

Phylogenomics of T4 cyanophages: lateral gene transfer in the ‘core’ and origins of host genes

J. Cesar Ignacio-Espinoza¹ and
Matthew B. Sullivan^{1,2*}

Departments of ¹Molecular and Cellular Biology, and
²Ecology and Evolutionary Biology, University of
Arizona, Tucson, AZ, USA.

Summary

The last two decades have revealed that phages (viruses that infect bacteria) are abundant and play fundamental roles in the Earth System, with the T4-like myoviruses (herein T4-like phages) emerging as a dominant ‘signal’ in wild populations. Here we examine 27 T4-like phage genomes, with a focus on 17 that infect ocean picocyanobacteria (cyanophages), to evaluate lateral gene transfer (LGT) in this group. First, we establish a reference tree by evaluating concatenated core gene supertrees and whole genome gene content trees. Next, we evaluate what fraction of these ‘core genes’ shared by all 17 cyanophages appear prone to LGT. Most (47 out of 57 core genes) were vertically transferred as inferred from tree tests and genomic synteny. Of those 10 core genes that failed the tree tests, the bulk (8 of 10) remain syntenic in the genomes with only a few (3 of the 10) having identifiable signatures of mobile elements. Notably, only one of these 10 is shared not only by the 17 cyanophages, but also by all 27 T4-like phages (thymidylate synthase); its evolutionary history suggests cyanophages may be the origin of these genes to *Prochlorococcus*. Next, we examined intragenic recombination among the core genes and found that it did occur, even among these core genes, but that the rate was significantly higher between closely related phages, perhaps reducing any detectable LGT signal and leading to taxon cohesion. Finally, among 18 auxiliary metabolic genes (AMGs, a.k.a. ‘host’ genes), we found that half originated from their immediate hosts, in some cases multiple times (e.g. *psbA*, *psbD*, *pstS*), while the remaining have less clear

evolutionary origins ranging from cyanobacteria (4 genes) or microbes (5 genes), with particular diversity among viral *TalC* and *Hsp20* sequences. Together, these findings highlight the patterns and limits of vertical evolution, as well as the ecological and evolutionary roles of LGT in shaping T4-like phage genomes.

Introduction

Bacteriophages (a.k.a. phages, or viruses that infect bacteria) are the most abundant biological entities on earth (Hendrix *et al.*, 1999; Rohwer, 2003), and play critical roles in the earth system by altering bacterial population structure, mortality and evolution (Fuhrman, 1999; Weinbauer and Rassoulzadegan, 2004; Suttle, 2005; 2007; Breitbart *et al.*, 2007). In the oceans, one group, the T4-like phages, are globally distributed and abundant as estimated either from single gene and metagenomic surveys (Fuller *et al.*, 1998; Breitbart *et al.*, 2002; Filee *et al.*, 2005; Short and Suttle, 2005; Angly *et al.*, 2006; DeLong *et al.*, 2006; Yooseph *et al.*, 2007; Williamson *et al.*, 2008) or culture-based methods, particularly for cyanobacteria (Suttle and Chan, 1993; Waterbury and Valois, 1993; Wilson *et al.*, 1993; Lu *et al.*, 2001; Sullivan *et al.*, 2003; 2006; 2008; 2010; Marston and Sallee, 2003; Wilhelm *et al.*, 2006).

Because of this abundance and ecological importance, there are now 27 T4-like phage genomes available: 17 ‘cyanophages’ that were isolated using diverse *Prochlorococcus* and *Synechococcus* host strains (Mann *et al.*, 2005; Sullivan *et al.*, 2005; 2010; Weigele *et al.*, 2007; Millard *et al.*, 2009), and 10 ‘non-cyanophages’ that were isolated using the heterotrophic bacteria *Escherichia coli*, *Vibrio* and *Aeromonas* (Miller *et al.*, 2003a,b; Nolan *et al.*, 2006; Petrov *et al.*, 2006). While T4-like phages, based on morphology, are thought to be a monophyletic group (Calendar, 2006) within the Myoviridae family in the order *Caudovirales* (van Regenmortel *et al.*, 2000), extensive genomic and genetic diversity occurs in the T4-likes (Miller *et al.*, 2003a; Sullivan *et al.*, 2005; 2010; Filee *et al.*, 2006; Nolan *et al.*, 2006; Weigele *et al.*, 2007; Millard *et al.*, 2009). This observation together with frequent mosaicism in phages, particularly siphoviruses (Proux *et al.*, 2002; Hendrix, 2003; Pedulla *et al.*, 2003), has led to the general acceptance of the paradigm that the

Received 14 July, 2011; revised 13 December, 2011; accepted 15 January, 2012. *For correspondence. E-mail mbsulli@email.arizona.edu; Tel. (+1) 520 626 6297; Fax (+1) 520 621 9903.

main forces shaping phage genomes are lateral gene transfer (LGT) and recombination (Hendrix *et al.*, 1999; Filee *et al.*, 2006).

Initial studies, based on DNA hybridization techniques between lambdoid phages (*Siphoviridae* family of phages) led to the modular theory of phage evolution whereby genomes are thought to evolve by the rearrangement of sets of functionally related genes (Botstein, 1980). Mechanistically, this implies non-vertical inheritance of genetic material, which violates the assumptions of phylogeny. In this respect Hendrix and colleagues (1999) documented LGT between phage families that could obscure taxonomic boundaries while other studies (e.g. whole genome sequencing of lambdoid *Streptococcus* phages) revealed conservation in gene content and synteny suggesting that at least some vertically inheritance among these mosaic siphoviruses (Desiere *et al.*, 1999).

Concurrently, environmental phage ecologists began the use of single protein phylogenies to infer relationships among wild phages. For example, T4-like phages were evaluated using the major capsid (gp23), tail sheath (gp18) and tail tube (gp19) proteins, which resulted in recognition of four groups of T4-like phages: (i) T4 (ii) Pseudo-T4 (iii) Schizo-T4, and (iv) Exo-T4 (Tétart *et al.*, 2001; Desplats and Krisch, 2003). While based on relatively limited data from isolates and environmental sequences it set up initial hypotheses about T4-like phage phylogenetic relationships. Further, within this framework, Nolan, Comeau and Filee examined non-cyano T4-like phage genomes and showed that they shared 24–82 genes (core genes, Filee *et al.*, 2006; Nolan *et al.*, 2006; Comeau *et al.*, 2007), which appeared predominantly vertically transferred (Filee *et al.*, 2006; Comeau *et al.*, 2007). We recently compared 26 T4-like phage genomes (Sullivan *et al.*, 2010) and found 38 genes shared across cyano and non-cyano T4 phages (this 'core' is now well sampled) complemented by a set of genes that are required for host- or niche-specific interactions (e.g. 25 genes shared only among the 16 cyanophage genomes, a.k.a. the 'cyano T4 core'). Beyond these hierarchical core gene sets, cyanophage genomes also encode auxiliary metabolic genes (AMGs, *sensu* Breitbart *et al.*, 2007), 'host' genes normally involved in atypical phage functions such as photosynthesis (e.g. *psbA*, *psbD*), phosphate stress (e.g. *pstS*, *phoA*, *phoH*) and carbon metabolism (e.g. *gnd*, *zwf*, *CP12*, *talC*; Mann *et al.*, 2003; 2005; Sullivan *et al.*, 2005; 2010; Weigle *et al.*, 2007; Millard *et al.*, 2009; Thompson *et al.*, 2011). These AMGs can be core (e.g. *psbA*, *cobS*, *phoH*, *hsp20* and *mazG*), but most are sporadically distributed and likely prone to LGT to provide ecological advantage in adapting to novel niches and hosts.

Here we evaluate the degree of vertical inheritance in T4-like phage genomes with a focus on the Exo-T4

marine cyanophages. To this end, we document the fraction of core gene single tree topologies that explain the evolutionary history of all core genes (e.g. concatenated core gene supertree), quantify intragenic recombination among T4 phage core genes, and explore the source of AMGs in cyanophage genomes.

