Helgoland, North Sea	0	1976–1978	M	–	–	–	7
HER321	H100	Helgoland, North Sea	0	1976–1978	P	–	–	–	7
HER322	H100	Helgoland, North Sea	0	1976–1978	M	–	–	–	7
HER327	11–68	Helgoland, North Sea	0	1976–1978	S	–	–	–	7
HER328	H105	Helgoland, North Sea	0	1976–1978	S	–	–	–	7

a. M, P and S represent the virus families *Myoviridae*, *Podoviridae* and *Siphoviridae* respectively, as determined by morphology. 'M?' indicates that the assignment is based solely on amplification and sequencing of a g20 PCR product and has not been confirmed with electron microscopy.

b. Reference where cultured isolate was originally described: 1, Sullivan and colleagues (2003); 2, this study; 3, Waterbury and Valois (1993); 4, Marston and Salee (2003); 5, Wilson and colleagues (1993); 6, Zhong and colleagues (2002); 7, Wichels and colleagues (1998).

'+' indicates positive PCR amplification; '–' indicates that there was no PCR product of the expected size. The new g20 sequences contributed in this study are shown in bold letters. CPS1.1/8.1 is the new primer set designed for this study, while CPS4GC/5 and CPS1/8 were published previously.

Fig. 1. Evolutionary relationships determined using 183 amino acids of the portal protein gene (g20) amplified from cultured phage isolates (names begin with 'S-' or 'P-' and are coloured orange or green for *Synechococcus* or *Prochlorococcus* phages respectively) from this study (italicized), as well as previous studies (non-italicized), and environmental g20 sequences (names in black) (Zhong *et al*., 2002; Marston and Sallee, 2003). Clusters defined by Zhong and colleagues (2002) are as follows: clusters I–III contain g20 sequences from cultured phage isolates, while clusters A–F represent only environmental g20 sequences. Clusters containing identical g20 protein sequences are numbered with alphanumeric numbers (1–13). For cultured phages, the phage isolate names are followed by black lettering that indicates the original host strain used for isolation, while the phage host range is indicated as high light-adapted *Prochlorococcus* (green circle or dash), low light-adapted *Prochlorococcus* (blue circle or dash) or *Synechococcus* (orange circle or dash). The circles represent cross-infection was observed within this group of hosts tested, whereas a dash indicates that no cross-infection was observed. Isolates not available for host range testing have no indication of their host range. The tree shown was inferred by neighbour-joining as described in the *Experimental procedures*. Support values shown at the nodes are neighbour-joining bootstrap/maximum parsimony bootstrap/maximum likelihood quartet puzzling support (only values > 50 are shown). Well-supported nodes (as defined in *Experimental procedures*) are designated by italicized support values, including six nodes that represent subclusters within the culture-containing clusters I–III. The g20 sequence from the non-cyanomyophage isolate T4 was used as an outgroup to root this tree.

and used to specifically amplify cyanomyophage g20 sequences from field samples (Fuller *et al*., 1998; Wilson *et al*., 1999; 2000; Zhong *et al*., 2002; Frederickson *et al*., 2003; Marston and Sallee, 2003; Dorigo *et al*., 2004; Wang and Chen, 2004; Sandaa and Larsen, 2006; Wilhelm *et al*., 2006). Each of these primer sets was designed based on a limited number of sequences from cultured isolates. Thus we wondered how well these primer sets would capture the diversity of g20 sequences in our relatively extensive *Prochlorococcus* and *Synechococcus* cyanophage collection (Table 1).

