# *The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'*



**Citation:** Sullivan, M. B., Krastins, B., Hughes, J. L., Kelly, L., Chase, M., Sarracino, D. and Chisholm, S. W. (2009), The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. Environmental Microbiology, 11: 2935–2951.

**As Published:** http://dx.doi.org/10.1111/j.1462-2920.2009.02081.x

**Persistent URL:** http://hdl.handle.net/1721.1/61354

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

**THE GENOME AND STRUCTURAL PROTEOME OF AN OCEAN CYANOBACTERIAL SIPHOVIRUS: A NEW WINDOW INTO THE CYANOBACTERIAL 'MOBILOME'**

Matthew B. Sullivan[1*], Bryan Krastins[2], Jennifer L. Hughes[3], Libusha Kelly[1], Michael Chase[2], David Sarracino[2] and Sallie W. Chisholm[1]

[1]*Department of Civil and Environmental Engineering and Department of Biology, MIT, 48-425, Cambridge MA 02139, USA*
[2]*Harvard Partners, Cambridge MA 02139*
[3] *Ecology and Evolutionary Biology Department, University of Arizona, Tucson, AZ 85721*
** Present address: Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721*

**ABSTRACT***: Prochlorococcus*, an abundant phototroph in the oceans, are infected by members of three families of viruses: myo-, podo- and siphoviruses. Genomes of myo-and podoviruses isolated on *Prochlorococcus* contain DNA replication machinery and virion structural genes homologous to those from coliphages T4 and T7, respectively. They also contain a suite of genes of cyanobacterial origin, most notably photosynthesis genes, which are expressed during infection, and appear integral to the evolutionary trajectory of both host and phage. Here we present the first genome of a cyanobacterial siphovirus, P-SS2, which was isolated from Atlantic slope waters using a *Prochlorococcus* host (MIT9313). The P-SS2 genome is larger than, and considerably divergent from, previously sequenced siphoviruses. It appears most closely related to lambdoid siphoviruses, with which it shares 13 functional homologs. The ~108kb P-SS2 genome encodes 131 predicted proteins and notably lacks photosynthesis genes which have consistently been found in other marine cyanophage, but does contain 14 other cyanobacterial homologs. While only six structural proteins were identified from the genome sequence, 35 proteins were detected experimentally; these mapped onto capsid and tail structural modules in the genome. P-SS2 is potentially capable of integration into its host as inferred from bioinformatically identified genetic machinery *int, bet, exo* and a 53bp attachment site. The host attachment site appears to be a genomic island that is tied to insertion sequence (IS) activity that could facilitate mobility of a gene involved in the nitrogen-stress response. The homologous region and a secondary IS-element hot-spot in *Synechococcus* RS9917 are further evidence of IS-mediated genome evolution coincident with a probable relic prophage integration event. This siphovirus genome provides a glimpse into the biology of a deep-photic zone phage as well as the ocean cyanobacterial prophage and IS element 'mobilome' .

**INTRODUCTION***:*

Phages (viruses that infect prokaryotes) represent the largest source of uncharacterized genetic diversity in the biosphere (Pedulla et al. 2003). One particular group of these phages, the ocean cyanophages, has been relatively well studied because of the global abundance of their cyanobacterial hosts (Partensky et al. 1999, Waterbury et al. 1979, 1986), for which a number of genome sequences are available (Rocap et al. 2003, Palenik et al. 2003, 2006, Dufresne et al. 2003, 2008, Coleman et al. 2006, Kettler et al. 2007). The abundance of ocean cyanophages often co-varies with cyanobacterial abundance in the wild (Waterbury & Valois 1993, Suttle & Chan 1994, Lu, Chen & Hodson 2001, Sullivan et al. 2003, Muhling et al. 2005). Though estimating the quantitative impact of cyanophages on mortality of their cyanobacterial hosts is challenging due to the current need to compare strain-specific cyanophage titers to total cyanobacterial counts, cyanophages are thought to be responsible for a small, but significant fraction of cell mortality (Waterbury & Valois 1993, Suttle et al. 1994, Fuhrman 2000).

Three morphologies of viruses – myo-, podo-, and siphoviruses – are known to infect ocean cyanobacteria. The myovirus and podovirus cyanophage families have been relatively well characterized by morphology, host range and genomics (Waterbury & Valois 1993, Chen et al. 2002, Sullivan et al. 2003, 2005, Mann et al. 2005), and almost universally contain homologs to their host's photosynthetic machinery, including the core reaction center genes of the photosystem (Mann et al. 2003, Millard et al. 2004, Lindell & Sullivan et al. 2004, Zeidner et al. 2005, Sullivan & Lindell et al. 2006). The core photosynthesis reaction center gene, *psbA*, has been shown to be expressed during infection for a podovirus (Lindell et al. 2005, 2007) and a myovirus (Clokie et al. 2006), and is hypothesized to play a role in cyanophage fitness (Lindell et al. 2007; Bragg & Chisholm 2008, Hellweger 2009). Not only do these genes appear to be important for the cyanophage, but sequence analysis has shown that sub-sections of the phage copy can be traced back to their host genome (Sullivan & Lindell et al. 2006). Thus, ocean cyanophages appear to influence the evolution of cyanobacterial genomes via horizontal gene transfer events, even at the level of the core reaction centers (Zeidner et al. 2005, Sullivan & Lindell et al. 2006).

Cyanophage studies to date have focused on lytic phages, which infect the host, use its machinery to replicate and burst the cell, releasing phage progeny. In contrast, temperate phages infect their hosts and may temporarily insert their DNA into the host genome as a prophage, which is replicated with the host genome as part of the cell cycle. Expression of prophage genes often fundamentally changes the host's physiology – a process known as lysogenic conversion (Calendar, 1988). For example, pathogen-associated toxin genes are commonly encoded by prophages (Boyd, Davis, and Hochhut, 2001; Miao and Miller, 1999; Wagner and Waldor, 2002), which act as a mechanism for horizontally transferring such toxins between microbial species (Banks, Beres, and Musser, 2002). More than 70% of sequenced bacterial genomes contain prophages (Canchaya et al., 2003, Casjens et al. 2003). They often represent the primary constituent of strain-to-strain variability (Baba et al., 2002; Simpson et al., 2000; Smoot et al., 2002; Beres et al., 2002), and their genes are among the most highly expressed genes in genome-wide expression studies (Smoot et al., 2001; Whiteley et al., 2001).

Curiously, the genomes of currently available freshwater and marine cyanobacterial genomes lack identifiable prophage (Canchaya et al., 2003; Casjens, 2003; Dufresne et al., 2003, 2008; Coleman et al., 2006; Kettler et al. 2007), in spite of two lines of indirect evidence that suggest prophages exist in cyanobacteria. First, strain-to-strain variability in marine cyanobacterial genomes is often clustered in genomic islands with signatures of phage and mobile element activity even including phage-like integrase genes (Palenik et al. 2003, Coleman et al., 2006; Kettler et al. 2007, DuFresne et al. 2008). Second, addition of inducing agents to natural seawater communities have yielded increases in culturable *Synechococcus* cyanophage thought to be induced prophage (McDaniel et al., 2002; Ortmann et al., 2002).

Here, we characterize the genome and proteome of an ocean siphovirus that was isolated from 83 meter deep Atlantic Ocean slope waters using *Prochlorococcus* MIT9313 as a host strain. The data are analyzed on their own, and in an evolutionary context using comparative genomics of *Prochlorococcus* and *Synechococcus* genomes. To this end, we uncover basic biology of the siphovirus genome, identify a possible integration site in its host, and explore the evolutionary link between insertion sequence activity and prophage integration in the *Prochlorococcus* and related *Synechococcus* host genomes.

## RESULTS AND DISCUSSION

### *The architecture of the P-SS2 particle and its genome*

P-SS2 has the morphology of a siphovirus, with a ~75nm diameter elongated (~140nm long) capsid and a ~325nm flexible, non-contractile tail (Fig. 1A). This is the largest siphovirus for which a complete genome has been sequenced (Table 1), and the size of its genome is also large: at 107,595bp (Fig. 1B; Table 1) it is surpassed only by the 122kb coliphage T5 genome (Table 1). Of the 131 predicted ORFs in the P-SS2 genome, only 38 have recognizable homologs (Table 2). This is proportionally fewer than in other siphoviruses, where often half the predicted proteins have recognizable homologs (Pedulla et al. 2003, Brussow & Dessiere 2001, Proux et al. 2002). It is, however, proportionally similar to the alpha-proteobacterial marine siphovirus phi-JL001 where only 17 of 91 ORFs had homologous proteins in the database (Lohr et al. 2005), and consistent with the idea that marine siphoviruses encode proteins that are under-represented in the database. Of the 38 ORFs in the P-SS2 genome that have homologs (Table 2), 24 have ascribed functions, eight are hypotheticals predominately from cyanobacteria or their phages, and six are ORFan proteins with a single database match.

Twenty-two of the P-SS2 ORFs appear phage-related, with 13 of these most similar to proteins of the lambdoid siphoviruses and nine most similar to other viral types (Fig. 1B). Six have sequence homology to lambdoid structural proteins (tail fiber, tail collar *gpH*, tail tape measure, host specificity *gpJ*), recombination (*bet*), and lysis (lysozyme) proteins. Another six include 'cyanobacterial' analogs of lambdoid proteins (dCTP deaminase, single-stranded DNA binding protein, integrase, thymidylate synthase, and the small and large subunits of terminase) and the last encodes a 'non-cyanobacterial' (exonuclease, *exo*) lambdoid analog. The remaining nine ORFs with phage-related homologs are most similar to unclassified prophage proteins (ORFs 007, 014, 025, 030, 092), T4-like myovirus proteins (ORFs 028, 067, 080), and a T7-like podovirus protein (ORF 045).

Gene expression in cyanobacteria is coomonly regulated at the level of transcription, so we investigated the potential transcriptional regulatory machinery available to siphovirus P-SS2 (Fig. 1B). A search for classic sigma-70 promoter sequences and rho-independent terminators revealed 19 and 22, respectively (details in methods, genome locations in Suppl. Table 1 and Fig. 1B, weblogo promoter consensus sequence in Suppl. Fig. 1). Notable among these were transcriptionally autonomous genes (termed "morons" by Hendrix et al. 1999) that are thought to be the basis for mosaicism among siphoviruses; examples here include the *Prochlorococcus* MIT9313 ORFan (P-SS2 ORF 053) and Syn5 cyanopodophage ORFan (P-SS2 ORF 045). As well, reverse promoters that were identified often provided support that predicted opposite strand ORFs may indeed be functional (e.g., P-SS2 ORFs 016-020). Sequence analysis also identified two sigma factors that are likely used by P-SS2 to modulate host RNAP activity during infection. The first, a group 2 sigma factor (ORF 113), most probably recognizes the canonical sigma-70 promoter sequences identified above, while the second, a group 3 sigma factor (ORF 003), likely recognizes sequences specific for a particular regulon (Lonetto et al. 1992). Such functionally-specific group 3 sigma factors are uncommon among the sequenced marine *Prochlorococcus* genomes to date; found only in ProMIT9303 and the original P-SS2 host strain, ProMIT9313. However, the P-SS2 group 3 sigma factor appears significantly diverged from that of its host (Suppl. Fig. 2), so if the phage or

host version was acquired from the other entity then it has greatly diverged or one of the two acquired the sigma factor from outside cyanobacteria and their phages.

While the bulk of P-SS2's phage-related proteins are most similar to those from lambdoid phages, P-SS2 is a distantly related lambdoid phage at best. First, the sequence similarity of the aforementioned six 'lambdoid' proteins is quite poor. Second, a phylogeny of the large terminase protein (diagnostic for DNA packaging characteristics; Casjens et al., 2005), suggests that the P-SS2 TerL and the homologs from remnants of marine *Synechococcus* prophage integration events (see discussion below) comprise a novel terminase class quite divergent from known phage terminases (Fig. 2). Third, most of the protein components that comprise the P-SS2 virus particle are unrecognizable; only six structural proteins could be assigned by sequence, as elaborated upon below.