Results and discussion

Towards a T4 phylogeny as a 'reference' topology for quantifying LGT

To broadly assess the evolutionary relationships among 27 T4-like phage genomes (Table 1), we first evaluated the variability in topologies of core protein supertrees and whole proteome trees to establish a 'reference' tree for subsequent analyses.

The T4 core tree. Thirty-five 'T4 core' proteins were shared among all 27 T4-like phage genomes (Sullivan *et al.*, 2010). The concatenated protein T4 core supertree (15 224 amino acids) was robust and well supported (Fig. 1A) with a topology that generally agrees with published single protein trees (gp23, major capsid protein; gp18, tail sheath protein; gp19, tail tube protein) where the groups T4, Pseudo-T4, Schizo-T4 and Exo-T4 were originally defined (Tétart *et al.*, 2001; Desplats and Krisch, 2003). While earlier works observed a monophyletic Schizo-T4, we found this group to be paraphyletic (Fig. 1A); however, with only two Schizo-T4 taxa available, this remains unresolved. Further, we observe that cyanophages form two deep-branching clades (labelled 'Clusters A and B' in Exo-T4 in Fig. 1A), not previously documented due to the lack of available cyanophage taxa.

The cyano T4 core tree. With 17 cyano T4 phage genomes, we expanded our core dataset to include all predicted proteins that are exclusive and universal among these genomes (the 'cyano T4 core'). This expanded dataset (35 T4 core proteins + 22 cyano T4 core proteins) represents 17–28% of the predicted proteome of any cyanophage genome, as compared with 10–16% for the T4 core set. The resulting cyano T4 core supertree (57 concatenated proteins, 20 638 amino acids) supports the two deep-branching cyanophage clades (Fig. 1B) and is broadly concordant with the T4 core supertree (compare Fig. 1A with Fig. 1B) in that only the topology of cluster A had small changes that are probably result of the insufficient sampling of the divergent phages within this group.

Whole genomes. Whole genome gene content has been used to infer evolutionary relationships between microbes (Snel *et al.*, 1999), phages (Rohwer and Edwards, 2002)

Table 1. T4-like phages used in this study.

Published name	Original host	Genome size (kb)	# ORFs	%G + C	Genbank accession #	Genome publication
Cyanophages						
P-SSM2	<i>Prochlorococcus</i> NATL1A	252.4	334	35.5%	AY939844	Sullivan <i>et al.</i> (2005)
P-SSM4	<i>Prochlorococcus</i> NATL2A	178.2	221	36.7%	AY940168	Sullivan <i>et al.</i> (2005)
P-HM1	<i>Prochlorococcus</i> MED4	181	241	38.0%	GU071101	Sullivan <i>et al.</i> (2010)
P-HM2	<i>Prochlorococcus</i> MED4	183.8	242	38.0%	GU075905	Sullivan <i>et al.</i> (2010)
P-RSM4	<i>Prochlorococcus</i> MIT9303	176.4	239	38.0%	GU071099	Sullivan <i>et al.</i> (2010)
P-SSM7	<i>Prochlorococcus</i> NATL1A	182.2	237	37.0%	GU071103	Sullivan <i>et al.</i> (2010)
S-PM2	<i>Synechococcus</i> WH7803	196.3	244	37.8%	AJ630128	Mann <i>et al.</i> (2005)
Syn9	<i>Synechococcus</i> WH8109	177.3	228	40.50%	DQ149023	Weigele <i>et al.</i> (2007)
Syn19	<i>Synechococcus</i> WH8109	175.2	215	41.0%	GU071106	Sullivan <i>et al.</i> (2010)
Syn33	<i>Synechococcus</i> WH7803	174.4	227	40.0%	GU071108	Sullivan <i>et al.</i> (2010)
Syn1	<i>Synechococcus</i> WH8101	191.2	234	41.0%	GU071105	Sullivan <i>et al.</i> (2010)
S-ShM2	<i>Synechococcus</i> WH8102	179.6	230	41.0%	GU071096	Sullivan <i>et al.</i> (2010)
S-SM2	<i>Synechococcus</i> WH8017	190.8	267	40.0%	GU071095	Sullivan <i>et al.</i> (2010)
S-SSM7	<i>Synechococcus</i> WH8109	232.9	319	39.0%	GU071098	Sullivan <i>et al.</i> (2010)
S-SSM5	<i>Synechococcus</i> WH8102	176.2	225	40.0%	GU071097	Sullivan <i>et al.</i> (2010)
S-SM1	<i>Synechococcus</i> WH6501	178.5	234	41.0%	GU071094	Sullivan <i>et al.</i> (2010)
S-RSM4	<i>Synechococcus</i> WH7803	194.5	238	41.0%	FM207411	Millard <i>et al.</i> (2009)
Non-cyanophages						
T4	<i>E. coli</i> B	168.9	278	35.3%	AG158101	Miller <i>et al.</i> (2003a)
RB32	<i>E. coli</i>	165.9	270	35.3%	DQ904452	http://phage.bioc.tulane.edu/
RB43	<i>E. coli</i> B	180.5	292	43.2%	AY967407	Nolan <i>et al.</i> (2006)
RB49	<i>E. coli</i> CAJ70	164	274	40.4%	AY343333	Nolan <i>et al.</i> (2006)
RB69	<i>E. coli</i> CAJ70	167.6	273	37.7%	AY303349	Nolan <i>et al.</i> (2006)
KVP40	<i>Vibrio parahaemolyticus</i>	244.8	381	42.6%	AY283928	Miller <i>et al.</i> (2003b)
44RR	<i>Aeromonas salmonicida</i> 170-68	173.6	252	43.9%	AY357531	Nolan <i>et al.</i> (2006)
Aeh1	<i>Aeromonas hydrophila</i>	233.2	352	42.8%	AY266303	Nolan <i>et al.</i> (2006)
PHG25	<i>Aeromonas salmonicida</i> 170-68	161.5	242	41.0%	DQ529280	Petrov <i>et al.</i> (2006)
PHG31	<i>Aeromonas salmonicida</i> 95-68	172.9	247	43.9%	AY962392	Petrov <i>et al.</i> (2006)

and even T4-like cyanophage (see below; Millard *et al.*, 2009). Thus we constructed a whole genome gene content phylogeny from the 27 T4-like phage genomes using distances calculated from Simpson Similarity Index (*sensu* Snel *et al.*, 1999; Kettler *et al.*, 2007). In this representation (Fig. 1C) the length of the tips leading to individual genomes is proportional to the number of unique genes in each genome, which in this dataset ranges from 5% for P-HM1 (anomalously low as it shares 83% of its proteins with the genome of co-isolated phage P-HM2, Sullivan *et al.*, 2010) to 50% for the most divergent phage among the 27 genomes (Aeh1). This approach (Fig. 1C) recovers the T4, Pseudo-T4, Schizo-T4 and Exo-T4 groups previously described (Tétart *et al.*, 2001; Desplats and Krisch, 2003), highlighting the relationship between genome content and genetic distance – an observation previously documented in T4 phages using shorter (10 kb) genomic regions (Hambly *et al.*, 2001). However, genome content trees document cyanophage Cluster A as polyphyletic due to the breakdown of the S-SM2/P-SSM2/S-SSM7 cluster. Notably, those phages represent the most divergent cyanophage genomes (Sullivan *et al.*, 2010). As with the cyano T4 core supertree, it is likely that further taxon sampling in this part of the tree should help resolve this polytomy.

Whole genome corollaries. Given this relative consistency across these trees, we first wondered whether any environmental factors correlated to our reference tree topology (Fig. 1B, Table S1). While little environmental data were directly available from the 'source waters' used to isolate these T4-like phages, we used data interpolated from MEGX (Kottmann *et al.*, 2010) using the latitude, longitude, date and depth for each water sample. Among nine variables tested using Unifrac (Lozupone and Knight, 2005) only host range, quantified as phylogenetic diversity (*sensu* Faith, 1992), was a factor that even modestly (P -value = 0.1) explained the topology branching patterns (note: host range data are not available for 3 of 17 cyanophage isolates; Table S1). While data are limiting, we speculate that host range structuring supertree topologies merely reflects coevolution (Kitchen *et al.*, 2011), where host niche specialization is accompanied by diversification of their viruses.