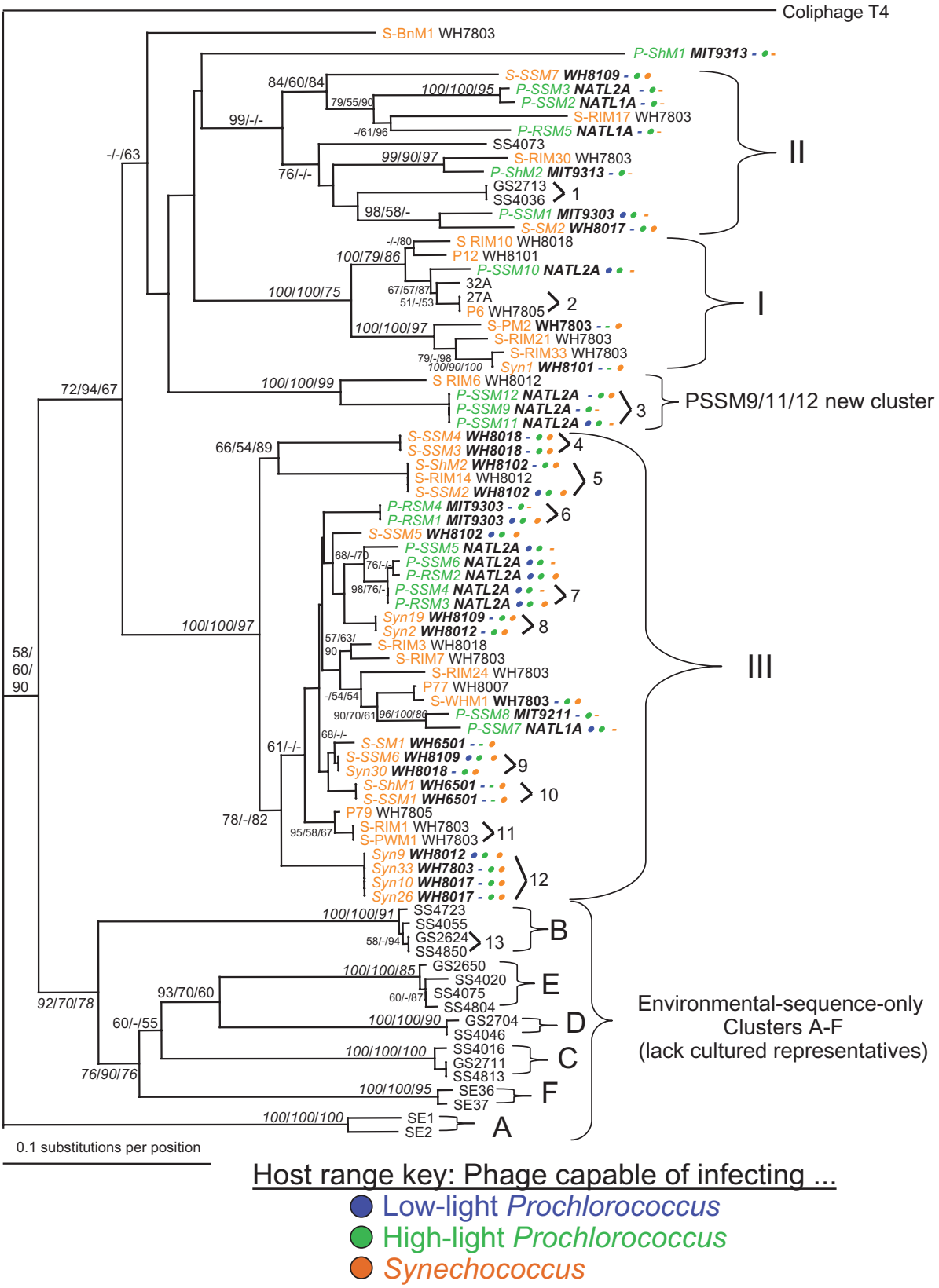
We found that the CPS4GC/5 primer set (Wilson *et al*., 1999) amplified g20 sequences from 80% of the cyanomyophages screened (bold entries in Table 1). This primer set, however, amplifies only a small region of this gene (~165 bp), thus its utility for subsequent phylogenetic analyses is limited. In contrast, the CPS1/8 primer set (Zhong *et al*., 2002), which captures a larger segment of the gene (~594 bp), amplified the g20 sequence of only 56% of the cyanomyophages screened (Table 1). Using genome sequence data from two *Prochlorococcus* cyanomyophages (Sullivan *et al*., 2005) that became available after these primer sets were designed, we modified the CPS1/8 primer set with the hope of amplifying g20 from all of our isolates for use in subsequent phylogenetic analyses. Indeed, the redesigned set (CPS1.1/8.1) captured g20 homologues from all cyanomyophage isolates screened (Table 1). Despite their degeneracy, the redesigned primer set remained specific only for cyanomyophage isolates as inferred from repeatedly negative PCR results against the sipho- and podo-cyanophage, as well as the non-cyanomyophages we examined (Table 1).

Phylogenetic relationships of g20 sequences

We next analysed how these new g20 sequences from cultured isolates compared with selected sequences (see *Experimental procedures*) from the databases (Fig. 1). Randomly paired g20 sequence identities from this data set ranged from 59% to 100% amino acid identity, notably with some identical g20 protein sequences observed multiple times (alphanumeric clusters #1–13 in Fig. 1). This is not unprecedented: even at the level of the gene, identical viral sequences have been previously reported from vastly different aquatic environments using two separate gene markers including g20 (Zhong *et al*., 2002; Marston and Sallee, 2003; Short and Suttle, 2005) and DNA polymerase (Breitbart *et al*., 2004b; Breitbart and Rohwer, 2005).

In phylogenetic analyses, 40 of 45 g20 sequences from cyanomyophages (38 new, 7 previously published) grouped within the clusters that contain cultured representatives (I, II and III), four fell into a new monophyletic cluster (indicated by 'PSSM9/11/12 new cluster' on Fig. 1), and one (P-ShM1) fell onto a long branch. None fell into the previously defined (by Zhong *et al*., 2002) environmental-sequence-only clusters A–F, which were thought to be from marine cyanomyophages because of the use of isolate-designed and -tested 'cyanophage-specific primers'. Thus either our phage culture collection is still not diverse enough to represent the g20 diversity of phages that infect marine cyanobacteria, or the sequences in the environmental-sequence-only clusters A–F represent myophages that infect other hosts. Observations made by Short and Suttle (2005) lend support to the latter. They found three g20 sequences in waters 3246 m deep in the Arctic Chukchi Sea, waters unlikely to contain cyanobacteria and their phages, which grouped with cluster A.

Given our extensive host range information for these cyanobacteria phage–host systems, we examined g20 clustering patterns for relationships with respect to the host strains upon which the phage were isolated or could cross-infect. None of the three culture-containing clusters (I, II, III) were comprised solely of g20 sequences from phages with similar hosts (Fig. 1), and no clear-cut patterns emerged when subclusters within these clusters were evaluated. This is consistent with the observations of Stoddard and colleagues (2007), who recently reported that g20 sequences could not predict the pattern of cross-resistance observed when selecting for cyanophage



Host range key: Phage capable of infecting ...

- Low-light *Prochlorococcus*
- High-light *Prochlorococcus*
- *Synechococcus*

Table 2. Origins of the g20 sequences used in 'meta' phylogenetic analyses shown in Fig. 2.

# Sequences	Description	PCR-based?	Sequence label in Fig. 2	Refs
512	Environmental sequences from 42 oceanic sample sites from the GOS	N	JC#	1
56	Environmental sequences from 19 globally distributed freshwater and marine sites	Y	AY705#	2
25	Environmental sequences from Rhode Island coastal waters, USA	Y	AY259#	3
43	Environmental sequences from Lake Erie, USA	Y	DQ318#	4
47	Environmental sequences from Lake Bourget, France	Y	AY426#	5
27	Environmental sequences and mixed lysates from coastal north-western Atlantic Ocean	Y	Variable	6
51	Cultured marine cyanomyophages of variable coastal and open ocean origins	N/A	Variable	3, 7
8	Cultured non-cyanomyophages from sewage	N/A	Variable	8

The 'PCR-based' column indicates whether the environmental sequence was obtained by PCR or metagenomic approaches (N/A indicates that this is not applicable for sequences from cultured phage isolates). Reference code: 1, Rusch and colleagues (2007); 2, Short and Suttle (2005); 3, Marston and Salee (2003); 4, Wilhelm and colleagues (2006); 5, Dorigo and colleagues (2004); 6, Zhong and colleagues (2002); 7, this study; 8, T4-like phage genomes website http://phage.bioc.tulane.edu/

resistance in *Synechococcus.* Conversely, they also found that *Synechococcus* DNA-dependent RNA polymerase genotypes were not related to phage sensitivities (Stoddard *et al*., 2007). Thus for the *Prochlorococcus*/*Synechococcus*/myophage system in Fig. 1, it appears that commonly used phage and host genetic markers lack the ability to predict either the range of hosts that a phage can infect, or the range of phages to which a host is susceptible.