***Structural proteins***

To expand our understanding of the genes encoding the P-SS2 structure, we identified the structural proteins in purified virus particles experimentally using mass spectrometry (see Methods). We detected 35 structural proteins (Table 2, hashed lines in ORFs in Fig. 1B), including all six that were identified by sequence. As is common in phage genomes (Brussow & Desiere 2001, Proux et al. 2002, Casjens 2003), these structural genes were clustered on the genome into 'modules' (red lines below genome in Fig. 1B). The largest cluster, consisting of 28 structural genes, included homologs to tail fiber structural genes (ORFs 067, 072, 074, 077, 091). Notably, one putative tail fiber gene (ORF 073) is predicted to encode a 1627 amino acid protein and has a significantly low %G+C content (Suppl. Fig. 3). This anomalous %G+C, and the fact that the ORF is most similar to a non-siphovirus tail fiber gene from the myovirus P-SSM2 genome, suggests possible horizontal transfer into P-SS2 from another phage class. If true, such tail fiber switching might have significant implications for host-range among cyanophages.

Another five proteins detected in the virus particle were mapped to a genome cluster that contained a capsid protein homolog (ORF 030) and a highly conserved marine prophage protein that was among the most abundant proteins in the proteomics analysis (ORF 025; averaged 33 detected peptides across biological replicates). This region likely defines proteins involved in capsid formation. The last two structural proteins detected in the virus particle are small proteins in the 5'-end of the P-SS2 genome (ORFs 009, 020) with unknown function.

***Cyanobacterial and marine features of the P-SS2 genome***

The P-SS2 genome is 108 kb whereas, with one exception, most of the other siphoviruses sequenced to date have genomes on the order of 20-50 kb (Table 1). What comprises this extra DNA? Unlike the majority of cultured marine myovirus and podovirus (Mann et al. 2003, Millard et al. 2004, Lindell & Sullivan et al. 2004, Lindell et al., 2005, 2007, Sullivan & Lindell et al., 2006), P-SS2 does not encode cyanobacterial photosynthesis genes. Because cyanomyoviruses were commonly isolated from similar deep-photic zone depths that also contain *psbA* (12 are documented in Sullivan & Lindell et al., 2006), we posit that the lack of such photosynthesis genes in siphovirus P-SS2 is more likely to be due to the hypothesized temperate phage lifestyle of this virus.

However, P-SS2 does encode 14 genes with homology to genes from ocean cyanobacteria. Six of these proteins are also phage-encoded in the lambda/*E.coli* system, and likely, as described above for cyanobacterial lambdoid analogs in the genome section have important DNA synthesis and packaging functions. The remaining eight, which have host but not phage parallels in the lambda/*E.coli* system, include a cyanobacterial DNA primase (paired with a phage-encoded non-cyanobacterial DNA helicase), ribonucleotide reductase (*RNR*), cobalamin synthesis gene *cobO*, three conserved marine cyanobacterial hypothetical proteins, and two cyanobacterial ORFan genes. The last six of these genes have not been seen previously in any phage genome, and their functional roles and importance to phage fitness remain unclear. In contrast, primase, helicase, and RNR-encoding genes are common in phage genomes. While

not found in lambda, primase and helicase genes are often present in other siphovirus genomes, including the divergent ocean siphovirus phi-JL001, suggesting that these genes encode critical protein functions not required in lambda. Further, while RNR-encoding genes are uncommon among siphoviruses (found in 12 of 107 siphoviruses at http://rnrdb.molbio.su.se), two lines of evidence suggest their importance in marine ecosystems. First, the other marine siphovirus, phi-JL001, contains a RNR-encoding gene (Lohr et al. 2005). Second, they are also found in non-siphovirus marine phage: all marine T4-like and T7-like phage sequenced to date contain them (Rohwer et al., 2000, Chen et al., 2002, Mann et al., 2005; Sullivan et al., 2005, Pope et al. 2007, Weigele et al., 2007) even though they are absent in non-marine T7-like phages. We hypothesize that the prevalence of RNR-encoding genes in marine phage of all types reflects the importance of scavenging nucleic acids for DNA synthesis during phage infection in the often nitrogen- and phosphorous-limited oceanic environments.

***Host genomic islands and a putative P-SS2 integration site***

The P-SS2 genome contains three of the four genes that are considered hallmark lysogeny genes in lambda: *int*, *exo* and *bet* (Table 2, Fig. 1B). If capable of integration as a prophage, then two more protein functions must be present as well. First, the function of the 4$^{th}$ hallmark lysogeny gene, excisionase, would need be filled by a host-encoded version as has been observed for other phages and plasmids that use host-encoded site-specific recombinases (Huber & Waldor 2002, Barre & Sherratt 2002). Second, a repressor is critical to remain integrated as a prophage to prevent expression of lytic genes that would induce the prophage out of its host genome. While P-SS2 lacks an identifiable repressor, these are small proteins that are highly divergent and often not recognizable even in known functional prophages (e.g., marine *Silicibacter* prophages; Chen et al. 2006). In addition to *int, exo* and *bet*, the P-SS2 genome contains a 53bp intragenic non-protein-coding sequence between *int* and *bet* that exactly matches an intergenic, non-coding region of host *Pro*MIT9313 and includes 36bp that exactly match a nearby tRNA-Met (Fig. 3). No other matches outside of the 36-bp coding region of the tRNA-Met were found in Genbank, the Global Ocean Survey microbial metagenomes, or the microbial genomes at Microbes Online. Prophages commonly integrate into conserved host genome sites such as tRNA (Campbell 2003) and tmRNA (Williams et al., 2002). Thus this exceedingly rare 53bp match to the non-coding host genome sequence which should be prone to amelioration by neutral mutation and that maintains partial identity to a tRNA in the host genome may represent the phage (*attP*) and host (*attB*) site-specific attachment sites. Near the putative integration site in the host genome are signatures of mobile genetic element activity including eight transposase genes, five pseudo-tRNA-Met genes and five copies of 36bp of the 53bp exact match (Fig. 3). However, these mobile genetic elements are likely relics of long-ago transposition events because the transposase genes are variously degraded and lack identifiable inverted-repeat "ends" (see Methods).

While future work is required to experimentally prove that this 53bp match is the integration site for P-SS2 in its host, we chose to examine the available marine *Prochlorococcus* and *Synechococcus* genomes in this tRNA-Met + *ansA* region. Across 21 genomes, this homologous region revealed a complex evolutionary story (Fig. 4A). In each examined host genome, this region is highly conserved right up to the tRNA-Met gene from the *ansA* gene side of this tRNA (blue bar in Fig. 4A). Among some genomes, synteny continues among sub-groups of these strains (*ProI, ProII, ProIII, Synechococcus* labels in Fig. 4A). In contrast, a number of genomes have hypervariable 'genomic island' regions (*sensu* Coleman et al. 2006) at this tRNA breakpoint (red squares in Fig. 4A). Some of these 'island' regions are small, as in the *Prochlorococcus* strains MED4 and MIT9515, and contain numerous *hli* genes, which appear to have been horizontally transferred to these genomes by phages (Lindell & Sullivan et al. 2004). Others are more extensively hypervariable, as in *Pro*MIT9313 (described above, detailed in Fig. 3) and *Syn*RS9917 (detailed in Fig. 4B). The *Syn*RS9917 island region is 41 kb in size and lacks *hli* genes, but contains six transposases and 2 P-SS2-like genes – lysozyme and ORF97 (Fig. 4B). Further, ~27kb of highly

syntenic genome away is a second genomic island of ~42kb, bounded by a tRNA-Ser (Fig. 4B). This island contains another transposase, assorted genes related to prophage induction including maintenance proteins, an anti-repressor and a possible repressor, as well as four P-SS2-like genes – large terminase, integrase, RNAP sigma factor, ORF25 structural gene, lysozyme. This is most certainly a relic prophage that shared some similarity with P-SS2.

In contrast to the *Pro*MIT9313 transposases described above, the bulk of the *Syn*RS9917 transposase genes appear as intact composite IS element with identifiable "ends" (Fig. 5). The presence of two RAD52 family proteins (Fig. 4B) suggests the need for double-stranded DNA repair as if this region were under heightened 'attack' from IS elements. Notably, a second hot-spot of IS elements occurs at *pyrE* in the *Syn*RS9917 genome. This region contains another tRNA-Met and RAD52 family protein, and is syntenic across all host genomes examined except for a 65kb genomic island in *Syn*RS9917 (Fig. 6A). This *Syn*RS9917 island contains 13 phage-related genes (seven of which are similar to P-SS2 genes, Fig. 6B) and, while clearly incomplete, represents the most intact marine cyanobacterial prophage observed to date. Notably, among the 21 *Prochlorococcus* and *Synechococcus* host strains available for testing at the time, P-SS2 infected only its original host used for isolation, *Prochlorococcus* MIT9313 (note: *Syn*RS9917 was not available for testing, Sullivan et al. 2003).

***IS-mediated genome evolution in Prochlorococcus MIT9313***

While the tRNA-Met + AnsA genomic island in ProMIT9313 described above contains variously degraded IS elements, this region may not simply be a "graveyard" of degraded genes and pseudogenes. All five transposases are bordered by paralogous genes with sequence similarity to a nif11-domain (Fig. 3) that suggests a nitrogen stress related function based upon annotation alone. Indeed, nitrogen stress, and not phosphate stress, alters the gene expression of four of five of these nif11-domain-containing genes (Tolonen et al., 2006; Martiny et al., 2007). Across diverse bacterial genomes, insertion sequence elements often alter expression of neighboring genes using outward facing promoters (Mahillion & Chandler 1998). IS elements are also capable of transporting 'cargo' genes around genomes (Bartosik et al., 2008, Poirel et al., 2005, Toleman et al., 2006), often with greater transposition efficiency than wild-type elements (Bartosik et al. 2008) and lead to fitness gains under experimental evolutionary conditions (Schneider & Lenski 2004) and in the generation of reduced symbiont genomes (Moran & Plague 2004, Plague et al. 2008). Thus we hypothesize that these *Pro*MIT9313 IS elements may both move around (as "cargo") and regulate (via outward-facing promoters) these proximal nitrogen stress genes thereby contributing to host niche-differentiation.

Further, this region is coincident with the putative P-SS2 siphovirus integration site, suggesting two layers of genome evolution. IS elements may mediate *intra*genomic innovation by bringing genes to this region as an evolutionary "sandpit" where selection challenges new combinations of alleles and genes as appears to be the case for the nif11-domain genes above. Then, if indeed P-SS2 is capable of integration at this site, the phage could obtain such IS-mediated genetic innovation through aberrant excision and distribute it to other host genomes (*inter*genomic innovation) via new infections. If this were occurring, then one might expect to occasionally observe IS elements from such evolutionary action in phage genomes. Indeed, in spite of very few environmental phage genomes available, IS transposase genes have been observed in marine phages (e.g., vibriophage VHML; Chibani-Chennoffi et al. 2004, Oakey et al. 2002) and freshwater cyanophages (Ma-LMM01 has 3 transposases; Yoshida et al. 2008). Further, IS elements are known to facilitate the spread and expression of genes enabling antibacterial resistance and degradation of toxic compounds (Berg & Howe 1989, Bushman 2002, Nojiri et al., 2004). Perhaps here nitrogen stress response is similarly tied to IS-mediated evolution in marine cyanobacteria, which are often N-limited.

***The acquisition of "host genes" into the cyanophage genome pool***

While "host genes" or "auxiliary metabolic genes" (AMGs) are commonly observed in cyanophage genomes (reviewed in Breitbart et al., 2007), it remains unclear how they are obtained by cyanophages. Notably, a second IS-element hot-spot in *Syn*RS9917 is again located at a tRNA-Met which in 21 other *Prochlorococcus* and *Synechococcus* genomes is proximal to *pyrE, cobS* and *purH* (Figs. 6A) – three genes that are found in lytic myovirus cyanophage genomes (Syn9, P-SSM2 and P-SSM4; Sullivan et al., 2005, Weigele et al. 2007). In contrast, this region in the *Syn*RS9917 genome has significant evidence of past prophage-integration activity including 13 phage-related genes, seven of which share sequence similarity with P-SS2 genes (Fig. 6B). While this prophage is clearly a relic (nine transposases, missing many genes), could such a prophage have introduced these three "host genes" or "auxiliary metabolic genes" (AMGs) into the phage genome pool? Induced prophages can improperly excise from the host genome and mispackage up to 10% of the host genome proximal to the integration site in place of part or all of the phage genome (Calendar 1988). Thus the remnant *Syn*RS9917 prophage(s) may have initially obtained such AMGs proximal to this site-specific integration site. These genes could then have been disseminated to super-infecting lytic cyanophages through recombination. Such prophage-to-lytic-phage recombination events are thought to be among the most probable means of spreading new genetic material through the phage genome pool as has been observed in *Streptococcus thermophilus* and *Lactococcal* phages (Brussow & Desiere 2001), mycobacteriophages (Pedulla et al. 2003), and more generally the siphoviruses (Hendrix et al. 1999).