Further exploring whole genome content, we observed that shared protein identities plotted against the fraction of the genome that was shared pairwise across the 27 T4-like phage genomes dataset revealed a strong correlation (linear regression, $r^2 = 0.95$, Fig. S1). This and the genome content phylogeny suggest that closely related phages share more genome content, and further that their

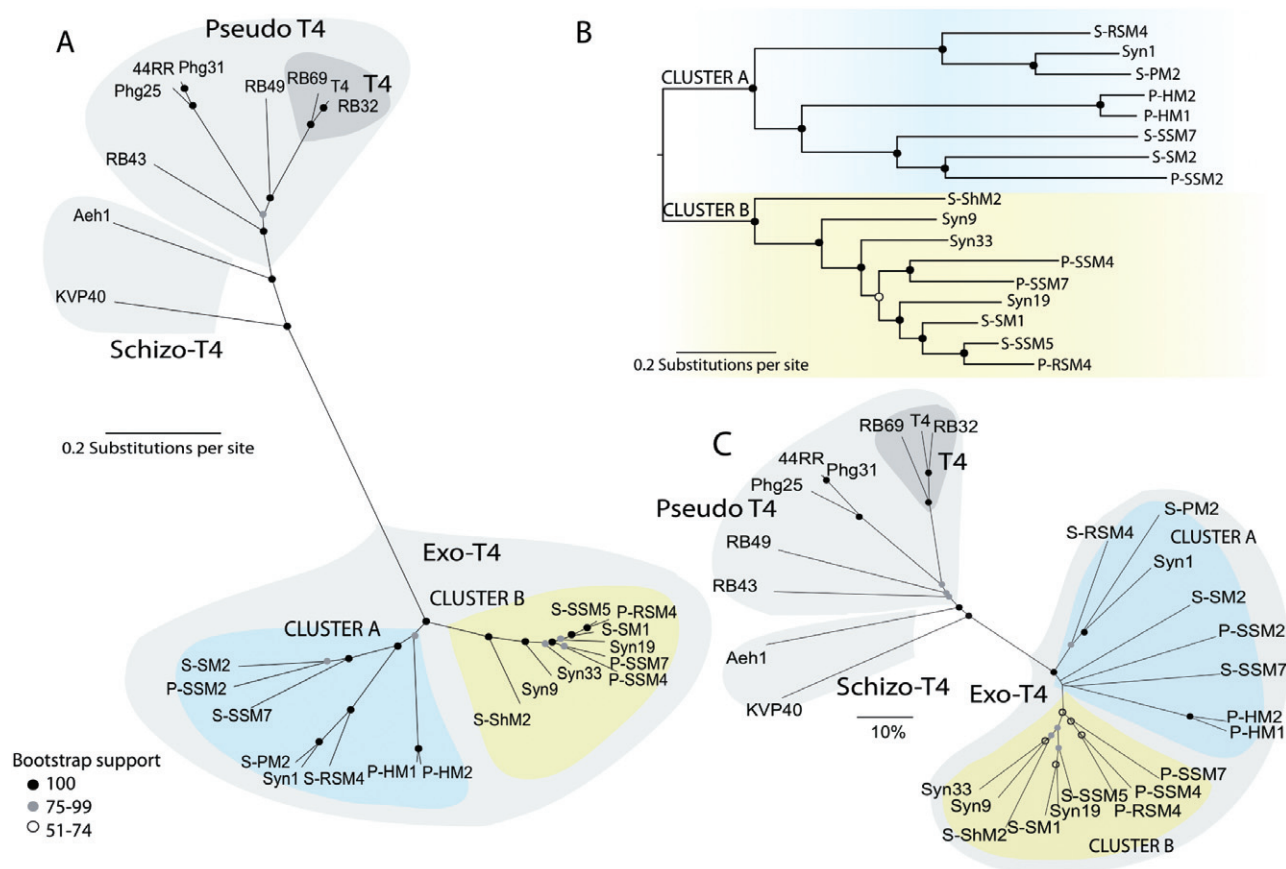


Fig. 1. Phylogenies representing the evolutionary relationships among the T4-like myoviruses. Reconstruction details are in *Experimental procedures*.

A. T4 core phylogeny (35 concatenated proteins totalling 15 224 AA), whereby these genes from the 17 cyanophage genomes delineate two deep-branching clusters (A and B) within the Exo-T4 lineage.
 B. Cyano T4 core phylogeny (57 concatenated proteins totalling 20 638 AA).
 C. Genome content phylogeny calculated using Simpson similarity index. In panels (A) and (C) previously defined groups (Tétart *et al.*, 2001; Desplats and Krisch, 2003) are noted. Bootstrap support (100 replicates) is indicated as grayed circles at the nodes of the tree.

genomes are predominantly vertically inherited. In microbes, LGT (Lerat *et al.*, 2003; Coscollá *et al.*, 2011) and recombination (Didelot *et al.*, 2010) are more common between closely related organisms, which is thought to reduce the likelihood of such events obscuring 'species' boundaries (Konstantinidis *et al.*, 2006; Didelot *et al.*, 2010). In the T4 phages, the relative consistency across single gene and supertree phylogenies suggests that even though these phages are prone to LGT and homologous and non-homologous recombination (Mosig, 1998; Mosig *et al.*, 2001), this may not prevent us from inferring evolutionary relationships.

Do the single gene trees agree with each other and with the reference supertrees?

We next wondered, how many single core gene tree topologies are congruent with each other and the 'reference' T4 core supertree topologies as an indicator of LGT.

To test this, we used Shimodaira-Hasegawa tests (SH, see *Experimental procedures*, where the null hypothesis states that the likelihoods of both topologies are not significantly different of each other) to assess whether each of the 57 core gene alignments could predict a reference tree or any of the statistically indistinguishable top-scoring single gene topologies (1538 trees). Broadly, we observed two scenarios (Fig. 2A): either the genes (i) shared an evolutionary history with each other (47 genes that were not able to reject the null hypothesis, i.e. same topology) or (ii) have more discretely shared or not shared evolutionary histories (10 genes that rejected equality in topologies with a *P*-value < 0.05). These SH results were robust against a variety of potential artefacts as they neither correlated to gene size (see histogram plot at right in Fig. 2A), nor were sensitive to alignment algorithms (Fig. S2). Given the paradigm in phage genomics of rampant LGT (Hendrix *et al.*, 1999; Filee *et al.*, 2006), we were surprised that most (47 of 57) single gene tree

topologies supported the reference tree topology – i.e. had a shared evolutionary history. Comparative genomic analysis revealed that 45 of those 47 core genes with shared evolutionary history were also syntenic in the phage genomes (the 2 exceptions = T4-GCs 043 and 250, Fig. S3), further supporting a vertical mode of evolution. Among their host genomes, we know that < 9% of core genes from *Prochlorococcus* (1250 genes) or *Synechococcus* (1570 genes) appear prone to LGT (Kettler *et al.*, 2007; Dufresne *et al.*, 2008). Thus T4-like cyanophage genomes demonstrate more LGT (10 out of 57 core genes) than their hosts. In spite of this, such a majority signal of shared evolutionary history in the T4-like cyanophage core genome suggests that future phage genomics work may benefit from quantitative evaluation of LGT where data permit to know the extent of LGT in other phage systems.

We next examined the 10 genes that failed the SH tree tests. Surprisingly, eight of them were located in syntenic genome positions, while only two were not (red vs dark lines in Fig. S3). Among these eight syntenic genes, two cases (T4-GCs 170 and 190) are notable as they remain syntenic (between either RegA-gp43 2–7 kb range or gp41-gp61 12–15 kb range, respectively), but are also associated with hypothesized mobile cassettes. The two genes that were non-syntenic (T4-GC71 and T4-GC49) are present in hypervariable regions (Sullivan *et al.*, 2005; 2010; Millard *et al.*, 2009), and at least for T4-GC49 may have been mobilized in three genomes where it sits near a hypothesized homing endonuclease. For the remaining six genes that are also syntenic, we found no clear scenarios for why these core genes might have evolutionary histories that conflicted with the reference tree. This is in spite of the fact that our analyses were extensive including examination of genomic location and proximity to known or suspected mobile elements, evaluation of transcriptionally autonomous ‘moron’ units (*sensu* Hendrix *et al.*, 1999), and even including the recently described ‘PeSL’ elements thought to drive LGT among T4-like phages (Table S2, Arbiol *et al.*, 2010). This latter analytical result, although surprising to us initially, likely stems from the fact that the 17 cyanophage genome sequences represent distantly related viruses (except for P-HM1 and P-HM2), which would make PeSL detection challenging.