We next added more recently published g20 sequences to this analysis, including those from the non-PCR-based GOS metagenomics database (Rusch *et al*., 2007) and all published PCR-based environmental sequences (Fig. 2, Table 2). Only sequences of sufficient length for phylogenetic analysis were used. The majority (464 of 769) of these environmental sequences, including 401 GOS sequences, grouped in culture-containing clusters I, II and III. First we found that 13 of the 38 GOS sample sites included in our analysis lack *Prochlorococcus* and *Synechococcus* (as determined by dot-blots in Rusch *et al*., 2007), yet 75 g20 sequences from these sites fell into clusters I, II and III (Fig. 2), thought, from earlier studies, to represent myophages that infect marine pico-cyanobacteria. Thus it appears that clusters I, II and III likely represent phages that infect a diversity of hosts and are not limited to pico-cyanobacteria-dominated environments. Second, these analyses revealed that cluster II contains ~10-fold more GOS sequences than clusters I and III (336 versus 32 and 33 respectively). If we ignore possible cloning bias, this suggests that cluster II sequences are by far the most abundant type in the environments sampled. Third, we note that a relatively tiny number of the GOS sequences fell into the environmental-sequence-only clusters – clusters A–F in Fig. 1 – that were defined by Zhong and colleagues (2002) (Fig. 2). The 12 that fell into cluster A originated from seven sites with different physicochemical characteristics (see colour rings, Fig. 2). Even fewer sequences fell into environmental-sequence-only clusters B–F, suggesting that these types of g20 sequences are either extremely rare in the environments sampled to date, or are sequencing artefacts.

This expanded data set lends support for three additional g20 lineages (Fig. 2). These include 93 sequences that group with the previously identified 'freshwater' cluster (Dorigo *et al*., 2004; Short and Suttle, 2005; Wilhelm *et al*., 2006; labelled as 'new cluster #1' in Fig. 2), 25 sequences that group with the new culture-containing P-SSM9/11/12 cluster (named after the original phage isolates forming this cluster in Fig. 1, labelled as 'new cluster #2' in Fig. 2) and 84 environmental sequences (74 GOS + 10 non-GOS environmental sequences, labelled as 'new cluster #3' in Fig. 2) of mixed biogeographic and habitat origin that form a new environmental-sequence-only cluster.

Relationship between g20 clusters and habitat

Using Unifrac distance metric statistical tools (Lozupone *et al*., 2006), we examined the meta-g20 data set for correlates between sequence clustering and habitat descriptors, such as the microbial community type, temperature and salinity of the original sample. As a first approximation of the microbial community type, we used previously defined environmental categories originally inferred from ribotype dot-blots and metagenomic sequence data (figs 9 and 10 in Rusch *et al*., 2007) for the GOS g20 sequences, then assigned such categories where reasonable assumptions could be made for non-GOS sequences (details in Table 3 legend). We found that the g20 sequence clusters were non-randomly distributed with respect to sequences that originated from freshwater, tropical freshwater, arctic/polar, estuarine, Sargasso and hypersaline environments, while eight other environments lacked statistically significant clustering (Table 3). Beyond habitat-related properties, we also

Table 3. Relationship between g20 sequence clusters and the microbial community types of the original habitats from which they were collected.

Environmental category	Unifrac *P*-value	# sequences	Qualitative relative abundance of dominant ribotypes inferred for the GOS samples in these categories[%]																			
			SAR11 – surface 1	SAR11 – surface 2	SAR11 – surface 3	*Prochlorococcus*	*Synechococcus*	Roseobacter	SAR86	SAR116	Archaea	Cytophaga	Rhodospirillaceae	Alphaproteobacteria	Gammaproteobacteria	Acidomicrobidae	Cellulomonadaceae	Chlorobi	Acidobacteria	Frankineae	Bdellovibrionales	Comamonadaceae
Temperate ocean – north	0.2736	84																				
Temperate ocean – south	0.0532	13																				
Tropical and Sargasso	0.5098	44																				
Tropical – open Ocean	0.8969	139																				
Tropical – near Galapagos	0.1411	116																				
May Sargasso Sea	0.0812	6																				
Coral reef atoll	0.0714	48																				
Fringing Reef	0.4238	34																				
Tropical freshwater	*0.0001*	*47*																				
Estuary	*0.0020*	*14*																				
Sargasso Sea	*0.0029*	*32*																				
Hypersaline	*0.0474*	*3*																				
Freshwater	*0.0009*	*99*		N	O		G	O	S		S	A	M	P	L	E		D	A	T	A	
Arctic/polar	*≤0.0001*	*19*		N	O		G	O	S		S	A	M	P	L	E		D	A	T	A	

a. Qualitative characterization of the relative abundance of dominant ribotypes using published data from the GOS (detailed data available in Rusch *et al.*, 2007). These data represent only those microbes captured in 0.1–0.8 μm size fraction samples, except for the Fringing Reef sample which is the 0.8–3.0 μm size fraction. No data are available for freshwater and arctic/polar samples because these were not part of the GOS sampling expedition.