***Prevalence of P-SS2-like siphoviruses in the surface oceans***

Given that siphoviruses have rarely been isolated in studies using a diversity of marine cyanobacterial hosts to isolate phage from seawater (Suttle & Chan 1994, Waterbury & Valois 1993, Lu, Chen & Hodson 2001, Marson & Sallee 2003, Sullivan et al. 2003), one wonders whether this was a function of isolation procedures or whether they occur in relatively low abundances in the wild. To begin to address this question, we used the P-SS2 genome to 'recruit' homologous fragments from the microbial fraction Global Ocean Survey surface ocean metagenomes (see Methods). There were only seven GOS reads with a best hit to the P-SS2 genome (Suppl. Table 3), and the alignment lengths of these hits were short, ranging from 47 to 242bp. Homologs of the *Pro*MIT9313 genome are also rare in the GOS dataset (<0.35% of the total hits in any given site, data not shown), which is not surprising as these LL-adapted *Prochlorococcus* cells are not abundant in surface waters (Johnson & Zinser et al. 2006). Thus this limited analysis suggests that siphoviruses similar to the one used in this study are not abundant in surface ocean waters, but may be in undersampled lower euphotic zone waters.

**CONCLUSIONS**

The ocean cyanobacterial siphovirus P-SS2 contains a large genome that is significantly divergent from the siphovirus genomes sequenced to date, so much so that even structural proteins required experimental validation to annotate. This contrasts with the classically lytic phages (e.g., T4-like myoviruses and T7-like podoviruses) which exemplify a cohesive genomic architecture ranging from non-marine coliphages (Miller et al. 2003, Nolan et al. 2006) to marine representatives of roseophages (Rohwer et al. 2000), vibriophages (Miller et al. 2003) and cyanophages (Chen et al. 2002, Mann et al. 2005, Sullivan et al. 2005, Weigele et al. 2007, Pope et al. 2007). The siphoviruses, however, are thought to be prone to extensive genetic module 'swapping' through intensive recombination (Hendrix et al. 1999, Juhala et al. 2000, Brussow & Desiere 2001, Proux et al. 2002, Pedulla et al. 2003), thus the divergence observed is not surprising. This intense mosaicism causes siphoviruses to display web-like phylogenies (Brussow & Desiere 2001) and to represent the most taxonomically challenging phage group (Hendrix et al. 1999, Edwards & Rohwer 2002, Lawrence et al. 2002, Proux et al. 2002). Beyond the P-SS2 genome, exploration of a putative phage integration site in the host genome revealed extensive genomic

islands in the host and IS elements among some ocean *Prochlorococcus* and *Synechococcus* genomes at this location. Particularly striking are the IS element hot-spots in *Syn*RS9917 where the co-mingling components of the cyanobacterial 'mobilome' revealed evidence of prophages under IS element attack, as well as a possible mechanism for phage-captured "host " AMGs central to cyanophage biology (reviewed in Breitbart et al. 2007).

**ACKNOWLEDGEMENTS**

The siphovirus P-SS2 genome was sequenced and assembled by the DOE JGI under the auspices of the Community Sequencing Program. This work was supported in part by grants to SWC from the Gordon and Betty Moore Foundation, NSF, DOE-GTL. We thank Maureen Coleman for generating the draft P-SS2 genome figure, Brian Binder and the crew of the R/V Endeavor for the sampling opportunity, and Sherwood Casjens, Patrick Degnan, Howard Ochman, Luke Thompson, Marcia Osburne, Maureen Coleman, Melissa Duhaime and Li Deng for the original terminase alignments (SC) and engaging discussion of insertion sequence elements (PD, HO), ribonucleotide reductases (LT) and siphovirus biology (MO, MC, MD, LD). We thank three anonymous reviewers whose comments also improved the manuscript.

**METHODS**

***Isolation of the phage and preparation for genomic sequencing***

The siphovirus P-SS2 was isolated from Atlantic Ocean slope waters (38°10'N, 73°09'W) collected on 17 September 2001 on the R/V Endeavor cruise number 360. The sampled water was from 83m depth with a salinity 36.6ppt and temperature 20.8°C. This water was 0.2µm filtered and stored at 4°C until it was used directly in a plaque assay with *Prochlorococcus* strain MIT9313 as a host on 12 Dec 2001. A large, well-resolved plaque was picked from the lawn of host cells on 29 Dec 2001, plaque purified two more times and stored as a lysate of a P-SS2 clonal stock isolate.

P-SS2 was prepared for sequencing as previously described (Lindell & Sullivan et al. 2004). Briefly, phage particles were concentrated from large volume (2L) lysates using polyethylene glycol. Concentrated DNA-containing phage particles were purified from other material in phage lysates using a density cesium chloride gradient. Purified phage particles were broken open (SDS/proteinase K), and DNA was extracted (phenol:chloroform) and precipitated (ethanol) yielding small amounts of DNA (<1ng). A custom 1-2kb insert linker-amplified shotgun library was constructed by Lucigen (Middletown, Wisconsin, United States) as described previously (Breitbart et al. 2002). Additional larger insert (3–8 kb) clone libraries were constructed from genomic DNA by the Department of Energy (Joint Genome Institute, Walnut Creek, California, United States) using a similar protocol to provide larger scaffolds during assembly. Inserts were sequenced by the Department of Energy Joint Genome Institute from all clone libraries and used for initial assembly of these phage genomes. The Stanford Human Genome Center Finishing Group (Palo Alto, California, United States) closed the genomes using primer walking.

***Genome annotation***

Gene identification and characterization was done as in Sullivan et al. (2005). Briefly, protein coding genes were predicted using GeneMark and manual curation. Translated ORFs were compared to known proteins in the nonredundant GenBank and in the KEGG databases using the BLASTp program. Where BLASTp e-values were high (>0.001) or no sequence similarity was observed, ORF annotation was aided by the use of PSI-BLAST, gene size, domain conservation, and/or synteny (gene order). Identification of tRNA genes was done using tRNAscan-SE. Additionally, rho-independent transcription terminators were identified with TransTermHP (Kingsford *et al.*, 2007) using default parameters. All terminators had a confidence score >80% with an energy score of <-15 and a tail score of <-6. Bacterial

$\sigma^{70}$ promoters were predicted using BPROM (Softberry, Mount Kisco, NY) using default parameters. All intergenic promoters with a linear discriminant function >3.5 were considered candidate promoters. Inverted repeats of IS elements were found using the Palindrome program in the EMBOSS software suite (Rice *et al.*, 2000). Approximately 200bp upstream and downstream of each putative IS element was fed into Palindrome and the output of many inverted repeats identified were screened manually to identify those that were exact matches and at least 7bp long. IS elements were classified using the ACLAME database blast tool (Leplae *et al.*, 2004). Genome visualizations were done in Artemis (Rutherford et al. 2000), while comparative genomics analyses were greatly aided by the tools available at MicrobesOnline (http://www.microbesonline.org). For figure labels where genes are denoted as "prophage" or "cyanobacterial", these assignments were made using NCBI taxonomy lineages of the top 5 blast hits. In all cases where these are denoted, all 5 top hits were of one of these two organismal types, cyanobacteria or known temperate phages or integrated prophages, with e-values < 0.001. The resulting genome sequence is deposited under Genbank accession #GQ334450.

***Ocean microbial metagenomic analyses***

To determine whether P-SS2 occurred in the wild, we queried the Global Ocean Survey (GOS, Rusch et al. 2007) microbial surface ocean water metagenomes. We created a database of all sequenced marine isolates, including Gordon and Betty Moore Foundation Marine Microbial Initiative genomes, NCBI marine isolates, and cyanophage available from the Genbank and CAMERA databases as of November 2008. Environmental metagenomic reads were BLASTed (blastall -p blastn -e 1e-5 -z 25000000000 -m 7 -a 4 -F "m L" -X 150 -U T) against this database. Best hits to each GOS read were retrieved and filtered by alignment length. Reads with best hits to P-SS2 and *Pro*MIT9313 (Genbank ID: NC_005071) are the focus of this study.

***Large terminase (TerL) and group 3 sigma factor protein phylogenies***

Protein alignments were generated using the Promals web server (Pei & Grishin, 2007, Pei *et al.*, 2007) using default parameters and manually edited as needed. Amino acid distance trees were constructed using the PAUP*4.0b10 software. Neighbor joining was used to reconstruct distance trees using minimum evolution as the objective function and uncorrected distances. Amino acid maximum likelihood trees were inferred using the CIPRES web portal RAxML rapid bootstrapping and ML search (Stamatakis, 2006, Stamatakis *et al.*, 2008) assuming the James-Taylor Thornton model of substitution using empirical base frequencies and estimating the proportion of invariable sites from the data.

***Virion structural proteomics***

Briefly, the samples were incubated in a denaturing solution of 8M Urea/1% SDS/100mM ammonium bicarbonate/10mM DTT pH 8.5 at 37 degrees for 1 hour. Next, the samples were alkylated for one hour by the addition of iodoacetamide to a final concentration of 40mM and then quenched with 2M DTT. Following the addition of 4X LDS loading buffer (Invitrogen), each sample was centrifuged at 14,000 rpm for 5 minutes at room temperature, and each sample was fractionated on a NuPAGE 10% Bis-Tris 10 lane gel (Invitrogen) for 2.5 hours at 125 volts, 50mA and 8W. Gels were shrunk overnight by the addition of 50% ethanol and 7% acetic acid, and then allowed to swell for 1 hour by the addition of deionized water. Gels were stained with SimplyBlue Safe Stain (Invitrogen) for 2-4 hours, imaged, and sliced horizontally into fragments of equal size based on the molecular weight markers.

In-gel digestion was performed after destaining and rinsing the gel sections with two washes of 50% ethanol and 7% acetic acid, followed by two alternating washes with 50 mM ammonium bicarbonate and acetonitrile. After removal of the last acetonitrile wash, 100uL of sequencing grade trypsin (Promega) was added to each gel slice at a concentration of 6.6 ng/uL in 50 mM ammonium bicarbonate/10% acetonitrile. The gel slices were allowed to swell for 30 minutes on ice, after which the

tubes were incubated at 37 degrees for 24 hours. Peptides were extracted with one wash of 100 uL of 50 mM ammonium bicarbonate/10% acetonitrile and one wash of 100 uL of 50% acetonitrile/0.1% formic acid. The extracts were pooled and frozen at -80 degrees, lyophilized to dryness and redissolved in 40uL of 5% acetonitrile, 0.1% formic acid.

Samples were then loaded into a 96-well plate (AbGene) for mass spectrometry analysis on a Thermo Fisher Scientific LTQ-FT. For each run, 10uL of each reconstituted sample was injected with a Famos Autosampler, and the separation was performed on a 75mM x 20cm column packed with $C_{18}$ Magic media (Michrom Biosciences) running at 250 nL/min provided from a Surveyor MS pump with a flow splitter with a gradient of 5-60% water 0.1% formic acid, acetonitrile 0.1% formic acid over the course of 120 minutes (150 min total run). Between each set of samples, standards from a mixture of 5 angiotensin peptides (Michrom Biosciences) were run for 2.5 hours to ascertain column performance and observe any potential carryover that might have occurred. The LTQ-FT was run in a top five configuration with one MS 200K resolution full scan and five MS/MS scans. Dynamic exclusion was set to 1 with a limit of 180 seconds with early expiration set to 2 full scans.