Notably 9 of the 10 genes that did not pass the SH test were cyano T4 core genes shared only among the 17 cyanophage genomes, with only the 10th (thymidylate synthase) being a core gene shared by all 27 T4-like phage genomes. Indeed this gene proved interesting. Thymidylate synthase catalyses the synthesis of thymidylate from uridylate using either ThyA (the canonical form) or ThyX (the alternative form; Murzin, 2002; Myllykallio *et al.*, 2002) with the latter reducing the number of reactions necessary in the pathway (Fig. S4). Among marine

picocyanobacteria, *Synechococcus* have the ThyA pathway, while *Prochlorococcus* have the simpler ThyX pathway. Curiously, all 17 cyano T4-like phages encode the latter ThyX pathway regardless of whether they were isolated using, or predominantly, infect *Prochlorococcus* or *Synechococcus* cultures. Phylogenetic reconstruction revealed that *Prochlorococcus* sequences formed two clades – one nested within viral sequences (Group B in Fig. 2B), and a second that was basal to these cyanophage and Group B sequences (Group A in Fig. 2B). Further, examination of the host genomes revealed that thymidylate synthase is located in two different locations (Fig. 2C) that correspond to the two clades observed within the phylogeny. Together these data suggest that phages mediated LGT of thymidylate synthase within *Prochlorococcus*, as previously documented for high-light inducible proteins (Lindell *et al.*, 2004), and may explain why the cyanophage copies of this T4 core gene caused the failure of the SH tree tests.

Patterns of homologous intragenic recombination

Sixteen of 17 T4-like cyanophages contain the recombination machinery (*uvsX*, *uvsY*, *uvsW* genes) required for homologous recombination (S-PM2 is missing *uvsX* and *UvsY*, Mann *et al.*, 2005; Sullivan *et al.*, 2010), which highlights the importance of this process in T4-like phage evolution (Desplats and Krisch, 2003; Marston and Amrich, 2009). To further investigate the extent of homologous recombination in reshaping these 27 T4-like phage genomes, we examined intragenic recombination among their shared proteins (see *Experimental procedures*).

First, recombination events appear separately bounded for the two major groups, cyanophages and non-cyanophages (Fig. 3); the frequencies *within* these groups are statistically higher than those *between* groups (Chi-squared, $P < 0.001$). This observation implies the logical mutual exclusion of niches where these two groups occur. Second, the barrier between cyanophage clusters A and B is less than that between cyanophages and non-cyanophages: the detected frequencies *within* cluster A and *between* clusters A and B are not statistically different, while the number of recombination events *within* cluster B is statistically increased relative to those observed in cluster A or between A and B (Chi-squared, $P < 0.001$). Finally, we constructed a dendrogram (Fig. 3) using the number of detected recombination events as a similarity metric (see *Experimental procedures*), such that phages that recombine more often will cluster closer than those that do not. This method is undoubtedly database-limited in that the ‘recombinant’ phage sequences could have been acquired from a third, unsampled phage, but does provide a first-level understanding of intragenic recombination-driven gene flow among the sampled

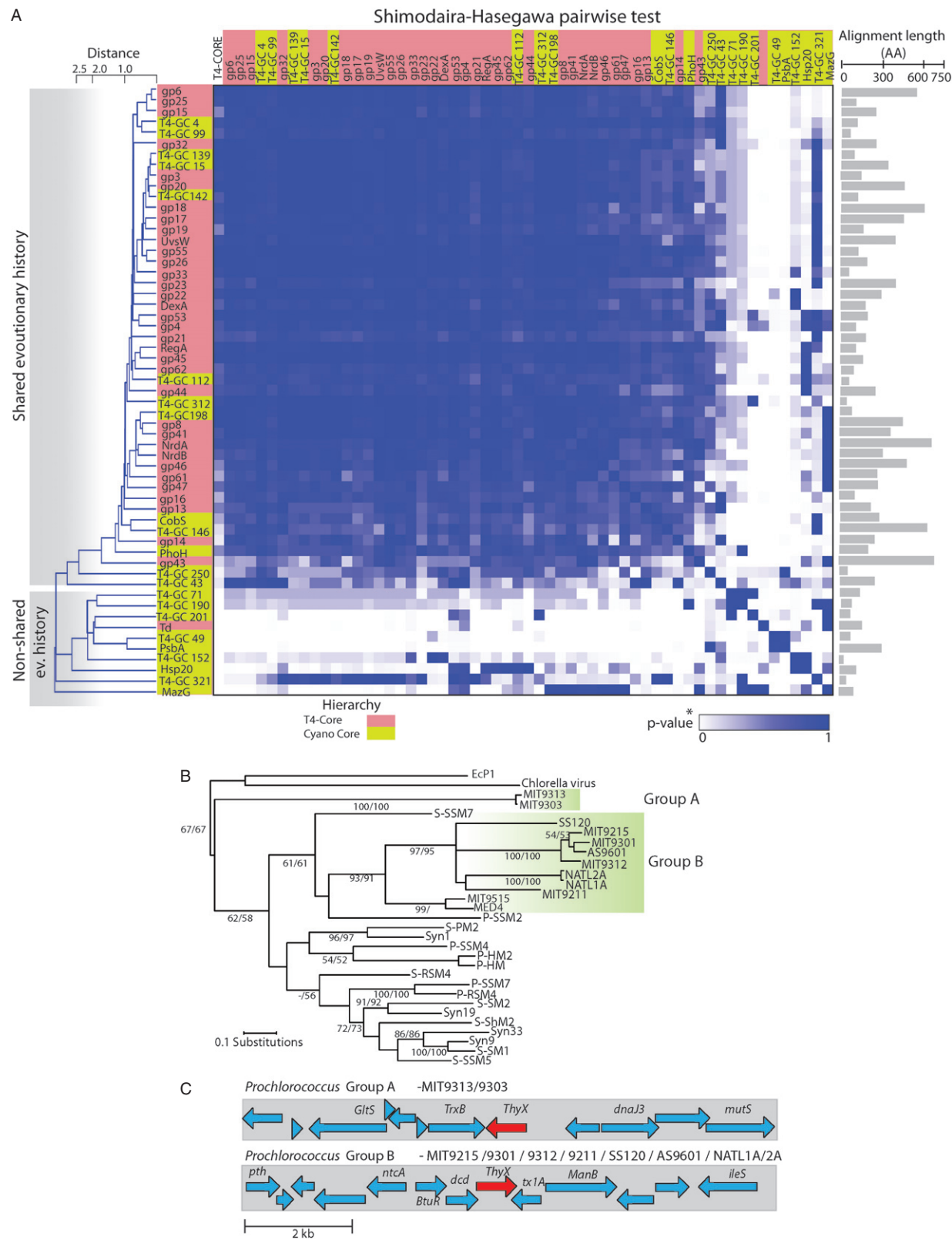


Fig. 2. Using phylogenetic reconstructions, tree tests and comparative genomics to evaluate lateral gene transfer in the T4 core genome.

A. All pairwise comparisons of single core gene trees to each other and to reference trees were evaluated using SH tests to detect incongruencies. Only significant *P*-values (> 0.05) are shown, which indicates that the compared genes had a shared topology, and suggests similar evolutionary histories. The SH test results are flanked by features from left to right: the dendrogram on the far left uses SH test probabilities to cluster the genes with respect to each other using Euclidean distances.

B. Phylogram of thymidylate synthase, the only T4 core gene to fail the SH tree test, was inferred from LogDet transformed distances with branch lengths and bootstraps calculated using a maximum likelihood framework, numbers at the nodes represent bootstrap support under NJ and ML. Cyanobacterial sequences are highlighted in boxes.