Unifrac distance metric (Lozupone and Knight, 2005) was used for the analysis. A *P*-value < 0.05 (italicized) indicates that sequences from that category are non-randomly distributed with respect to habitat in the phylogenetic analysis. In the Unifrac analysis presented here, we used the environmental categories given to the GOS sample g20 sequences by Rusch and colleagues (2007) (inferred using ribotype dot-blots and shared metagenomic content; figs 9 and 10 in Rusch *et al.*, 2007), whereas we assumed which environmental category non-GOS sequences belonged to as follows: (i) Woods Hole, Plymouth, NE Providence Channel, Rhode Island waters were considered 'temperate ocean – north' (akin to GOS sample 8, Newport Harbor, RI), (ii) freshwater, the Sargasso Sea or estuaries were considered 'freshwater', 'Sargasso Sea' or 'estuary' respectively, (iii) arctic or polar water sequences were given their own category. We did not assume an environmental category for non-GOS samples originating from the Red Sea, Atlantic Ocean continental shelf and slope waters, Dauphin Island and Gulf Stream so they were not used in this analysis (temperature and salinity data were available for many of these samples, so they were used in subsequent analyses). A total of 698 categorized sequences were used in the Unifrac analysis. To provide an overall picture of the microbial community for each environmental category, we provide qualitative relative abundance microbial community data for each environmental category inferred from the ribotype data published for the GOS samples in Rusch and colleagues (2007) as follows: dark squares, highly dominant ribotypes; lighter squares, ribotypes that are present but not dominant; white squares, ribotype was not detected.

observed non-random g20 sequence distributions relative to abiotic factors, such as salinity (four of five categories significant, Table 4) and temperature (three of five categories significant, Table 5). In both cases, the outermost categories (e.g. 'cold' and 'hot', but not 'medium' for temperature) were significantly structured, but median categories were not. Qualitatively, some of these clustering patterns are also evident in the colour-coded rings in Fig. 2.

Notably, however, clustered sequences, when significantly correlated with a habitat characteristic, always contained exceptions. For example, the 'freshwater' category was one of the most significantly non-random sequence categories (Fig. 2, Tables 3–5). In spite of this, the 'freshwater' cluster also contained 6 sequences from brackish waters, while 68 additional freshwater sequences were distributed elsewhere in the tree (light blue in the outer circle in Fig. 2). Similarly, while sequences in the 'tropical freshwater' category were found to be non-randomly distributed (Table 3), this is likely driven by the 24 sequences that form a well-defined subcluster within cluster II (GOS site 20 subclus-

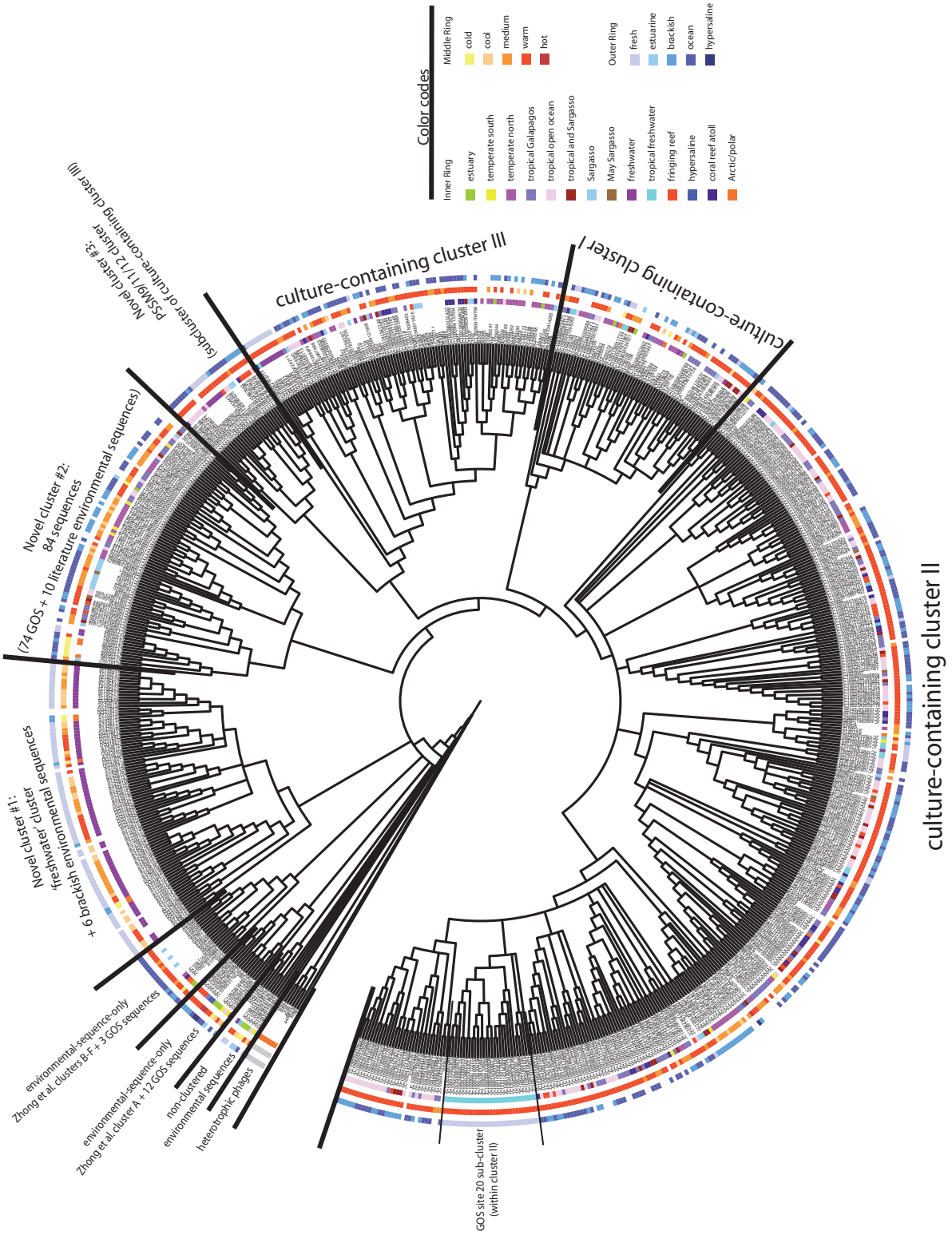


Table 4. Probability that g20 sequence clusters are non-random with respect to the salinity at the site from which they were collected.