Peptide identifications were made using SEQUEST (ThermoFisher Scientific) through the Bioworks Browser 3.3. The data was searched with a 10ppm window on the MS precursor with 0.5 Dalton on the fragment ions with no enzyme specificity. A reverse database strategy (Elias et al. 2007) was employed with a six frame translation of the genomic sequence reversed and concatenated with the forward sequences supplemented with common contaminates and filtered to obtain a false discovery rate of less than or equal to 1%. Peptides passing the filters were mapped back onto the genome and compared to predicted open reading frames.

**FIGURE LEGENDS**

Figure 1: The morphology (A) and genome / structural proteome (B) of *Prochlorococcus* siphovirus P-SS2.
A. Electron micrograph of uranyl acetate negative-stained, purified P-SS2 viral particle.
B. The open reading frames (ORFs) are indicated either on the positive (above grey line) or negative (below grey line) DNA strand. Bioinformatically determined promoters and terminators are indicated, as is a putative host integration site (see Fig. 2). Structural proteins detected using mass-spectrometry are indicated by the diagonal lines in the corresponding ORFs, with structural modules indicated by the red lines and the text underneath the genome. For further detail, the number of virion structural peptides detected per ORF is provided in Table 2. The genome sequence is deposited in Genbank under accession #GQ334450.

Figure 2: Phylogenetic relationships of the large terminase protein across diverse phage types. This protein is diagnostic of phage DNA packaging mechanisms (Casjens et al. 2005), and was here used to initially characterize the P-SS2 large terminase protein relative to known phage terminases. (*) Denotes marine phage and cyanobacterial host genomes. Notably, the terminase from the other marine siphovirus whose genome is sequenced (phi-JL001) clusters separately from known terminases, while that from cyanophage P-SS2 clusters with terminases from marine cyanobacterial host genomes (likely remnant prophages, see text). The tree shown is a maximum likelihood tree constructed from 1,513 positions (significantly divergent protein and gapped alignment) as described in Methods. Numbers above and below branches represent bootstrap values over 75 from maximum likelihood and distance analyses, respectively. Numbers in parentheses with taxa labels represent number of taxa in collapsed nodes.

Fig. 3. Schematic representation of genome regions surrounding the putative phage (P-SS2) and host (*Prochlorococcus* MIT9313, Genbank ID: NC_005071) integration sites. This site consists of a 53-bp exact match between the phage sequence downstream of its integrase gene at position 90,836 - 90,888, and the non-coding sequence in the host genome at position 912,261 - 912,313. This general region of the host genome is a genomic island, and thus hypervariable (see text). Numbers at the genome ends represent the nucleotide position in the respective genomes.

Figure 4: Genome arrangement at the tRNA-Met + *ansA* locus across (A) *Prochlorococcus* and *Synechococcus* genomes, and (B) detailed for *Synechococcus* RS9917.
A. Comparative genomics of marine *Prochlorococcus* and S*ynechococcus* at the tRNA-Met + *ansA* locus identified as the putative P-SS2 integration site in *Pro*MIT9313. Across the marine cyanobacteria, this region is highly syntenic with four basic genome patterns observed – denoted as *ProI, ProII, ProIII*, and *Synechococcus* in the figure. However, some strains lack synteny and have hypervariable or 'genomic islands' regions, indicated by the red boxes in the figure. MED4 has a small ~ 8kb island with phage high-light inducible genes (this is equivalent to ISL2, Coleman et al. 2006), while MIT9515 has a slightly larger and similar island to MED4's then a region that is a large genome rearrangement (red dashed line) that is syntenic to another region of the MED4 genome (647,805 - 687,505). The eMIT9313 variability in this region is detailed in Fig. 3.
B. The tRNA-Met + AnsA region in *Synechococcus* RS9917 that is homologous to the putative attB integration site in *Pro*MIT9313 from Fig. 3. This ~41kb 'island' region contains four transposases, an antitoxin gene, and two PSS2-like genes –lysozyme and structural protein ORF97. Genomic synteny to all the other marine cyanobacteria then continues for ~27kb until reaching a second ~42kb 'island' that is bounded on the other side by tRNA-Ser, and contains a transposase, as well as numerous prophage-related genes including a possible repressor, anti-repressor, prophage maintenance protein, RNAP sigma

factor, and four PSS2-like genes – large terminase, integrase, ORF25 structural gene, lysozyme. The COG categories refer to those at Microbes Online.

Figure 5: Characterization of insertion sequence (IS) elements in *Synechococcus* RS9917. The 22 transposase genes and surrounding regions (the IS element) revealed four groups of multi-copy IS elements, and five unique or degraded IS elements in the SynRS9917 genome. Using the ACLAME database, we classified these IS elements as follows. The two multi-copy groups "A1" and "A2" are IS3-like mobile elements and have identical inverted repeats, identical lengths and >87% sequence identity. IS group "B", is also IS3-like element but has a shorter inverted repeat. IS group "C" is longer and most similar to IS21-like elements.
A. Location and orientation of IS elements, represented by colored arrows, in relation to the proposed P-SS2 like phage integration sites (regions represented by black bars).
B. Diagrams of IS elements including size of inverted repeat (size in bp indicated above the 5'-end of the yellow box), position and orientation of ORFs, and size of flanking non-coding regions (shown in yellow, with size in bp indicated below the yellow box).

Figure 6: Genomic arrangement of the tRNA-Met + *pyrE* site in *Synechococcus* RS9917 identified as a secondary hot-spot for insertion sequence elements.
A. Schematic of the highly syntenic tRNA-Met + *pyrE* region from representative *Prochlorococcus* and *Synechococcus* genomes. Minor insertions in *Pro*MIT9303 and *Pro*MIT9313 (*rffM* insertion, small hypothetical ORFs) and the marine *Synechococcus* (*rffM* + large hypothetical ORF) are the only deviations from complete synteny in this region, except for the genomic island detailed for *Synechococcus* RS9917 (see Fig. 6B). Gene names are listed for the top genome only, and homologues across the genomes are similarly colored. Red gene names have been previously observed in myovirus cyanophage genomes (Sullivan et al. 2005). Nine other genomes are similar to the *Prochlorococcus* MED4 arrangement, one other for the MIT9313 arrangement, and eight other for the *Synechococcus* WH8102 arrangement (details in Supplementary Table 2).
B. In contrast to the genome conservation observed in other *Prochlorococcus* and *Synechococcus* genomes, *Syn*RS9917 contains a ~65kb genomic island region that contains 9 transposases, 7 P-SS2-like genes and 7 phage-like genes. This is the most intact prophage in any marine *Prochlorococcus* or *Synechococcus* genome, but it is still significantly degraded.

Supplementary Figure 1: Weblogo representation of the consensus promoter sequences predicted across the siphovirus P-SS2 genome. Genomic locations of the predicted promoter sequence locations are presented in Suppl. Table 1 (along with predicted terminators).

Supplementary Figure 2: Phylogenetic relationships of group 3 sigma factors among phages and microbes. In contrast to group 1 sigma factors which are universal among microbes, these group 3 sigma factor transcriptional regulatory proteins are uncommon among microbes. This is particularly notable among the marine *Prochlorococcus* where they are only found in ProMIT9313 and ProMIT9303. Tree details are as in the Fig. 2 legend, and methods, while in-figure table contains taxa names.

Supplementary Figure 3: %G+C plot of the siphovirus P-SS2 genome. The black line indicates a sliding base-pair (100bp) window of %G+C along the genome, while the red line indicates 2.5 times the standard deviation. Notably the major deviation from the genome average (highlighted with the gray box) is where the anomalous tail fiber protein of putative lateral gene transfer origin.

Table 1: Genome-wide characteristics of marine siphoviruses P-SS2 (this study) and phi-JL001 (Lohr et al. 2005) relative to other recognized phage groups within the Siphoviridae. Siphoviruses are all non-enveloped and contain double-stranded DNA genomes, non-contractile, flexible tails, and are distinguished by different combinations of alleles of structural and DNA replication proteins.

| | Genome features | | | Particle features | |
|---|---|---|---|---|---|
| Phage Genus[&] | Size (kb) | # ORFs | % G+C | Capsid diameter (nm) | Tail (nm) – L x W |
| *Marine, non-classified siphoviruses* | | | | | |
| cyanophage P-SS2 | 108 | 131 | 52.3 | 75 | 325 x 12 |
| Alpha-protebacteria ϕJL001 | 63 | 91 | 62 | 75 | 125 x N.D. |
| *Lambda-likes* [v] | | | | | |
| Enterobacteria phage λ | 48.5 | 92 | 49 | 60 | 150 x 8 |
| Enterobacteria phage HK022 | 40.8 | 57 | 49 | 51 | 106 x N.D. |
| Enverobacteria phage HK97 | 39.7 | 62 | 49 | 54 | 179 x N.D. |
| *T1-likes* | | | | | |
| Enterobacteria phage T1 | 48.8 | 78 | 45 | 60 | 150 x 8 |
| Enterobacteria phage TLS | 49.9 | 87 | 42 | 50 | N.D. |
| Enterobacteria phage RTP | 46.2 | 75 | 44 | 60 | 160 x N.D. |
| *L5-likes* | | | | | |
| Mycobacterium phage L5 | 52.3 | 88 | 62 | 60 | 135 x 8 |
| Mycobacterium phage D29 | 49.1 | 84 | 63 | N.D. | N.D. |
| Mycobacterium phage Bxb1 | 50.6 | 86 | 63 | 60 | 135 x N.D. |
| *ϕC31-likes* | | | | | |
| Streptomyces phage ϕC31 | 41.5 | 54 | 63 | 53 | 100 x 5 |
| Streptomyces phage ϕBT1 | 41.8 | 56 | 62 | N.D. | N.D. |
| *N15-likes* | | | | | |
| Enterobacteria phage N15 | 46.4 | 60 | 51 | 60 | 140 x 8 |
| *T5-likes* | | | | | |
| Enterobacteria phage T5 | 121.7 | 195 | 39 | 80 | 180 x 9 |
| *c2-likes* | | | | | |
| Lactococcus phage bIL67 | 22.2 | 37 | 35 | 41 | 98 x 9 |
| Lactococcus phage c2 | 22.2 | 41 | 36 | N.D. | N.D. |
| *ψM1-likes* | | | | | |
| Methanobacterium phage *ψM1* | 26.1 | 31 | 46 | 55 | 210 x 10 |

[&] genus as recognized by the International Committee on the Taxonomy of Viruses (van Regenmortel et al. 2000) and recently described Sfi21-like siphovirus families (Proux et al. 2002)

[v] There are 19 sequenced genomes currently recognized as part of the lambda supergroup. Here we present a representative genome from each major group.