C. Genomic localization of thymidylate synthase in the two *Prochlorococcus* clades delineated in the tree from Fig. 2B.

genomes. Broadly, the congruence of this dendrogram to the phylogenetic 'reference trees' (compare Fig. 3 with Fig. 1A and 1B) highlights that such intragenic recombination events tend to occur between closely related phages. Those results are not completely surprising as it has been observed that sequence similarity is a prerequisite to recombination (Majewski, 2001) and it is well known that these rates decrease as sequence similarity does (Shen and Huang, 1986). Thus although homologous recombination does not appear to work as a source of entirely novel DNA, it is a relevant mechanism of diversification. Along these lines, previous work suggested that homologous recombination among cyanophage isolates generates microdiversity, but without blurring their defined strains (Marston and Amrich, 2009). Together these findings suggest that, as in prokaryotes (Didelot *et al.*, 2010), homologous recombination does occur but that it may not preclude our ability to define 'species' as it occurs more often among closely related strains (Konstantinidis *et al.*,

2006). Undoubtedly such LGT and non-homologous recombination could blur species boundaries where it is being transferred from distant groups as observed in microbial genomics (e.g. Lerat *et al.*, 2003; Lawrence and Retchless, 2009).

The origin of AMGs

Auxiliary metabolic genes (*sensu* Breitbart *et al.*, 2007) represent genomic adaptations to different hosts and environments, and are commonly distributed among the 17 sequenced T4-like cyanophage genomes (Sullivan *et al.*, 2010). Perhaps the best known AMGs are photosynthesis related genes, where multiple acquisitions have been documented for the core photosynthesis genes (*psbA*, *psbD*, Millard *et al.*, 2004; Sullivan *et al.*, 2006), the expansion of a gene family is driven by cyanophages (high-light inducible proteins, Lindell *et al.*, 2004), and even sporadically distributed, highly divergent phage

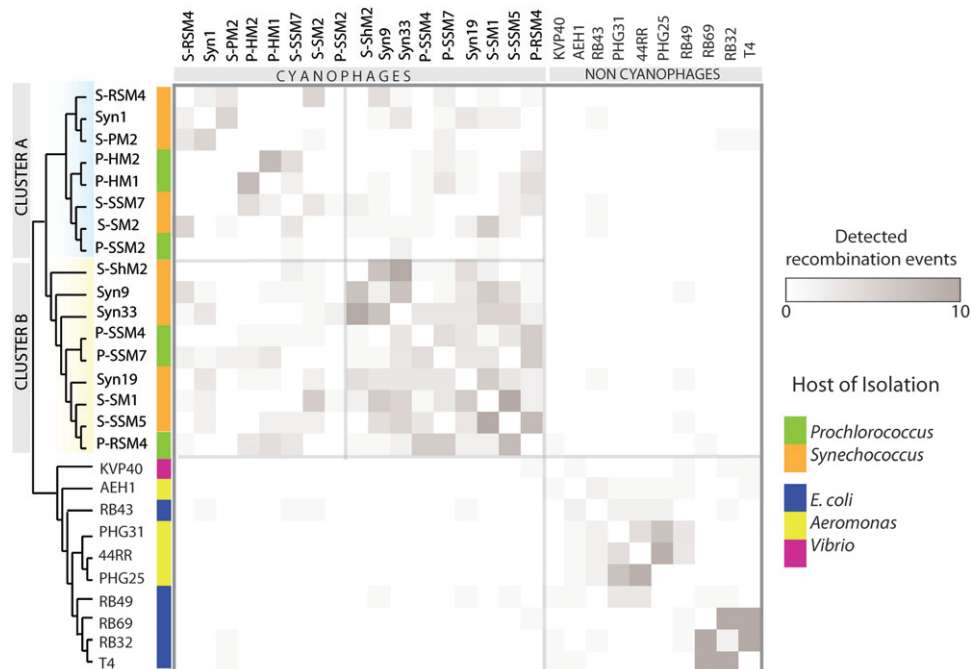


Fig. 3. Patterns of intragenic recombination inferred using RDP. Heat map summarizing the number of intragenic recombination events detected for each pairwise comparison normalized to the total number of protein clusters shared by the two phages. The genomes are ordered using the presented cladogram (on left) that was calculated using the number of recombination events that occurred between any two phages. Cluster A and B refer to the clusters delineated in the cyano T4 core gene tree presented in Fig. 1B.

Table 2. AMGs and their evolutionary origin.

Gene	Protein	# Viral sequences	Evolutionary origin ^a	Average mean distance ^b	Previous assessments
<i>TalC</i>	Transaldolase	16	Divergent from any microbial group	2.30	–
<i>CP12</i>	Carbon metabolic regulator	16	Cyanobacteria Host ^c	1.54	–
<i>Gnd</i>	6-phosphogluconate dehydrogenase	8	Divergent from any microbial group	1.21	Millard <i>et al.</i> (2009)
<i>Zwf</i>	Glucose 6-phosphate dehydrogenase	6	Cyanobacteria Host	1.23	Millard <i>et al.</i> (2009)
<i>PetF</i>	Ferredoxin	5	Cyanobacteria Not host	1.12	–
<i>PetE</i>	Plastocyanin	11	Cyanobacteria Not host	1.01	Lindell <i>et al.</i> (2004)
<i>SpeD</i>	S-adenosylmethionine decarboxylase	7	Cyanobacteria Host	2.66	–
<i>CpeT</i>	CpeT-like protein	13	Cyanobacteria Host	3.24	–
<i>PebS</i>	Phycocerythrobilin biosynthesis	4	Cyanobacteria Not host	1.47	–
<i>PcyA</i>	Phycobillin biosynthesis	3	Cyanobacteria Host	1.50	–
<i>PstS</i>	ABC-type phosphate transport system	9	Cyanobacteria Host	1.72	–
<i>PhoH</i>	P-starvation inducible protein	17	Divergent from any microbial group	1.10	–
<i>Hsp20</i>	Homologue to heat shock protein	17	Divergent from any microbial group	2.50	–
<i>CobS</i>	Cobalamin biosynthesis protein	17	Divergent from any microbial group	1.00	–
<i>PsbD</i>	Photosystem II D2 protein	13	Cyanobacteria Host	0.63	Sullivan <i>et al.</i> (2006)
<i>PsbA</i>	Photosystem II D1 protein	17	Cyanobacteria Host	1.13	Sullivan <i>et al.</i> (2006)
<i>PTOX</i>	Plastoquinol terminal oxidase	10	Cyanobacteria Host	1.50	Millard <i>et al.</i> (2009)
<i>MazG</i>	Pyrophosphatase	17	Cyanobacteria Not host	2.11	Bryan <i>et al.</i> (2008); Sullivan <i>et al.</i> (2010)

a. Trees in File S1.

b. In substitutions per site. As calculated in Webb (2000).

c. Host refers to marine picocyanobacteria of the genera *Prochlorococcus* and *Synechococcus*; Not host refers to other cyanobacteria.

photosynthesis genes appear functional (*pebS*, Dammeyer *et al.*, 2008). It is thought that these AMGs enable viruses to develop capabilities beyond DNA replication, packaging and virion assembly – most likely to modulate bottlenecks in microbial metabolism that limit their production during infection. For example, cyanophage photosynthesis genes are expressed during infection (Clokier *et al.*, 2006; Lindell *et al.*, 2007; Dammeyer *et al.*, 2008), and they are predicted to improve phage fitness (Bragg and Chisholm, 2008; Hellweger, 2009). However, it is likely that LGT plays an important role in the presence/absence of AMGs in cyanophage genomes as they (i) tend to be non-syntenic and cluster within hypervariable regions of cyanophage genomes (Millard *et al.*, 2009; Sullivan *et al.*, 2010), and (ii) their presence often (e.g. *cobS*, *pstS*, *phoH*) correlates with environmental factors in metagenomes (Williamson *et al.*, 2008). Thus we chose to explore the origin and evolutionary history of cyanophage AMGs, with a focus on those that are understudied, as many have been explored elsewhere (MazG by Bryan *et al.*, 2008 and Sullivan *et al.*, 2006; 2010; PsbA and PsbD by ; PTOX, Gnd and Zwf by Millard *et al.*, 2009; PetE and HliP by Lindell *et al.*, 2004; PebS by Dammeyer *et al.*, 2008).