Environmental category	Salinity (ppt)	# Sequences	Unifrac *P*-value
Sewage	*N/A*	*6*	*≤ 0.0001*
Fresh	*< 0.50*	*149*	*≤ 0.0001*
Estuarine	*0.5–17.99*	*6*	*0.0096*
Brackish	18–32.99	183	0.1456
Ocean	*33–38*	*286*	*0.0006*
Hypersaline	*> 38*	*8*	*0.0474*

The Unifrac distance metric (Lozupone *et al*., 2006) was used for the analysis. Salinity values, when not available from the published work, were obtained from the communicating author of the paper in which the g20 sequence was first reported. All freshwater samples were assumed to have a salinity of < 0.50 ppt. All but the sequences from brackish waters clustered non-randomly ($P < 0.05$) with respect to the habitat type as defined by salinity.

ter in Fig. 2). However, another 18 sequences from this same sample are scattered throughout the rest of the tree (11 in cluster II, 4 in cluster I and 3 in other clusters).

In other words, while some patterns emerge, exceptions are so frequent that one must conclude that the g20 sequence is not a good predictor of the habitat from which the phage originated. This is perhaps not surprising given the sheer abundance of phages on the planet (10^{31} phages) and the apparent promiscuity of viral–host interactions allow a lot of 'rule breakers' to persist. For example, not only can viral particles survive the physical challenges of extreme environmental shifts (Breitbart *et al*., 2004c), but viruses from one environment (e.g. freshwater Great Lakes) are also readily capable of infecting hosts from another environment (e.g. oceanic *Synechococcus*; (Wilhelm *et al*., 2006). Further, in coliphage T4, the g20 gene encodes a portal protein (Marusich and Mesyanzhinov, 1989) involved in functions quite removed from the direct interaction between

Table 5. Probability that g20 sequence clusters are non-random with respect to the temperature at the site from which they were collected.

Environment	Temperature (°C)	# Sequences	Unifrac *P*-value
Sewage	*N/A*	*6*	*≤ 0.0001*
Cold	*< 4.99*	*20*	*≤ 0.0001*
Cool	*5–14.99*	*57*	*0.2209*
Medium	15–21.99	141	0.2296
Warm	*22–29.99*	*467*	*0.0003*
Hot	*> 30*	*3*	*0.0394*

The Unifrac distance metric (Lozupone *et al*., 2006) was used for the analysis. Temperature values, when not available from the published work, were obtained from the communicating author of the paper in which the g20 sequence was first reported. All but the sequences from moderate temperatures clustered non-randomly ($P < 0.05$) with respect to the habitat type as defined by temperature.

Fig. 2. Evolutionary relationships determined using 554 base pairs of the portal protein gene (g20) from 769 available g20 sequences. Clusters defined by Zhong and colleagues (2002) are identified as culture-based clusters I–III and environmental-sequence-only clusters A–F. New clusters defined since Zhong and colleagues (2002) are indicated with the preface 'new cluster', a number and a brief description. The tree shown is the consensus (majority rules) tree from 11 GARLI iterations inferred using the maximum likelihood criterion (see *Experimental procedures*), with the Aeromonas phage Aeh1 g20 sequence used as an outgroup to root the tree. Three colour rings reflect the habitat type from which the g20 sequence originated. For most of these sequences (GOS sequences), there is ribotype dot-blot and metagenomic information about the microbial community structure at the site, while for non-GOS sequences such information was assumed where reasonable to do so (see Table 3 legend). The inner ring is the microbial community structure information listed as Rusch and colleagues (2007)-defined environmental categories, while the other two rings reflect the temperature and salinity of the original sampling site.

phage and host. In contrast, the distal tail fibre gene is known to be the direct determinant of host range in T-even coliphages (Henning and Hashemolhosseini, 1994; Tetart *et al*., 1998). Thus, g20 sequence patterns might no longer correlate to host range at the fine scales (e.g. cyanobacteria and their phages) where host range 'jumps' could more commonly occur (e.g. by simple tail-fibre-switching *sensu* Tetart *et al*., 1998) that would de-couple host properties from vertically evolved g20 sequence lineages.

Concluding remarks

Taken together, these data reveal that oceanic phage g20 sequence clustering patterns are, at a fine level (e.g. cyanobacteria-cyanophages), largely uncorrelated to host factors. As one zooms out to more generally consider the relationship between g20 sequences from the wild and the habitat characteristics from which they were collected, we find that they are non-randomly distributed, reflecting in some cases a connection between habitat properties, microbial community structure and phage community composition as defined by the g20 gene. We posit that the latter patterns, when evident, reflect host range-limited vertical evolution of g20 sequences, while the former reflects highly specific 'tip-of-the-tree' phage–host interactions that are evolutionarily disconnected from that of the g20 protein product.

Experimental procedures

Phage isolates

Forty-five cyanomyophages were isolated (Table 1) as described previously (Waterbury and Valois, 1993; Wilson *et al*., 1993; Marston and Sallee, 2003; Sullivan *et al*., 2003). S-PM2 and S-WHM1 were provided by W. Wilson and all S-RIM phages were provided by M. Marston. The specificity of cyanomyophage g20 primers was tested using five marine *Pseudoalteromonas* spp. bacteriophages (HER320, HER321, HER322, HER327, HER328; Wichels *et al*., 1998) that were purchased from the Felix d'Herelle Reference Center for Bacterial Viruses (contact H. Ackermann) as well as seven heterotrophic bacteriophages (IH6-ϕ1, IH6-ϕ7, IH11-ϕ2, IH11-ϕ5, CB8-ϕ2, CB8-ϕ6, CB-ϕ8; Zhong *et al*., 2002) kindly provided by F. Chen.