* genome sizes are from the classified siphovirus genomes from the NCBI TaxBrowser database

Table 2: Summary table of P-SS2 predicted proteins that contained relevant annotation information as determined from (a) significant BLASTP hits (e-value < e-3) against the Genbank non-redundant database, (b) experimental proteomics on the virus particle, or (c) detection in viral metagenomes. For each protein, the genome locus information is paired with our annotations, as well as the top e-value and the average number of peptides detected from 3 biological replicate proteomic analyses (see text and methods).

| P-SS2 ORF # | Strand | LeftEnd | RightEnd | Size (aa) | gene | Putative function | e-value | Avg # peptides detected |
|---|---|---|---|---|---|---|---|---|
| 001 | + | 1 | 528 | 176 | terS | terminase - small subunit | $e^{-8}$ | 0 |
| 002 | + | 525 | 2018 | 498 | terL | terminase - large subunit | $e^{-63}$ | 0.5 |
| 003 | + | 2091 | 2732 | 214 | type III rpoS | cyanobacterial type III RNAP sigma factor | $e^{-13}$ | 0 |
| 005 | + | 3417 | 3608 | 64 | | unknown protein in metagenomes | no hits | 0 |
| 009 | + | 6004 | 6192 | 63 | | structural protein | no hits | 1.5 |
| 010 | + | 6423 | 6713 | 97 | thioredoxin | thioredoxin | $e^{-3}$ | 0 |
| 011 | + | 6853 | 9423 | 857 | nrd | cyanobacterial class II ribonucleotide reductase | e = 0 | 0 |
| 014 | + | 10792 | 11481 | 230 | | hypothetical protein | $e^{-5}$ | 0 |
| 020 | + | 13316 | 13579 | 88 | | unknown structural protein | no hits | 2.5 |
| 025 | + | 15118 | 16674 | 519 | | structural prophage protein | $e^{-46}$ | 33 |
| 028 | + | 17092 | 17280 | 63 | | conserved T4-like protein in metagenomes | $e^{-5}$ | 0 |
| 030 | + | 17648 | 22147 | 1500 | | major capsid protein | $e^{-18}$ | 96.5 |
| 031 | - | 22144 | 22347 | 68 | | unknown structural protein, also in metagenomes | no hits | 1.5 |
| 032 | + | 22350 | 22535 | 62 | | unknown structural protein | no hits | 1 |
| 033 | + | 22567 | 22767 | 67 | | unknown structural protein | no hits | 2.5 |
| 036 | + | 23802 | 24836 | 345 | cobO | cyanobacterial *cobO* | $e^{-98}$ | 0 |
| 038 | + | 25557 | 25730 | 58 | | conserved marine cyanobacterial protein | $e^{-12}$ | 0 |
| 045 | + | 27442 | 27702 | 87 | Syn5_026 | cyanopodophage Syn5 ORFan protein (gp26) | $e^{-13}$ | 0 |
| 049 | + | 28259 | 28486 | 76 | | conserved marine *Synechococcus* protein | $e^{-4}$ | 0 |
| 053 | + | 29861 | 30058 | 66 | 9313_1008 | *Prochlorococcus* eMIT9313 ORFan protein | $e^{-3}$ | 0 |
| 058 | + | 31797 | 32363 | 189 | kinase | possible phage kinase | $e^{-5}$ | 0.5 |
| 061 | + | 32988 | 33524 | 179 | | unknown structural protein | no hits | 16 |
| 062 | + | 33526 | 33993 | 156 | | unknown structural protein | no hits | 3 |
| 063 | + | 33993 | 34499 | 169 | | unknown structural protein | no hits | 5 |
| 066 | + | 36202 | 36831 | 210 | | unknown structural protein | no hits | 2.5 |
| 067 | + | 36831 | 37625 | 265 | fiber | cyanophage T4-like fiber | $e^{-7}$ | 9.5 |
| 068 | + | 37635 | 42854 | 1740 | fiber | unknown structural protein, tail fiber | e=0.015 | 6 |
| 069 | + | 42886 | 43926 | 347 | | unknown structural protein | no hits | 13 |
| 071 | + | 44422 | 44988 | 189 | | unknown structural protein | no hits | 1 |
| 072 | + | 44988 | 46352 | 455 | fiber | lambdoid phage tail fiber | $e^{-11}$ | 3 |
| 073 | + | 46354 | 51234 | 1627 | fiber | tail fiber with low %G+C | $e^{-35}$ | 5 |
| 074 | + | 51512 | 52852 | 447 | capsid decoration protein | lambdoid tail collar/fiber decoration protein (gpH) | $e^{-45}$ | 2 |
| 076 | + | 53164 | 53838 | 225 | | unknown structural protein | no hits | 9.5 |
| 077 | + | 54104 | 59761 | 1886 | tail tape measure | lambdoid tail tape measure protein | $e^{-36}$ | 102 |
| 078 | + | 59795 | 60211 | 139 | | unknown structural protein | no hits | 3 |
| 079 | + | 60216 | 63218 | 1001 | | unknown structural protein | no hits | 43 |

Table 2, continued

| P-SS2 ORF # | Strand | LeftEnd | RightEnd | Size (aa) | gene | Putative function | e-value | Avg # peptides detected |
|---|---|---|---|---|---|---|---|---|
| 080 | + | 63255 | 63593 | 113 | | cyanophage T4-like hypotheticals | $e^{-5}$ | 9 |
| 081 | + | 63603 | 64040 | 146 | | unknown structural protein | no hits | 7.5 |
| 082 | + | 64040 | 64306 | 89 | M2_082 | cyanophage P-SSM2 ORFan protein (gp082) | $e^{-6}$ | 3.5 |
| 083 | + | 64460 | 64657 | 66 | | unknown structural protein | no hits | 2 |
| 084 | + | 64656 | 72495 | 2613 | | unknown structural protein, also in metagenomes | no hits | 43 |
| 085 | + | 72530 | 73087 | 186 | | unknown protein in metagenomes | no hits | 0 |
| 086 | + | 73129 | 78666 | 1846 | | structural protein similar to marine siphophage JL001 ORFan protein (gp88) | $e^{-4}$ | 45 |
| 087 | + | 78666 | 80936 | 757 | | structural protein similar to cyanophage MaTMM01 ORFan protein (gp105) | $e^{-10}$ | 11 |
| 088 | + | 80960 | 82057 | 366 | | unknown structural protein | no hits | 17.5 |
| 089 | + | 82057 | 82920 | 288 | | unknown structural protein | no hits | 6 |
| 090 | + | 82920 | 83348 | 143 | | unknown structural protein | no hits | 4 |
| 091 | + | 83401 | 84141 | 247 | J | lambdoid host specificity protein (gpJ) | $e^{-4}$ | 14 |
| 092 | + | 84331 | 85251 | 307 | | structural cyanobacterial prophage protein | $e^{-65}$ | 43.5 |
| 093 | + | 85317 | 85643 | 109 | | unknown structural protein | no hits | 12.5 |
| 095 | + | 86065 | 86304 | 80 | | unknown protein in metagenomes | no hits | 0 |
| 097 | - | 86560 | 87231 | 224 | hyp_Syn | SynRS9917 ORFan protein | $e^{-23}$ | 0 |
| 098 | + | 87302 | 88528 | 409 | lysozyme | lysozyme | $e^{-11}$ | 0 |
| 101 | - | 89241 | 90614 | 458 | int | site-specific integrase (int) | $e^{-12}$ | 0 |
| 102 | - | 91145 | 92104 | 320 | bet | recombination protein (bet) | $e^{-15}$ | 0 |
| 103 | - | 92216 | 92956 | 247 | | conserved cyanobacterial protein | $e^{-5}$ | 0 |
| 108 | + | 94460 | 96085 | 542 | helicase | DNA helicase | $e^{-8}$ | 0 |
| 109 | + | 96089 | 97663 | 525 | primase | cyanobacterial DNA primase | $e^{-58}$ | 0.5 |
| 111 | + | 98065 | 98655 | 197 | dcd | cyanobacterial dCTP deaminase (dcd) | $e^{-19}$ | 0 |
| 113 | + | 98987 | 99841 | 285 | type II rpoS | type II RNAP sigma factor (rpoS) | $e^{-17}$ | 0 |
| 114 | + | 99889 | 100242 | 118 | ssb | cyanobacterial single-stranded DNA binding protein (ssb) | $e^{-22}$ | 0 |
| 123 | + | 103253 | 104206 | 318 | exo | 5'-3' exonuclease recombination protein (exo) | $e^{-11}$ | 0 |
| 126 | + | 105166 | 106119 | 318 | thy1 | cyanobacterial thymidylate synthase | $e^{-57}$ | 0 |

Supplementary Table 1: Genomic locations of predicted promoters and terminators.

| PROMOTERS | Start | End | strand | -35 seq | -10 seq |
|---|---|---|---|---|---|
| PROM_1 | 2025 | 2053 | + | ttgaca | tggtatcag |
| PROM_2 | 2751 | 2778 | + | ttgaag | ttttttcat |
| PROM_3 | 5658 | 5689 | + | ttctcc | tttttttat |
| PROM_4 | 9431 | 9459 | + | ttgaca | cgccatgat |
| PROM_5 | 10083 | 10111 | + | ttgaca | cgccatgat |
| PROM_6 | 10735 | 10763 | + | ttgaca | cgccatgat |
| PROM_7 | 24917 | 24943 | + | ctgtca | tcttataat |
| PROM_8 | 25784 | 25811 | + | ttgaca | atgtattct |
| PROM_9 | 26384 | 26413 | + | ttgaca | ccctatctt |
| PROM_10 | 26927 | 26955 | + | ttgaca | aggtaactt |
| PROM_11 | 29803 | 29835 | + | ttgata | cggtagaat |
| PROM_12 | 32692 | 32721 | + | tcgtta | gggtagact |
| PROM_13 | 84300 | 84327 | + | tttctt | aggtaattt |
| PROM_14 | 88983 | 89015 | + | ttgctc | aggtagcct |
| PROM_15 | 12296 | 12281 | - | ttgaca | agtcatatt |
| PROM_16 | 12945 | 12930 | - | ttgaca | tgtcatact |
| PROM_17 | 13268 | 13253 | - | ttgaca | ggtcatact |
| PROM_18 | 23168 | 23150 | - | ttaata | tcttatact |
| PROM_19 | 92159 | 92141 | - | ttgcct | cgctatgat |

| TERMINATORS | Start | End | strand | sequence |
|---|---|---|---|---|
| TERM_1 | 2756 | 2769 | + | gggactacggtccc |
| TERM_2 | 5662 | 5679 | + | cccccggcccaccgggga |
| TERM_3 | 6744 | 6759 | + | gcccctcttgaggggc |
| TERM_4 | 9462 | 9480 | + | gggagggttaagccctccc |
| TERM_5 | 10114 | 10132 | + | gggagggttaagccctccc |
| TERM_6 | 10766 | 10784 | + | gggagggttaagccctccc |
| TERM_7 | 11506 | 11527 | + | gggacgcccctagtcgcgtccc |
| TERM_8 | 15047 | 15063 | + | gcccctctaggaggggc |
| TERM_9 | 22802 | 22831 | + | gaccctgggagagtccatgctcctggggtc |
| TERM_10 | 25749 | 25765 | + | ccccgcctagtgcgggg |
| TERM_11 | 26113 | 26138 | + | agccctaagggtggtgccttagggcc |
| TERM_12 | 26893 | 26909 | + | gccccccctcgaggggggc |
| TERM_13 | 27149 | 27162 | + | gggcctacgggccc |
| TERM_14 | 27886 | 27902 | + | gggctccctaggggccc |
| TERM_15 | 29772 | 29789 | + | gcctcccgcaaggggggc |
| TERM_16 | 30163 | 30176 | + | ccccctcgaggggg |
| TERM_17 | 30228 | 30240 | + | gagagctcctctc |
| TERM_18 | 53859 | 53874 | + | gggggcttcggccccc |
| TERM_19 | 84167 | 84186 | + | ggccctgcaacagcagggtc |
| TERM_20 | 85665 | 85682 | + | gcctcccttacggggggc |
| TERM_21 | 106155 | 106172 | + | gcccccatatctgggggc |
| TERM_22 | 107068 | 107084 | + | ggccccttcaaggggcc |

Supplementary Table 2: Details of genomic regions sharing synteny with representative genomes presented in Fig. 6A.

| ***Isolates*** | ***Genome region (nt position)*** |
|---|---|
| *Same genome arrangement as MED4* | |
| MIT9215 | 253,135 - 275,635 |
| AS9601 | 256,220 - 278,720 |
| MIT9211 | 267,649 - 290,149 |
| MIT9301 | 255,803 - 278,720 |
| MIT9312 | 246,647 - 269,147 |
| MIT9515 | 265,543 - 288,043 |
| NATL1A | 309,245 - 331,745 |
| NATL2A | 291,439 - 313,939 |
| SS120 | 280,207 - 302,707 |
| *Same genome arrangement as MIT9313* | |
| MIT9303 | 2,174,404 - 2,196,904 |
| *Same genome arrangement as WH8102* | |
| BL107 | 72,445 - 94,945 |
| CC307 | 1,997,682 - 2,020,182 |
| CC9311 | 283,817 - 306,317 |
| CC9605 | 235,407 - 257,907 |
| CC9902 | 277,436 - 298,686 |
| RS9916 | 57,065 - 79,565 |
| WH7803 | 302,592 - 324,992 |
| WH7805 | 220,710 - 252,960 |

Supplementary Table 3: Environmental sequence reads from the Global Ocean Survey (Rusch et al. 2007) that were best hits to the P-SS2 genome.

| **Read name** | **GOS site** | **e-value** | **Alignment length (bp)** |
|---|---|---|---|
| JCVI_READ_1091145058945 | GS004 | 2.03E-08 | 47 |
| JCVI_READ_1091145481933 | GS004 | 2.21E-08 | 47 |
| JCVI_READ_1091141253121 | GS015 | 1.22E-31 | 242 |
| JCVI_READ_1095522140668 | GS033 | 3.04E-07 | 73 |
| JCVI_READ_1092961193753 | GS051 | 1.23E-06 | 72 |
| JCVI_READ_1108830208947 | GS114 | 4.41E-09 | 40 |
| JCVI_READ_1105297207090 | GS110b | 5.58E-06 | 95 |

References:


Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K., Oguchi, A. et al. (2002) Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* **359**: 1819-1827.