We examined 18 cyanophage AMGs (Table 2, see *Experimental procedures*), and found that they clustered in one of three ways: (i) with their cyanobacterial hosts (9 proteins; CP12, Zwf, SpeD, CpeT, PcyA, PstS, PsbD, PsbA and PTOX) (ii) with cyanobacteria but not necessarily with their *Prochlorococcus* or *Synechococcus* hosts (4 proteins; PetF, PetE, PebS and MazG), or (iii) were too

divergent to assign a microbial origin (5 proteins; TalC, PhoH, Hsp20, Gnd and CobS). While the implications of their prevalence in the viral genomes have been discussed elsewhere (Williamson *et al.*, 2008; Sullivan *et al.*, 2010; Clokier *et al.*, 2011), their diverse origin is remarkable as it suggests myriad LGT events and highlights the ecological importance and plasticity of the niche-defining genes. Where previously examined (see references above), our phylogenetic conclusions are congruent (Table 2). Among those remaining AMGs, we highlight three stories of AMG evolution.

First, we were struck by the finding that proteins involved in similar pathways have different evolutionary origins (Table 2 and File S1, but also see Thompson *et al.*, 2011). For instance, there are four enzymes involved in the pentose phosphate pathway (PPP) found in cyanophages. Transaldolase (TalC), which plays a role in the use of stored carbon (Weigle *et al.*, 2007; Sullivan *et al.*, 2010), was obtained by viruses once but was passed among viruses including both podo- and myoviruses (see monophyletic viral sequences clade in TalC tree, File S1); the origin of the gene is less clear as viral TalC has significantly diverged from any potential 'source' microbial taxon. In contrast, the regulatory protein CP12, which promotes the switch to the PPP to move carbon away from the Calvin cycle (Tamoi *et al.*, 2005; Zinser *et al.*, 2009), occurs only among the T4-like myoviruses and was clearly obtained from within *Prochlorococcus* (see CP12 tree, File S1). Further, two other cyanophage-encoded PPP proteins, Zwf and Gnd, also appear restricted to the T4-like myoviruses, but to have been obtained from their

hosts (Zwf) or to be divergent so as to be of unknown origin (Gnd; see Zwf and Gnd trees in File S1 and see Millard *et al.*, 2009).

In contrast to the above AMGs that were acquired in a single evolutionary event, some AMGs have been acquired multiple times. For example, PstS, which encodes a periplasmic phosphate binding protein and is prevalent in phages isolated from low-nutrient waters (Sullivan *et al.*, 2010), originates from the marine picocyanobacteria, but appears to be predominantly transferred from *Prochlorococcus* hosts to *Prochlorococcus* phages and *Synechococcus* hosts to *Synechococcus* phages – i.e. the viruses have obtained the gene twice (see PstS tree, File S1). While this contrasts previous findings where PstS data from phages were limited to two available sequences (Scanlan *et al.*, 2009), it is not unprecedented among cyanophage-encoded AMGs. Similar multiple acquisitions have been documented for PTOX (Millard *et al.*, 2009), PsbA and PsbD (Sullivan *et al.*, 2006) and high-light inducible proteins (Lindell *et al.*, 2004). These host-related LGT events (at least PsbD, PsbA, PstS and PTOX) together with the geography-linked distribution of AMGs (Dinsdale *et al.*, 2008; Williamson *et al.*, 2008; Sullivan *et al.*, 2010) highlight the strong ecosystem/niche specific selection towards the acquisition of new AMGs not unlike that seen in their cyanobacterial host cells (Coleman and Chisholm, 2010).

Finally, beyond the origin of AMGs, we note that viral AMGs have different levels of diversity (Table 2), indicating divergent evolutionary rates and presumed selective pressures. Many of these genes (e.g. PetE, PetF, PhoH, CobS, PsbA, PsbD) appear highly conserved with mean pairwise distances of ~ 1 (Webb, 2000). In contrast, some AMGs are much more diverse (e.g. TalC, Hsp20, SpeD, CpeT) with as much as 2.3, 2.5, 2.7 and 3.2 substitutions per site respectively. Such low mean pairwise distances are suggestive of either strong purifying selection or a rapid exchange of this gene between phages; we favour the former as these proteins seem to be central in photosynthesis and phosphate metabolism.

Together, these stories paint a relatively wholistic picture of evolution-in-action for AMGs whereby they have been obtained from varying sources and once in viruses their sequences have diverged under varying selective pressure.

Conclusions

The paradigm in phage genomics is one of rampant LGT. However, here we show that most core T4-like phage genes are vertically descendent, but appear more prone to LGT than microbial core genes. These observations warrant further study in integrating genomics into phage systematics (*sensu* Rohwer and Edwards, 2002), and

hold promise for interpreting fragmented metagenomic data from wild viral communities. We posit that just as single cell microbial genomics is likely to transform our understanding of genome evolution in microbes, so too will examinations of genome level variation from wild viral populations further our understanding of the processes driving phage genome evolution.

Experimental procedures

Sequences and GenBank files

Twenty-seven T4-like phage genomes were downloaded from GenBank with annotations updated as of 1 December 2010 (Table 1). The T4-GC (T4-Gene Cluster) protein clusters (*sensu* Sullivan *et al.*, 2010) were used to refine the annotation from the 27th genome that was not available for the previous study (S-RSM4; Millard *et al.*, 2009), where 3 cyano T4 core proteins were added to the GenBank annotation at the following positions: T4-GC43 at 150806–150988; T4-GC112 at 110411–110626 and T4-GC4 at 193925–194266.

Intragenic recombination analysis and removal of proteins from phylogenetic study

Members of the T4-GC protein clusters were examined for signatures of recombination as done previously (Sullivan *et al.*, 2006). Briefly, GENECONV (Padidam *et al.*, 1999), MAXCHI (Smith, 1992), CHIMERA (Posada and Crandall, 2001) and RDP (Martin and Rybicki, 2000) were implemented with default settings within RDP3 (Martin, 2009). Recombination events were only considered if detected by at least two programs with a *P*-value < 0.05 after a Bonferroni correction for multiple comparisons. Sequences with signals of recombination were removed from the phylogenetic analysis. A total of six sequences were removed but counted for recombination. These intragenic recombination analysis results are summarized in the form of a heat map using GnuPlot after normalizing all frequencies to the maximum in the dataset, setting this to 1.0. Finally, recombination similarity was calculated using the proportion of detected intragenic recombination events divided by the maximum count of possible events if all shared proteins had recombined, and converted to a distance (1.0 minus this recombination similarity metric) to be visualized as a cladogram using a hierarchical clustering algorithm in Matlab.

Reference tree phylogenetic reconstruction

Concatenated core gene supertrees. Concatenated gene supertrees were created from various 'core protein' datasets as follows. First, each protein in core T4-GC protein clusters were examined for detectable intragenic recombination and removed where detected resulting in removal of six protein clusters: three in the T4 Core (NrdC, Glutaredoxin; gp5, baseplate and lysozyme; gp48, baseplate tail tube cap) and three in the Cyano T4 Core (two hypothetical proteins, and Hli, high-light inducible proteins). The remaining protein clusters

were aligned with MAFFT (Katoh *et al.*, 2005) using the matrix BLOSSUM 30, gap opening penalty of 1.5, and extension of 0.5, with the option – globalpair. Alignments were trimmed in regions where obvious ambiguities were present, e.g. ragged ends and large gaps, using GBLOCKS (Castresana, 2000; settings: Minimum number of sequences $n \times 0.5 + 1$, maximum number of non-conserved columns 50, and minimum length of a block 5) and inspected manually. The phylogenetic reconstruction was done with RAxML (Stamatakis *et al.*, 2005) using the WAG substitution matrix and a gamma distribution in four discrete categories and a set of invariable sites (WAG + Γ_4 + I).