Primer redesign

To obtain g20 PCR amplicons from myophage that would not amplify using published primers, we added degeneracies to both CPS1 and CPS8, and shifted the CPS8 primer based upon genomic sequence data from two *Prochlorococcus* myophage isolates, P-SSM2 and P-SSM4 (Sullivan *et al*., 2005), to design CPS1.1 5′-GTAGWATWTTYTAYATTGAYG TWGG-3′ and CPS8.1 5′-ARTAYTTDCCDAYRWAWGGW TC-3′.

PCR amplification and sequencing

Previous g20 PCR primer sets [non-degenerate CPS4GC/CPS5 (Wilson *et al*., 1999) and degenerate CPS1/CPS8 (Fuller *et al*., 1998; Zhong *et al*., 2002] were designed to amplify ~200 bp and ~592 bp fragments, respectively, of the T4 g20 homologue in myophages.

The PCR reactions for CPS4GC/CPS5 and CPS1/CPS8 were conducted as described previously (Wilson *et al*., 1999; Zhong *et al*., 2002). Briefly, 2 μl of cyanophage lysate was added as DNA template to a PCR reaction mixture (total volume 50 μl) containing the following: 20 pmol each of a forward and reverse primers, 1× PCR buffer (50 mM Tris-HCl, 100 mM NaCl, 1.5 mM $MgCl_2$), 250 μM of each dNTP and 0.75 U of Expand high-fidelity DNA polymerase (Roche, Indianapolis, IN). The PCR amplification was carried out with a PTC-100 DNA Engine Thermocycler (MJ Research, San Francisco, CA). Optimized thermal cycling conditions varied slightly from those reported as follows: CPS4GC/CPS5 required an initial denaturation step of 94°C for 3 min, followed by 35 cycles of denaturation at 94°C for 1 min, annealing at 50°C for 1 min, ramping at 0.3°C s^{-1}, and elongation at 73°C for 1 min with a final elongation step at 73°C for 4 min, whereas both primer sets CPS1/CPS8 and CPS1.1/CPS8.1 required an initial denaturation step of 94°C for 3 min, followed by 35 cycles of denaturation at 94°C for 15 s, annealing at 35°C for 1 min, ramping at 0.3°C s^{-1}, and elongation at 73°C for 1 min with a final elongation step at 73°C for 4 min. Systematic PCR screening using various primer sets was conducted using the same PCR reaction conditions and amplification protocol, but replacing the high-fidelity DNA polymerase with the less-expensive Taq DNA polymerase (Invitrogen, Carlsbad, CA) and only using 20 μl reactions as replicate (range 3–8) PCR reactions were pooled before sequencing to decrease PCR bias (Polz and Cavanaugh, 1998). In all cases, a 5–10 μl aliquot of PCR product was analysed in a 1.5% TAE gel stained with EtBr. The gel image was captured and analysed with an Eagle Eye II gel documentation system (Stratagene, La Jolla, CA). For purification and sequencing, replicate PCR reactions were combined, run out on a 1.5% TAE gel and purified using the QIAGEN QIAquick gel extraction kit (Qiagen, Valencia, CA). The purified PCR products were sequenced directly on both strands using the degenerate PCR primers used to obtain the product (CPS1, CPS8, CPS1.1, CPS8.1) with best results at primer concentrations ~10-fold those suggested by the sequencing facility (40 pmol per reaction). To have greater confidence in negative PCR results, templates that did not produce amplified product were tested against optimized primer sets multiple times (data not shown). To confirm that our correctly sized amplicons from 'positive' PCR reactions were in fact g20 sequences, we sequenced the products. In all cases, the amplicon sequences were from g20 homologues

Where identical g20 sequences were observed in our study, we confirmed that the match was real and not the result of PCR contamination by re-amplifying and sequencing directly from fresh phage isolates (e.g. for P-SSM4, P-RSM3, S-SSM2 and 'Syn' phages Syn2, Syn9, Syn10, Syn26, Syn30, Syn33, Syn1, Syn19), many of which were obtained from stocks kept at a separate institution.

Phylogenetic analysis

For the new sequences presented in Fig. 1 of this study, paired sequence data were aligned using ClustalW (Thompson *et al*., 1997) and corrected manually using the sequence chromatograms. Consensus sequences for each cyanophage isolate were then translated in-frame into amino acids. Published g20 sequences from PCR-amplified environmental clone libraries and phage isolates were screened by building preliminary neighbour-joining trees to select representative sequences that spanned the known g20 diversity and added to this data set. Multiple sequence alignments of translated amino acid consensus sequences were done with ClustalW using the Gonnet protein weight matrix, a gap opening penalty of 15 and gap extension penalty of 0.30 (although changing these penalties did not significantly alter the alignments). Phylogenetic reconstruction was done using PAUP 4.0 (Swofford, 2002) for parsimony and distance trees and Tree-Puzzle 5.0 (Schmidt *et al*., 2002) for maximum likelihood trees. Evolutionary distances for neighbour-joining trees were calculated based on mean character distances, while evolutionary distances for maximum likelihood trees were calculated using the JTT model of substitution assuming a gamma-distributed model of rate heterogeneities with 16 gamma-rate categories empirically estimated from the data. A heuristic search with 10 random addition replicates using the tree-bisection-reconnection branch swapping algorithm was used for parsimony trees. Bootstrap analysis was used to estimate node reproducibility and tree topology for neighbour-joining (1000 replicates) and parsimony (100 replicates) trees, while quartet puzzling (10 000 replicates) indicates support for the maximum likelihood tree. The g20 sequence from coliphage T4 was used as the outgroup taxon for all analyses.

Phylogenetic analyses of 183 amino acids from viral g20 sequence from 79 taxa yielded robust, similar trees using both algorithmic (neighbour-joining) and tree-searching (parsimony and maximum likelihood) methods. The translated g20 sequences contained phylogenetically informative regions (e.g. for parsimony analyses, 41 positions were constant, 25 were parsimony uninformative and 117 were parsimony informative). Differences between the parsimony, distance and maximum likelihood trees were limited to the branching order of the terminal nodes in a given cluster. To evaluate whether g20 sequence diversity correlated to the host-related properties presented in Fig. 1, we empirically defined a 'well supported node' as one where the average support across all three phylogenetic methods was 80% or greater.