Banks, D.J., Beres, S.B., and Musser, J.M. (2002) The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends in Microbiology* **10**: 515-521.

Barre, F.X., and Sherratt, D.J. (2002) In *Mobile DNA II*. Craig, N.L., Craigie, R., Gellert, M., and Lambowitz, A.M. (eds). Washington DC: ASM Press, pp. 149-161.

Bartosik, D., Putyrski, M., Dziewit, L., Malewska, E., Szymanik, M., Jagiello, E. et al. (2008) Transposable modules generated by a single copy of insertion sequence ISPme1 and their influence on structure and evolution of natural plasmids of Paracoccus methylutens DM12. *J Bacteriol* **190**: 3306-3313.

Beres, S.B., Sylva, G.L., Barbian, K.D., Lei, B., Hoff, J.S., Mammarella, N.D. et al. (2002) Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Natl Acad Sci U S A* **99**: 10078-10083.

Berg, D. E., and M. M. Howe. 1989. Mobile DNA. ASM Press, Washington, D.C.

Boyd, E.F., Davis, B.M., and Hochhut, B. (2001) Bacteriophage-bacteriophage interactions in the evolution of pathogenic bacteria. *Trends in Microbiology* **9**: 137-144.

Breitbart, M., L.T. Thompson, C.A. Suttle and M.B. Sullivan. 2007. Exploring the vast diversity of marine viruses. *Oceanography.* 20: 135-139.

Brussow, H., and Desiere, F. (2001) Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. *Mol Microbiol* **39**: 213-222.

Bushman, F. 2002. Lateral DNA transfer: mechanisms and consequences. Cold Spring Harbor University Press, Cold Spring Harbor, N.Y.

Calendar, R. (1988) *The Bacteriophages*. New York: Plenum.

Campbell, A. (2003) Prophage insertion sites. *Res Microbiol* **154**: 277-282.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003) Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**: 238-276.

Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology* **49**: 277-300.

Casjens, S., Gilcrease, EB, Winn-Stapley, DA, Schicklmaier, P, Schmieger, H, Pedulla, ML, Ford, ME, Houtz, JM, Hatfull, GF, Hendrix, RW. (2005) The generalized transducing *Salmonella* bacteriophage ES18: Complete genome sequence and DNA packaging strategy. *Journal of Bacteriology* **187**: 1091-1104.

Chen, F., and Lu, J. (2002) Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Applied Environmental Microbiology* **68**: 2589-2594.

Chen, F., Wang, K., Stewart, J., and Belas, R. (2006) Induction of multiple prophages from a marine bacterium: a genomic approach. *Applied Environmental Microbiology* **72**: 4995-5001.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.L., and Brussow, H. (2004) Phage-host interaction: an ecological perspective. *J Bacteriol* **186**: 3677-3686.

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. *Environmental Microbiology* **8**: 827-835.

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.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V. et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**: 10020-10025.

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.

Edwards, R.A., Olsen, G.J., and Maloy, S.R. (2002) Comparative genomics of closely related salmonellae. *Trends in Microbiology* **10**: 94-99.

Elias J.E., and Gygi S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 4:207-14.

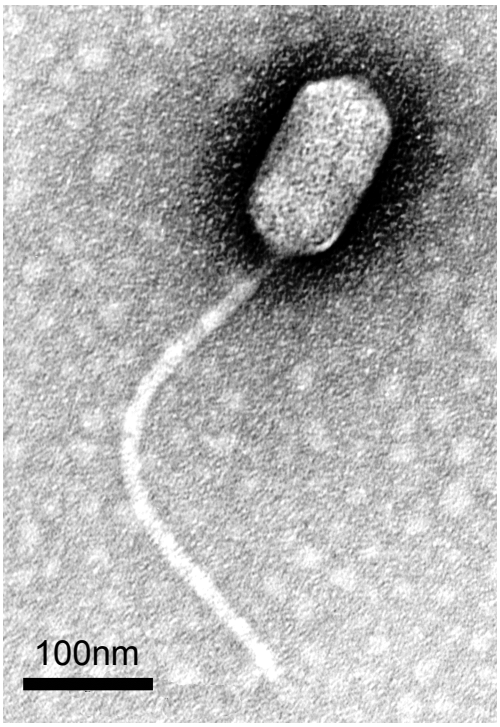
Fuhrman, J.A. (2000) Impact of viruses on bacterial processes. In *Microbial Ecology of the Oceans*. Kirchman, D.L. (ed): Wiley-Liss, Inc., pp. 327-350.

Hellweger. (2009) Carrying photosynthesis genes increases ecological fitness of cyanophage *in silico* . *Environmental Microbiology.* e-pub ahead of print. PMID*:* 19175665.
Hendrix, R.W., Smith, M.C., 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 U S A* **96**: 2192-2197.
Huber, K.E., and Waldor, M.K. (2002) Filamentous phage integration requires the host recombinases XerC and XerD. *Nature* **417**: 656-659.
Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., and Hendrix, R.W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **299**: 27-51.
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.
Kingsford, C. L., Ayanbule, K. & Salzberg, S. L. 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* **8**:R22.
Lawrence, J.G., Hatfull, G.F., and Hendrix, R.W. (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriology* **184**: 4891-4905.
Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. 2004. ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res* **32**:D45-9.
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.
Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86-89.
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 U S A* **101**: 11013-11018.
Lohr, J.E., Chen, F., and Hill, R.T. (2005) Genomic analysis of bacteriophage PhiJL001: insights into its interaction with a sponge-associated alpha-proteobacterium. *Applied Environmental Microbiology* **71**: 1598-1609.
Lonetto, M., Gribskov, M., and Gross, C.A. (1992) The sigma-70 family: Sequence conservation and evolutionary relationships. *J. Bacteriology* **174**: 3843-3849.
Lu, J., Chen, F., and Hodson, R.E. (2001) Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Applied Environmental Microbiology* **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. Bacteriology* **187**: 3188-3200.
Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Bacterial photosynthesis genes in a virus. *Nature* **424**: 741.
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. *Applied Environmental Microbiology* **69**: 4639-4647.
Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci U S A* **103**: 12552-12557.
McDaniel, L., Houchin, L.A., Williamson, S.J., and Paul, J.H. (2002) Lysogeny in marine *Synechococcus*. *Nature* **415**: 496.
Miao, E.A., and Miller, S.I. (1999) Bacteriophages in the evolution of pathogen-host interactions. *Proc Natl Acad Sci U S A* **96**: 9452-9454.
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 U S A* **101**: 11007-11012.
Miller, E.S., Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Durkin, A.S., Ciecko, A. et al. (2003) Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriology* **185**: 5220-5233.
Moran, N.A., and Plague, G.R. (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* **14**: 627-633.
Muhling, 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. *Environmental Microbiology* **7**: 499-508.
Nojiri, H., Shintani, M., and Omori, T. (2004) Divergence of mobile genetic elements involved in the distribution of xenobiotic-catabolic capacity. *Appl Microbiol Biotechnol* **64**: 154-174.
Oakey, H.J., Cullen, B.R., and Owens, L. (2002) The complete nucleotide sequence of the *Vibrio harveyi* bacteriophage VHML. *J Appl Microbiol* **93**: 1089-1098.
Ortmann, A.C., Lawrence, J.E., and Suttle, C.A. (2002) Lysogeny and lytic viral production during a bloom of the

cyanobacterium *Synechococcus* spp. *Microb. Ecol.* **43**: 225-231.
Palenik, B., Brahamsha, B., McCarren, J., Waterbury, J., Allen, E., Webb, E.A. et al. (2003) The genome of a motile marine *Synechococcus*. *Nature* **424**: 1037-1041.
Palenik, B., Ren, Q., Dupont, C.L., Myers, G.S., Heidelberg, J.F., Badger, J.H. et al. (2006) Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc Natl Acad Sci U S A* **103**: 13555-13559.
Partensky, F., Hess, W.R., and Vaulot, D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Molc. Biol. Rev.* **63**: 106-127.
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.
Pei, J. & Grishin, N. V. 2007. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* **23**:802-8.
Pei, J., Kim, B. H., Tang, M. & Grishin, N. V. 2007. PROMALS web server for accurate multiple protein sequence alignments. *Nucleic Acids Res* **35**:W649-52.
Plague, G.R., Dunbar, H.E., Tran, P.L., and Moran, N.A. (2008) Extensive proliferation of transposable elements in heritable bacterial symbionts. *J Bacteriol* **190**: 777-779.
Poirel, L., Lartigue, M.F., Decousser, J.W., and Nordmann, P. (2005) ISEcp1B-mediated transposition of blaCTX-M in Escherichia coli. *Antimicrob Agents Chemother* **49**: 447-450.
Pope WH, W.P., Chang J, Pedulla ML, Ford ME, Houtz JM, Jiang W, Chiu W, Hatfull GF, Hendrix RW, King J. (2007) Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a "horned" bacteriophage of marine *Synechococcus*. *J Mol Biol* **368**: 966-981.
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**: 6026-6036.
Rice, P., Longden, I. & Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**:276-7.
Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A. et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042-1047.
Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F., and Azam, F. (2000) The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* 45: 408-418.
Rutherford K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*.16: 944-5.
Schneider, D., and Lenski, R.E. (2004) Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res Microbiol* **155**: 319-327.
Semsey, S., Blaha, B., Koles, K., Orosz, L., and Papp, P. (2002) Site-specific integrative elements of Rhizobiophage *16-3* can integrate into proline tRNA(CGG) genes in different bacterial genera. *J. Bacteriology* **184**: 177-182.
Simpson, A.J., Reinach, F.C., Arruda, P., Abreu, F.A., Acencio, M., Alvarenga, R. et al. (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature* **406**: 151-157.
Smoot, J.C., Barbian, K.D., Van Gompel, J.J., Smoot, L.M., Chaussee, M.S., Sylva, G.L. et al. (2002) Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* **99**: 4668-4673.
Smoot, L.M., Smoot, J.C., Graham, M.R., Somerville, G.A., Sturdevant, D.E., Migliaccio, C.A. et al. (2001) Global differential gene expression in response to growth temperature alteration in group A *Streptococcus*. *Proc Natl Acad Sci U S A* **98**: 10416-10421.
Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688-90.
Stamatakis, A., Hoover, P. & Rougemont, J. 2008. A Rapid Bootstrap Algorithm for the RAxML Web-Servers. *Systematic Biology*:in press.
Sullivan, M.B., Coleman, M., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biology* **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 Biology* **4**: e234.
Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium

*Prochlorococcus*. *Nature* **424**: 1047-1051.
Suttle, C.A., and Chan, A.M. (1994) Dynamics and distribution of cyanophages and their effects on marine *Synechococcus* spp. *Applied. Environ. Microbiol.* **60**: 3167-3174.
Toleman, M.A., Bennett, P.M., and Walsh, T.R. (2006) ISCR elements: novel gene-capturing systems of the 21st century? *Microbiol Mol Biol Rev* **70**: 296-316.
Tolonen, A.C., Aach, J., Lindell, D., Johnson, Z.I., Rector, T., Steen, R. et al. (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol Syst Biol* **2**: 53.
Wagner, P.L., and Waldor, M.K. (2002) Bacteriophage control of bacterial virulence. *Infect Immun* **70**: 3985-3993.
Waterbury, J.B., and Valois, F.W. (1993) Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophage abundant in seawater. *Applied and Environmental Microbiology* **59**: 3393-3399.
Waterbury, J.B., Watson, S.W., Guillard, R.R.L., and Brand, L.E. (1979) Widespread occurrence of a unicellular marine planktonic cyanobacterium. *Nature* **277**: 293-294.
Waterbury, J.B., Watson, S.W., Valois, F.W., and Franks, D.G. (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.* **214**: 71-120.
Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F. et al. (2007) Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol* **9**: 1675-1695.
Whiteley, M., Bangera, M.G., Bumgarner, R.E., Parsek, M.R., Teitzel, G.M., Lory, S., and Greenberg, E.P. (2001) Gene expression in *Pseudomonas aeruginosa* biofilms. *Nature* **413**: 860-864.
Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nuc. Acids Res.* **30**: 866-875.
Yoshida, T., Nagasaki, K., Takashima, Y., Shirai, Y., Tomaru, Y., Takao, Y. et al. (2008) Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *J Bacteriol* **190**: 1762-1772.
Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., and Beja, O. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environmental Microbiology* **7**: 1505-1513.