Whole proteome phylogenies. The genome content phylogeny was reconstructed using the Simpson Similarity Index as established in Snel and colleagues (1999). Briefly, the distance between any two genomes was calculated as the number of genes shared divided by the number of genes in the smallest genome. Then from the distance matrix the tree is built using the Neighbor Joining algorithm, bootstrap values were calculated by re-sampling (100 times), without replacement, half of the genome content matrix. Distance, bootstraps and the tree were created using custom MatLab scripts (freely available at <http://www.eebweb.arizona.edu/faculty/mbsulli/scripts/espinoza.htm>).

Single protein tree reconstructions and tree topology comparisons

Single core gene trees. All proteins used in the concatenated core gene supertrees were also examined singly. Briefly, the protein sequences from members of a T4-GC were aligned then trimmed and refined as stated above. Finally their ML topologies were calculated in RAxML (Stamatakis *et al.*, 2005) under the WAG + I + Γ_4 model of evolution.

To minimize phylogenetic artefacts due to base compositional differences or lineage specific rates we used the LogDet transformed distances to infer the topologies. Then branch lengths and bootstrap support were optimized and calculated by Maximum likelihood as implemented in RAxML using the following model GTR + I + Γ_4 .

Statistical tree topology comparisons. To compare tree topologies (e.g. single gene trees to supertrees), Shimodaira-Hasegawa (SH) tests were used. These SH tests were done in CONSEL (Shimodaira and Hasegawa, 1999; 2001) using the output of the phylogenetic reconstructions of RAxML (Stamatakis *et al.*, 2005). In a SH test the null hypothesis is that the topologies compared are not significantly different (i.e. both topologies are equally good explanations for the aligned data). Thus low *P*-values indicate rejection of this hypothesis. For the comparison, we used all statistically indistinguishable top-scoring ML trees totalling 1538 that were found during the search of the ML tree and bootstraps. We therefore finished with 87 666 SH test *P*-values (57×1538). The heat map was built using custom perl and gnuplot scripts, it reflects the probability of two genes sharing a topology, which was calculated as stated in Filee and colleagues (2006) from the original *P*-values. Rows and columns were reordered to match a cladogram of the similarity between SH test *P*-value profiles that compared topolo-

gies of all considered genes. To test the sensitivity of the SH test results to the input alignments we compared output from each MUSCLE (Edgar, 2004) and MAFFT (Katoh *et al.*, 2005), and found only minor differences (Figs S3 and 4) that do not alter our conclusions. The cut-off between shared and not shared evolutionary history was done based on the congruence or incongruence of individual trees to the reference T4-cyano core tree (Fig. 2A). Furthermore this split in two groups is confirmed by a hierarchical clustering based on the Euclidean distances between all probabilities associated to each gene.

Informatic search for PeSL [Promoter (early) Stem Loop]. PeSLs (Arbiol *et al.*, 2010) in a canonical form have a Promoter and Stem loop on both sides of a gene. We used FindTerm v. 2.8 and Bprom with a threshold of –11.0 and 0.20, respectively, both from Softberry (<http://linux1.softberry.com/berry.phtml>), and included only output that also occurred in intergenic regions. Putative PeSLs were considered if they contained at least three out of four of these transcriptional elements.

Auxiliary metabolic genes. To retrieve the cellular version of phage AMG, all phage genes in an AMG T4-GC were used as queries against the Microbes Online (<http://www.microbesonline.org/>), and NCBI NR (<http://www.ncbi.nlm.nih.gov>) databases. Taxa were kept if their E-value was lower than 10^{-5} and redundant sequences were removed from the resulting retrieved sequence file. Protein trees were reconstructed as stated above initially using all retrieved sequences, but iterating upon the presented subtrees that contain only the viral sequences and neighbouring sequences.

Statistical evaluation of phylogenetic clustering

A *P*-test (Martin, 2002) was implemented in Unifrac (Lozupone and Knight, 2005) to evaluate whether phylogenetic clustering patterns correlated to various environmental metadata. Where environmental factors were not available for the data, interpolated data from Megx.net (Kottmann *et al.*, 2010) were used. The phylogenetic diversity of phage host ranges was calculated using the additive distances from the phylogeny from Roca and colleagues (2002).

Acknowledgements

We thank the members of the Tucson Marine Phage Lab, particularly Melissa Duhaime, for rich discussions provided during the preparation of this manuscript. We acknowledge funding support from the US Department of State for a Fulbright Fellowship to J.C.I.E., as well as BIO5, NSF DBI-0850105 and the Gordon and Betty Moore Foundation to M.B.S.

References

- Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

- Arbiol, C., Comeau, A.M., Kutateladze, M., Adamia, R., and Krisch, H.M. (2010) Mobile regulatory cassettes mediate modular shuffling in t4-type phage genomes. *Genome Biol Evol* **2**: 140–152.
- Botstein, D. (1980) A theory of modular evolution for bacteriophages. *Ann NY Acad Sci* **354**: 484–490.
- Bragg, J.G., and Chisholm, S.W. (2008) Modelling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS ONE* **3**: e3550.
- 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., Thompson, L.R., Suttle, C.S., and Sullivan, M.B. (2007) Exploring the vast diversity of marine viruses. *Oceanography* **20**: 353–362.
- Bryan, M.J., Burroughs, N.J., Spence, E.M., Clokie, M.R.J., Mann, N.H., and Bryan, S.J. (2008) Evidence for the intense exchange of mazG in marine cyanophages by horizontal gene transfer. *PLoS ONE* **3**: e2048.
- Calendar, R. (2006) Classification of bacteriophages. The Bacteriophages. Oxford University Press. 8–16.
- Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.
- Clokie, M.R., Millard, A.D., Letarov, A.V., and Heaphy, S. (2011) Phages in nature. *Bacteriophage* **1**: 31–45.
- Clokie, M.R.J., Shan, J., Bailey, S., Jia, Y., and Krisch, H.M. (2006) Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ Microbiol* **8**: 827–835.
- Coleman, M.L., and Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci USA* **107**: 18634–18639.
- Comeau, A.M., Bertrand, C., Letarov, A., Tétart, F., and Krisch, H.M. (2007) Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**: 384–396.
- Coscollá, M., Comas, I., and González-Candelas, F. (2011) Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila*. *Mol Biol Evol* **28**: 985–1001.
- Dammeyer, T., Bagby, S.C., Sullivan, M.B., Chisholm, S.W., and Frankenberg-Dinkel, N. (2008) Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol* **18**: 442–448.
- 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.
- Desiere, F., Lucchini, S., and Brüssow, H. (1999) Comparative sequence analysis of the DNA packaging, head, and tail morphogenesis modules in the temperate cos-site *Streptococcus thermophilus* bacteriophage Sfi21. *Virology* **260**: 244–253.
- Desplats, C., and Krisch, H.M. (2003) The diversity and evolution of the T4-type bacteriophages. *Res Microbiol* **154**: 259–267.
- Didelot, X., Lawson, D., Darling, A., and Falush, D. (2010) Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**: 1435–1449.
- Dinsdale, E., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.
- Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., *et al.* (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.
- Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol Conserv* **61**: 1–10.
- Filee, J., Tetart, F., Suttle, C.A., and Krisch, H.M. (2005) Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci USA* **102**: 12471–12476.
- Filee, J., Baptiste, E., Susko, E., and Krisch, H.M. (2006) A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol* **23**: 1688–1696.
- Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.
- 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.
- 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.
- Hellweger, F.L. (2009) Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. *Environ Microbiol* **11**: 1386–1394.
- Hendrix, R.W. (2003) Bacteriophage genomics. *Curr Opin Microbiol* **6**: 506–511.
- Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* **96**: 2192–2197.
- Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511–518.
- Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., *et al.* (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**: e231.
- Kitchen, A., Shackelton, L.A., and Holmes, E.C. (2011) Family level phylogenies reveal modes of macroevolution in RNA viruses. *Proc Natl Acad Sci USA* **108**: 238–243.
- Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**: 1929–1940.
- Kottmann, R., Kostadinov, I., Duhaime, M.B., Buttigieg, P.L., Yilmaz, P., Hankeln, W., *et al.* (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38**: D391–D395.
- Lawrence, J.G., and Retchless, A.C. (2009) The interplay of homologous recombination and horizontal gene transfer in