GOS g20 identification, filtering and phylogenetic analyses

Using the 549 bp g20 fragment from all available cultured isolates as queries (Table 1), we retrieved 553 sequence reads with similarity (bit score > 100) to this region of the g20 gene from the GOS databases (downloaded from http://camera.calit2.net/), then combined these GOS sequences with available published g20 sequences. The combined sequences were aligned using Clustal X and filtered to remove short, phylogenetically uninformative sequences, as well as sequences with poor quality at the ends. This manual curation left 769 total sequences (512 GOS sequences, details in Table 2) with 554 aligned nucleotide positions. Eleven maximum likelihood trees were generated using GARLI (Zwickl, 2006), starting from a neighbour-joining topology calculated in PAUP v4b10 (Swofford, 2002). Tree searching was terminated after 100 000 generations with no significantly better scoring topology, and a score improvement threshold for termination of 0.05. Topology mutation proportions were 0.1–0.2 nearest neighbour interchange and 0.8–0.9 limited SPR (subtree pruning-regrafting), with the maximum SPR range of 8–10 branches. From the 11 resulting trees, a majority-rule consensus tree (threshold 50% agreement) was generated in PAUP and is presented in Fig. 2.

Statistical analyses to evaluate whether g20 clustering patterns uncovered in the phylogenetic reconstructions were related to the habitat features of the original sample (e.g. microbial community type, temperature and salinity) were carried out using the Unifrac distance metric statistical tools available at http://bmf2.colorado.edu/unifrac/index.psp (Lozupone and Knight, 2005). The database and the tree file used for the analysis are provided in Supplementary Information (Files S1 and 2). Briefly, all g20 sequences were assigned to environmental categories using meta data for each sequence, with some assumptions made as described in Table 3 legend. Missing meta data for published g20 sequences were obtained where possible from the authors of the original work, as indicated in Tables 4 and 5. The patterns of these meta data were evaluated for 'each environment separately' in the context of a single neighbour-joining tree that included branch lengths (File S2) using Unifrac; all statistical results were similar using the *P*-test (also available at the Unifrac site, data not shown).

Nucleotide sequence accession numbers

The nucleotide sequences determined in this study were submitted to GenBank and assigned accession numbers EU715778–15813.

Acknowledgements

This research was supported in part by funding from NSF (CMORE contribution #87), DOE, The Seaver Foundation and the Gordon and Betty Moore Foundation Marine Microbiology Program to S.W.C.; an NIH Bioinformatics Training Grant supported M.B.S.; MIT Undergraduate Research Opportunities Program supported V.Q., J.A.L., G.T., R.F. and J.E.R.; Howard Hughes Medical Institute funded MIT Biology Department Undergraduate Research Opportunities Program supported A.S.D.; NSERC (Canada) Discovery Grant (DG 298394) and a Grant from the Canadian Foundation for Innovation (NOF10394) to J.P.B.; NSF Graduate Fellowship funding supported M.L.C. F. Chen and M. Marston kindly provided phage isolates used in testing of PCR primer sets. M.B.S. thanks C. Lozupone and M. Hamady for interpretive and technical support using Unifrac, as well as F. Chen, M. Marston, U. Dorigo, J. Waterbury and S. Wilhelm for providing unpublished meta-data for published g20 sequences to make the meta-g20 data set analyses as comprehensive as possible. The comments of V. Rich greatly improved the manuscript.

References

Ahlgren, N.A., and Rocap, G. (2006) Culture isolation and culture-independent clone libraries reveal new marine *Synechococcus* ecotypes with distinctive light and nitrogen physiologies. *Appl Environ Microbiol* **72:** 7193–7204.

Bergh, O. (1989) High abundance of viruses found in aquatic environments. *Nature* **340:** 467–468.

Bratbak, G., Heldal, M., Norland, S., and Thingstad, T.F. (1990) Viruses as partners in spring bloom microbial trophodynamics. *Appl Environ Microbiol* **56:** 1400–1405.

Breitbart, M., and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13:** 278–284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99:** 14250–14255.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2004a) Diversity and population structure of a near-shore marine-sediment viral community. *Proc R Soc Lond B Biol Sci* **271:** 565–574.

Breitbart, M., Miyake, J.H., and Rohwer, F. (2004b) Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett* **236:** 249–256.

Breitbart, M., Wegley, L., Leeds, S., Schoenfeld, T., and Rohwer, F. (2004c) Phage community dynamics in hot springs. *Appl Environ Microbiol* **70:** 1633–1640.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311:** 1768–1770.

Coombs, D., and Eiserling, F.A. (1977) Studies on the structure, protein composition and assembly of the neck of bacteriophage T4. *J Mol Biol* **116:** 375–407.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496–503.

Dorigo, U., Jacquet, S., and Humbert, J.-F. (2004) Cyanophage diversity, inferred from g20 gene analyses, in the Largest Natural Lake in France, Lake Bourget. *Appl Environ Microbiol* **70:** 1017–1022.

van Driel, R., and Couture, E. (1978) Assembly of the scaffolding core of bacteriophage T4 proheads. *J Mol Biol* **123:** 713–719.

Frederickson, C.M., Short, S.M., and Suttle, C.A. (2003) The physical environment affects cyanophage communities in British Columbia Inlets. *Microbial Ecology* **46:** 348–357.

Fuller, N.J., Wilson, W.H., Joint, I.R., and Mann, N.H. (1998) Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl Environ Microbiol* **64:** 2051–2060.

Fuller, N.J., Marie, D., Partensky, F., Vaulot, D., Post, A.F., and Scanlan, D.J. (2003) Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* **69:** 2430–2443.