Figure 1A



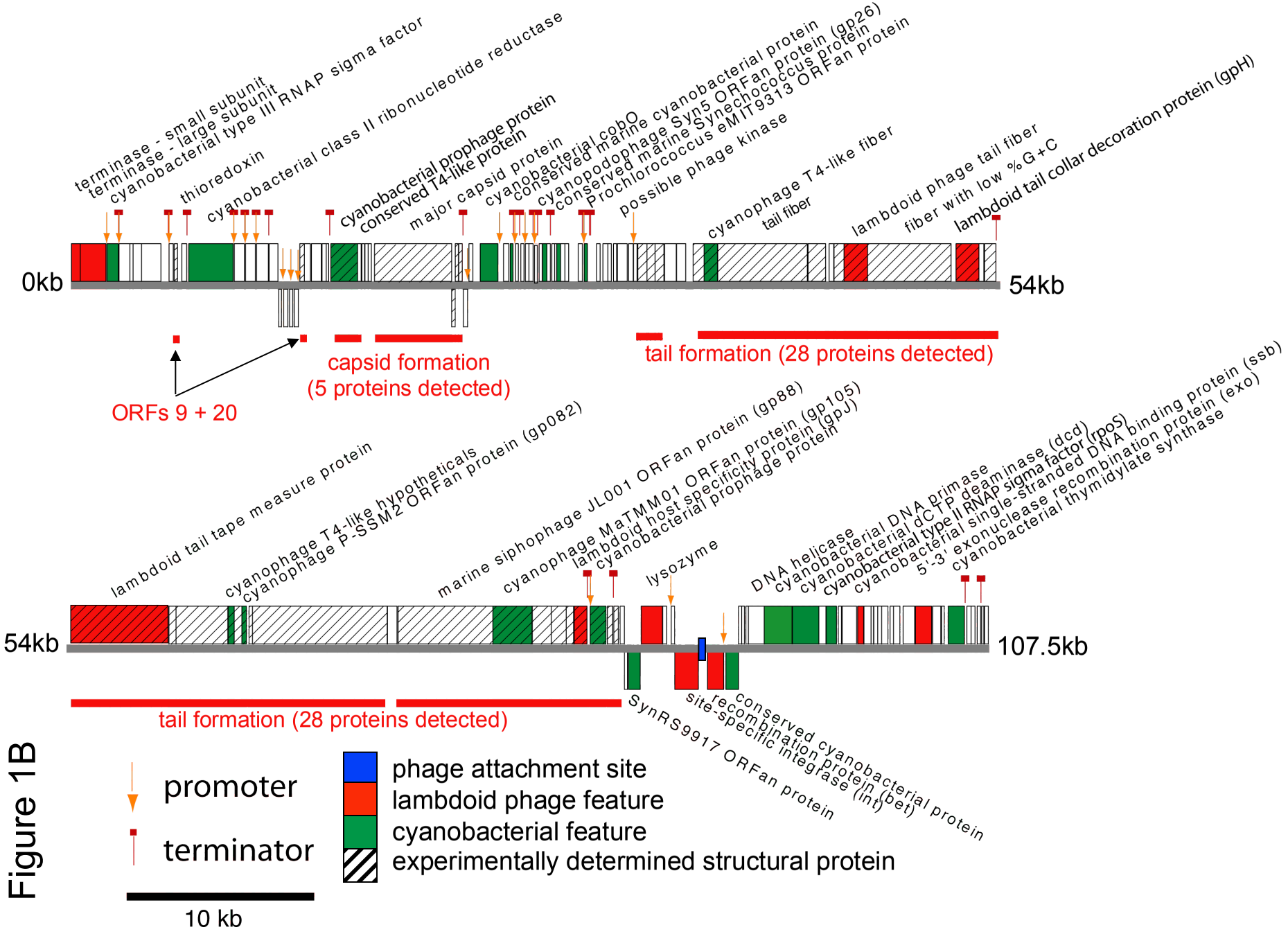

terminase - small subunit
terminase - large subunit
cyanobacterial type III RNAP sigma factor
thioredoxin
cyanobacterial class II ribonucleotide reductase
cyanobacterial prophage protein
conserved T4-like protein
major capsid protein
cyanobacterial cobO
conserved marine cyanobacterial protein
cyanopodophage Syn5 ORFan protein (gp26)
conserved marine Synechococcus protein
Prochlorococcus eMIT9313 ORFan protein
possible phage kinase
cyanophage T4-like fiber
tail fiber
lambdoid phage tail fiber
fiber with low %G+C
lambdoid tail collar decoration protein (gpH)
0kb
54kb
capsid formation
(5 proteins detected)
tail formation (28 proteins detected)
ORFs 9 + 20
lambdoid tail tape measure protein
cyanophage T4-like hypotheticals
cyanophage P-SSM2 ORFan protein (gp082)
marine siphophage JL001 ORFan protein (gp88)
cyanophage MaTMM01 ORFan protein (gp105)
lambdoid host specificity protein (gpJ)
cyanobacterial prophage protein
lysozyme
DNA helicase
cyanobacterial DNA primase
cyanobacterial dCTP deaminase (dcd)
cyanobacterial type II RNAP sigma factor (rpoS)
cyanobacterial single-stranded DNA binding protein (ssb)
5'-3' exonuclease recombination protein (exo)
cyanobacterial thymidylate synthase
54kb
107.5kb
tail formation (28 proteins detected)
SynRS9917 ORFan protein
site-specific integrase (int)
recombination protein (bet)
conserved cyanobacterial protein
promoter
terminator
10 kb
phage attachment site
lambdoid phage feature
cyanobacterial feature
experimentally determined structural protein

Figure 1B

Figure 2

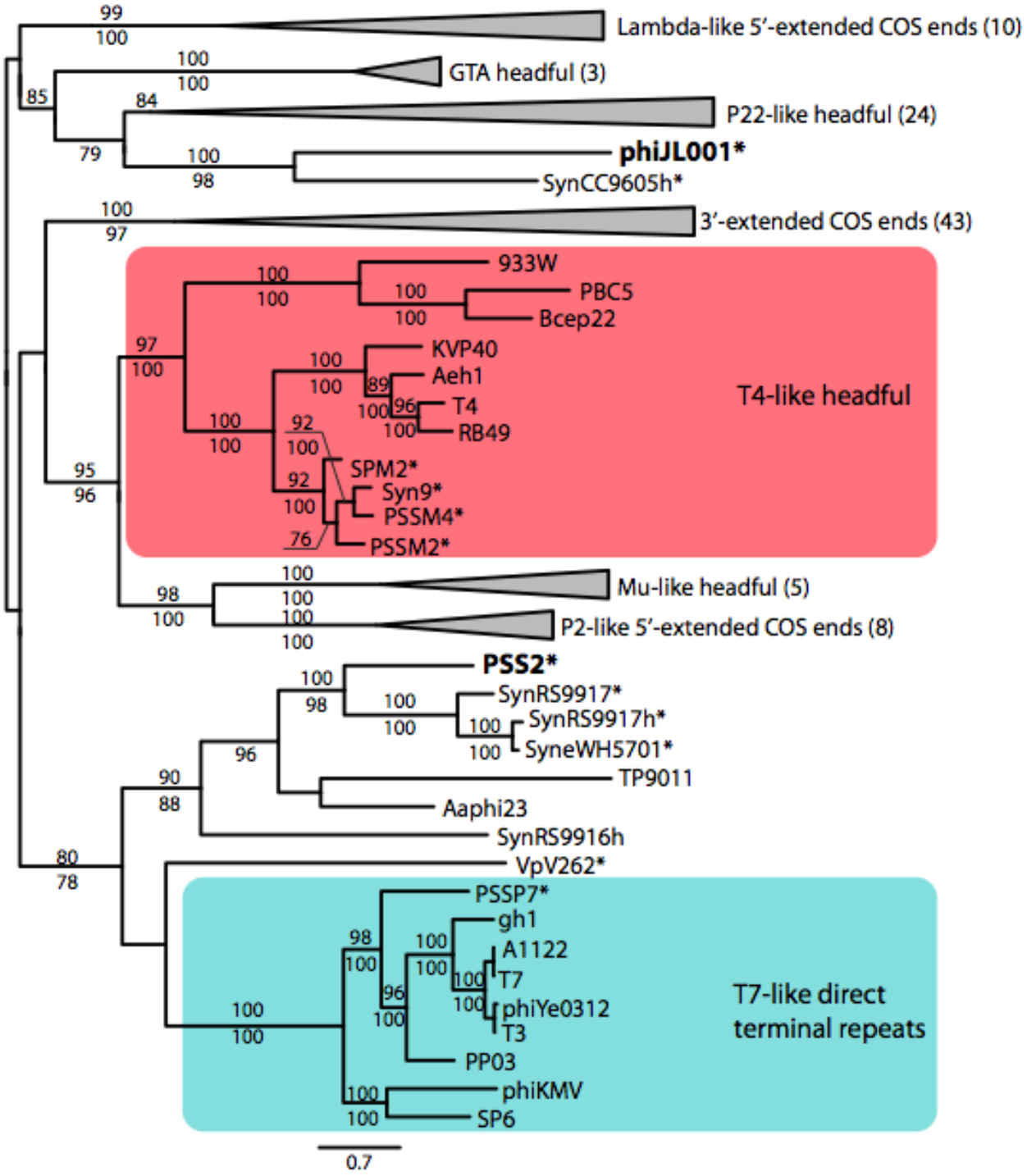
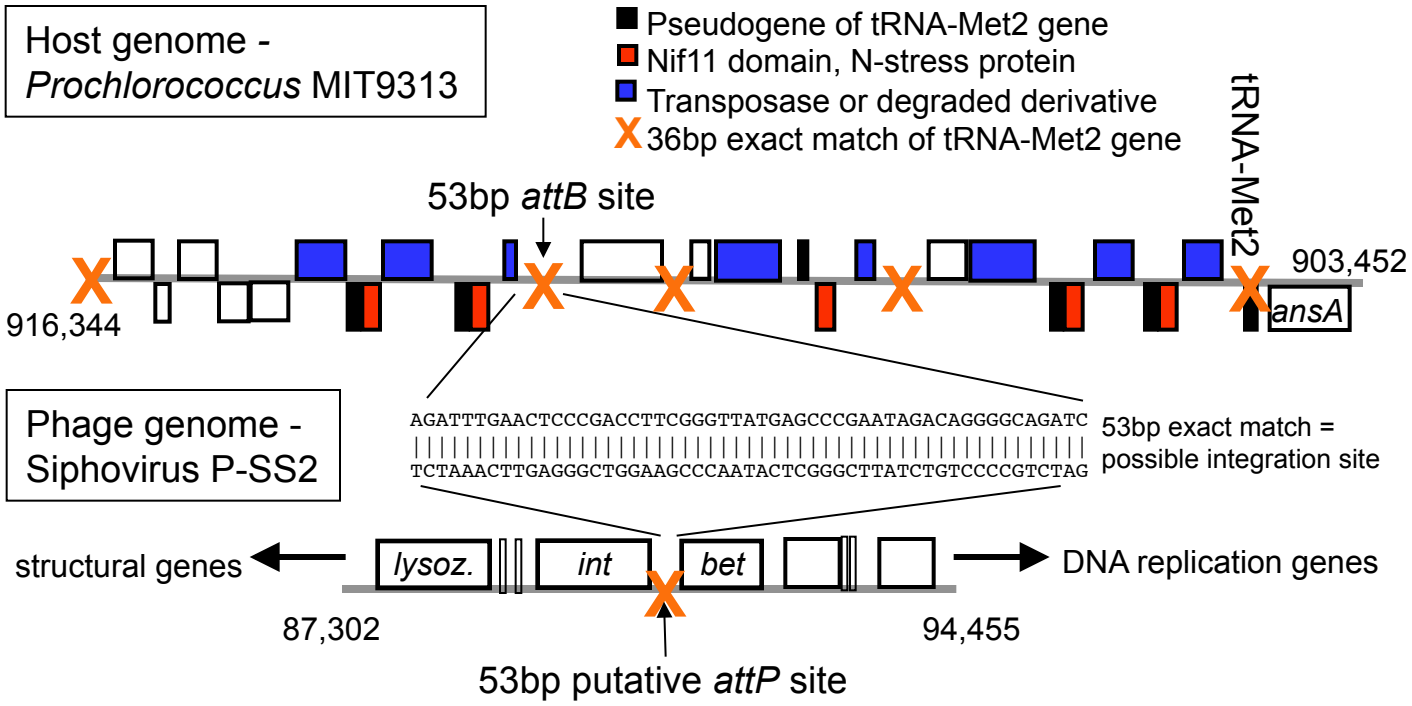