- bacterial speciation. In *Horizontal Gene Transfer: Genomes in Flux*. Gogarten, M.B., Gogarten, J.P., and Olendzenski, L. (eds). New York: Humana Press, pp. 29–53.
- Lerat, E., Daubin, V., and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol* **1**: E19.
- 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.
- Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., *et al.* (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**: 83–86.
- Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.
- Lozupone, C., Hamady, H., and Knight, R. (2006) UniFrac an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7**: 371–385.
- 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.
- Majewski, J. (2001) Sexual isolation in bacteria. *FEMS Microbiol Lett* **199**: 161–169.
- Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Bacterial photosynthesis genes in a virus. *Nature* **424**: 741.
- 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, F.M., and Amrich, C.G. (2009) Recombination and microdiversity in coastal marine cyanophages. *Environ Microbiol* **11**: 2893–2903.
- 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.
- Martin, A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol* **68**: 3673–3682.
- Martin, D., and Rybicki, E. (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**: 562–563.
- Martin, D.P. (2009) Recombination detection and analysis using RDP3. *Methods Mol Biol* **537**: 185–205.
- Millard, A., Clokie, M.R., Shub, D.A., and Mann, N.H. (2004) Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* **101**: 11007–11012.
- Millard, A.D., Zwirgmaier, K., Downey, M.J., Mann, N.H., and Scanlan, D.J. (2009) Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. *Environ Microbiol* **11**: 2370–2387.
- Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Ruger, W. (2003a) Bacteriophage T4 genome. *Microbiol Mol Biol Rev* **67**: 86–156.
- Miller, E.S., Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Durkin, A.S., Ciecko, A., *et al.* (2003b) Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol* **185**: 5220–5233.
- Mosig, G. (1998) Recombination and recombination-dependent DNA replication in bacteriophage T4. *Annu Rev Genet* **32**: 379–413.
- Mosig, G., Gewin, J., Luder, A., Colowick, N., and Vo, D. (2001) Two recombination-dependent DNA replication pathways of bacteriophage T4, and their roles in mutagenesis and horizontal gene transfer. *Proc Natl Acad Sci USA* **98**: 8306–8311.
- Murzin, A.G. (2002) Biochemistry. DNA building block reinvented. *Science* **297**: 61–62.
- Myllykallio, H., Lipowski, G., Leduc, D., Filee, J., Forterre, P., and Liebl, U. (2002) An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* **297**: 105–107.
- Nolan, J.M., Petrov, V., Bertrand, C., Krisch, H.M., and Karam, J.D. (2006) Genetic diversity among five T4-like bacteriophages. *Virology* **33**: 30.
- Padidam, M., Sawyer, S., and Fauquet, C.M. (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**: 218–225.
- Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., *et al.* (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**: 171–182.
- Petrov, V.M., Nolan, J.M., Bertrand, C., Levy, D., Desplats, C., Krisch, H.M., and Karam, J.D. (2006) Plasticity of the gene functions for DNA replication in the T4-like phages. *J Mol Biol* **361**: 46–68.
- Posada, D., and Crandall, K.A. (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* **98**: 13757–13762.
- Proux, C., van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G.F., *et al.* (2002) The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like *Siphoviridae* in lactic acid bacteria. *J Bacteriol* **184**: 6026–6023.
- van Regenmortel, M.H., Mayo, M.A., Fauquet, C.M., and Maniloff, J. (2000) Virus nomenclature: consensus versus chaos. *Arch Virol* **145**: 2227–2232.
- 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. (2003) Global phage diversity. *Cell* **113**: 141.
- Rohwer, F., and Edwards, R. (2002) The phage proteomic tree: a genome-based taxonomy for phage. *J Bacteriol* **184**: 4529–4535.
- Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., *et al.* (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.
- Shen, P., and Huang, H.V. (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112**: 441–457.

- Shimodaira, H., and Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16**: 1114–1116.
- Shimodaira, H., and Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**: 1246–1247.
- 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.
- Smith, M. (1992) Analyzing the mosaic structure of genes. *J Mol Evol* **34**: 126–129.
- Snel, B., Bork, P., and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nat Genet* **21**: 108–110.
- Stamatakis, A., Ludwig, T., and Meier, H. (2005) RaxML-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**: 456–463.
- 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.
- Sullivan, M.B., Coleman, M.L., Quinlivan, V., Rosenkrantz, J.R., DeFrancesco, A.S., Tan, G.P., *et al.* (2008) Portal protein diversity and phage ecology. *Environ Microbiol* **10**: 2810–2823.
- Sullivan, M.B., Huang, K., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R., *et al.* (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12**: 3035–3305.
- Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356–361.
- Suttle, C.A. (2007) Marine viruses – major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.
- 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.
- Tamoi, M., Miyazaki, T., Fukamizo, T., and Shigeoka, S. (2005) The Calvin cycle in cyanobacteria is regulated by CP12 via the NAD(H)/NADP(H) ratio under light/dark conditions. *Plant J* **42**: 504–513.
- Tétart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H.W., and Krisch, H.M. (2001) Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J Bacteriol* **183**: 358–366.
- Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Coleman, M.L., Singer, A.U., *et al.* (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci USA* **108**: 757–764.
- 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.
- Webb, C.O. (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am Nat* **156**: 145–145.
- Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F., *et al.* (2007) Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol* **9**: 1675–1695.
- Weinbauer, M.G., and Rassoulzadegan, F. (2004) Are viruses driving microbial diversification and diversity? *Environ Microbiol* **6**: 1–11.
- 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.
- Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3**: e1456.
- Wilson, W.H., Joint, I.R., Carr, N.G., and Mann, N.H. (1993) Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH7803. *Appl Environ Microbiol* **59**: 3736–3743.
- Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., *et al.* (2007) The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.
- Zinser, E.R., Lindell, D., Johnson, Z.I., Futschik, M.E., Steglich, C., Coleman, M.L., *et al.* (2009) Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. *PLoS ONE* **4**: e5135.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Relationship between genome content and shared protein similarity among 27 T4-like phage genomes. All pairwise comparisons ($n = 351$) were done, and linear regression statistics are presented in the figure.

Fig. S2. Whole genome plot highlighting which of the 57 core genes used in supertree reconstruction are syntenic across the 17 cyano T4-like phage genomes. Lines drawn from one genome to another connect shared core gene homologues. The names of all non-syntenic genes, represented by the darker lines (red or blue), are highlighted at the top of the plot.

Fig. S3. Thymidylate synthase pathways involved in the synthesis of Thymidine monophosphate (dTMP). Notably, the marine picocyanobacteria have adopted different pathways, with *Synechococcus* encoding the canonical ThyA pathway, while *Prochlorococcus* encodes the alternative ThyX pathway.

Fig. S4. Clustering pattern observed using MUSCLE as the alignment program. If the pattern was identical to that from MAFFT generated alignments (the underlying data in Fig. 2A), then all numbers from 1 to 57 would be sequential. Only three changes occur (highlighted with a box around the number), and these changes do not alter our inferences with respect to genes with shared versus unique evolutionary histories.

Table S1. Summary of the statistical tests to evaluate the association of any environmental factor to the phylogenetic patterns observed in the reference supertree presented in Fig. 1A. The tests were done using P test (Martin, 2002) as implemented in Unifrac (Lozupone *et al.*, 2006). Phylogenetic diversity was calculated as the sum of the branch lengths for the subtree from Rocap and colleagues (2002) that included all the hosts per each phage.

Table S2. Prevalence of PeSL-like elements in association with core genes. The number of PeSLs listed is the number identified for each T4-GC across the total of 17 genomes.

File S1. Maximum likelihood topologies of the AMGs assessed in this work, the topologies presented correspond to protein data treated as described in the *Experimental procedures* section. All available cyanobacterial sequences were included in the trees except where lack of sequence similarity prevented the determination of homologous positions in alignments; in these cases, unalignable taxa were excluded from the analyses.

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