Hambly, E., Tetart, F., Desplats, C., Wilson, W.H., Krisch, H.M., and Mann, N.H. (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc Natl Acad Sci USA* **98:** 11411–11416.

Henning, U., and Hashemolhosseini, S. (1994) Receptor recognition by T-even-type coliphages. In *Molecular Biology of Bacteriophage T4*. Karam, J. (ed.). Washington DC, USA: American Society for Microbiology Press, pp. 291–298.

Hsiao, C.L., and Black, L.W. (1978) Head morphogenesis of bacteriophage T4. III. The role of g20 in DNA packaging. *Virology* **91:** 26–38.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101:** 11013–11018.

Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71:** 8228–8235.

Lozupone, C., Hamady, M., and Knight, R. (2006) UniFrac – an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7:** 371.

Lu, J., Chen, F., and Hodson, R.E. (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Appl Environ Microbiol* **67:** 3285–3290.

Mann, N.H., Clokie, M.R., Millard, A., Cook, A., Wilson, W.H., Wheatley, P.J., *et al.* (2005) The genome of S-PM2, a 'photosynthetic' T4-type bacteriophage that infects marine *Synechococcus*. *J Bacteriol* **187:** 3188–3200.

Marston, M.F., and 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.

Marusich, E.I., and Mesyanzhinov, V.V. (1989) Nucleotide and deduced amino acid sequence of bacteriophage T4 gene 20. *Nucleic Acids Res* **17:** 7514.

Mühling, M., Fuller, N.J., Millard, A., Somerfield, P.J., Marie, D., Wilson, W.H., *et al.* (2005) Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environ Microbiol* **7:** 499–508.

Partensky, F., Hess, W.R., and Vaulot, D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63:** 106–127.

Paul, J.H., Sullivan, M.B., Segall, A.M., and Rohwer, F. (2002) Marine phage genomics. *Comp Biochem Physiol B Biochem Mol Biol* **133:** 463–476.

Polz, M.F., and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64:** 3724–3730.

Proctor, L.M., and Fuhrman, J.A. (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature* **343:** 60–62.

Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68:** 1180–1191.

Rohwer, F., and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184:** 4529–4535.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B.,

Williamson, S., Yooseph, S. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5:** e77.

Sandaa, R.A., and Larsen, A. (2006) Seasonal variations in virus–host populations in Norwegian coastal waters: focusing on the cyanophage community infecting marine *Synechococcus* spp. *Appl Environ Microbiol* **72:** 4610–4618.

Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502–504.

Short, C.M., and Suttle, C.A. (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol* **71:** 480–486.

Stoddard, L.I., Martiny, J.B., and Marston, M.F. (2007) Selection and characterization of cyanophage resistance in marine *Synechococcus* strains. *Appl Environ Microbiol* **73:** 5516–5522.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424:** 1047–1051.

Sullivan, M.B., Coleman, M., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3:** e144.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* **4:** e234.

Suttle, C.A. (2000) Cyanophages and their role in the ecology of cyanobacteria. In *The Ecology of Cyanobacteria: Their Diversity in Tiime and Space*. Whitton, B.A., and Potts, M. (eds). Boston, USA: Kluwer Academic Publishers, pp. 563–589.

Suttle, C.A., and Chan, A.M. (1993) Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar Ecol Prog Ser* **92:** 99–109.

Suttle, C.A., and Chan, A.M. (1994) Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Appl Environ Microbiol* **60:** 3167–3174.

Swofford, D.L. (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, *Version* 4. Sunderland, MA, USA: Sinauer Associates.

Tetart, F., Desplats, C., and Krisch, H.M. (1998) Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: recombination between conserved motifs swaps adhesin specificity. *J Mol Biol* **282:** 543–556.

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

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876–4882.

Wang, K., and Chen, F. (2004) Genetic diversity and population dynamics of cyanophage communities in the Chesapeake Bay. *Aquat Microb Ecol* **34:** 105–116.

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

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

Wichels, A., Biel, S.S., Gelderblom, H.R., Brinkhoff, T., Muyzer, G., and Schutt, C. (1998) Bacteriophage diversity in the North Sea. *Appl Environ Microbiol* **64:** 4128–4133.

Wilhelm, S.W., Carberry, M.J., Eldridge, M.L., Poorvin, L., Saxton, M.A., and Doblin, M.A. (2006) Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of g20 genes. *Appl Environ Microbiol* **72:** 4957–4963.

Wilson, W.H., Fuller, N.J., Joint, I.R., and Mann, N.H. (1999) Analysis of cyanophage diversity and population structure in a south-north transect of the Atlantic Ocean. In *Marine Cyanobacteria*. Sharpy, L., and Larkum, A.W.D. (eds). Monaco: Bulletin de l'Institut oceanographique, pp. 209–216.

Wilson, W.H., Fuller, N.J., Joint, I.R., and Mann, N.H. (2000) Analysis of cyanophage diversity in the marine environment using denaturing gradient gel electrophoresis. In *Microbial Biosystems: New Fronteirs. Proceedings of the 8th International Symposium on Microbial Ecology*. Bell, C.R., Brylinksky, M., and Johnson-Green, P. (eds). Halifax, Nova Scotia, Canada: Atlantic Canada Society for Microbial Ecology, pp. 565–570.

Wilson, W.H., Joint, I.R., Carr, N.G., and Mann, N.H. (1993) Isolation and molecular characterization of five marine cyanophages propogated on *Synechococcus* sp. strain WH 7803. *Appl Environ Microbiol* **59:** 3736–3743.

Zhong, Y., Chen, F., Wilhelm, S.W., Poorvin, L., and Hodson, R.E. (2002) Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. *Appl Environ Microbiol* **68:** 1576–1584.

Zwickl, D.J. (2006) *Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion*. Austin, TX, USA: University of Texas.

Supplementary material

The following supplementary material is available for this article online:

File S1. Salinity categories used in Unifrac analyses.

File S2. Treefree used in Unifrac analyses.

This material is available as part of the online article from http://www.blackwell-synergy.com

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