Lambda-like 5'-extended COS ends (10)
GTA headful (3)
P22-like headful (24)
phiJL001*
SynCC9605h*
3'-extended COS ends (43)
933W
PBC5
Bcep22
KVP40
Aeh1
T4
RB49
SPM2*
Syn9*
PSSM4*
PSSM2*
T4-like headful
Mu-like headful (5)
P2-like 5'-extended COS ends (8)
PSS2*
SynRS9917*
SynRS9917h*
SyneWH5701*
TP9011
Aaphi23
SynRS9916h
VpV262*
PSSP7*
gh1
A1122
T7
phiYe0312
T3
PP03
phiKMV
SP6
T7-like direct terminal repeats
0.7

Figure 3


Host genome -
*Prochlorococcus* MIT9313
Pseudogene of tRNA-Met2 gene
Nif11 domain, N-stress protein
Transposase or degraded derivative
36bp exact match of tRNA-Met2 gene
53bp *attB* site
tRNA-Met2
916,344
903,452
*ansA*
Phage genome -
Siphovirus P-SS2
AGATTTGAACTCCCGACCTTCGGGTTATGAGCCCGAATAGACAGGGGCAGATC
TCTAAACTTGAGGGCTGGAAGCCCAATACTCGGGCTTATCTGTCCCCGTCTAG
53bp exact match =
possible integration site
structural genes
*lysoz.*
*int*
*bet*
DNA replication genes
87,302
94,455
53bp putative *attP* site

# Figure 4A

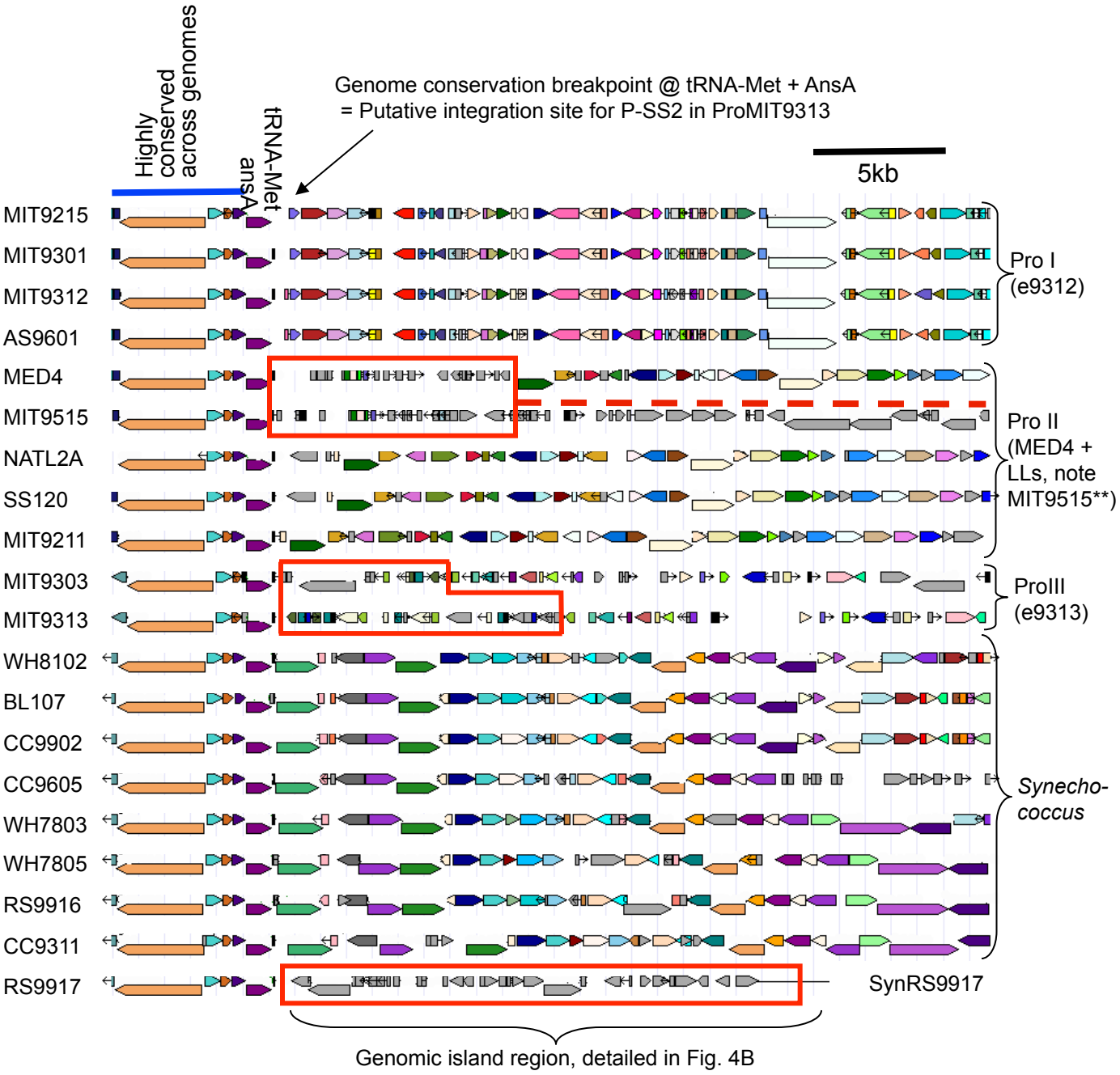
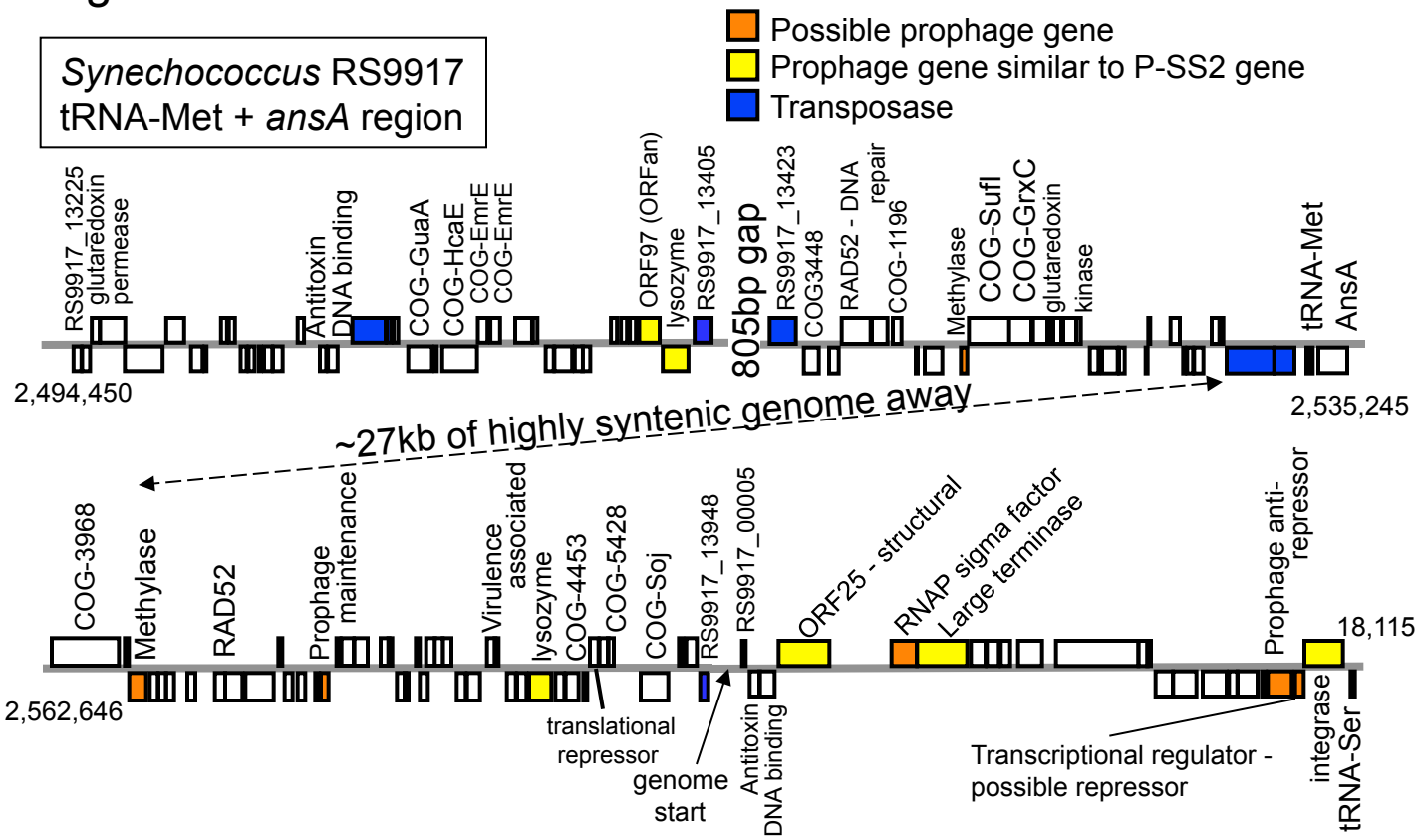

Highly conserved across genomes
tRNA-Met
ansA
Genome conservation breakpoint @ tRNA-Met + AnsA
= Putative integration site for P-SS2 in ProMIT9313
5kb
MIT9215
MIT9301
MIT9312
AS9601
MED4
MIT9515
NATL2A
SS120
MIT9211
MIT9303
MIT9313
WH8102
BL107
CC9902
CC9605
WH7803
WH7805
RS9916
CC9311
RS9917
Pro I (e9312)
Pro II (MED4 + LLs, note MIT9515**)
ProIII (e9313)
Synecho-coccus
SynRS9917
Genomic island region, detailed in Fig. 4B

Figure 4B


*Synechococcus* RS9917
tRNA-Met + *ansA* region
Possible prophage gene
Prophage gene similar to P-SS2 gene
Transposase
RS9917_13225 glutaredoxin permease
Antitoxin DNA binding
COG-GuaA
COG-HcaE
COG-EmrE
COG-EmrE
ORF97 (ORFan)
lysozyme
RS9917_13405
805bp gap
RS9917_13423
COG3448
RAD52 - DNA repair
COG-1196
Methylase
COG-SufI
COG-GrxC
glutaredoxin kinase
tRNA-Met
AnsA
2,494,450
2,535,245
~27kb of highly syntenic genome away
COG-3968
Methylase
RAD52
Prophage maintenance
Virulence associated lysozyme
COG-4453
COG-5428
COG-Soj
RS9917_13948
RS9917_00005
ORF25 - structural
RNAP sigma factor
Large terminase
Prophage anti-repressor
18,115
2,562,646
translational repressor
genome start
Antitoxin DNA binding
Transcriptional regulator - possible repressor
integrase
tRNA-Ser

# Figure 5

*IS element classifications in* Synechococcus *RS9917 genome*

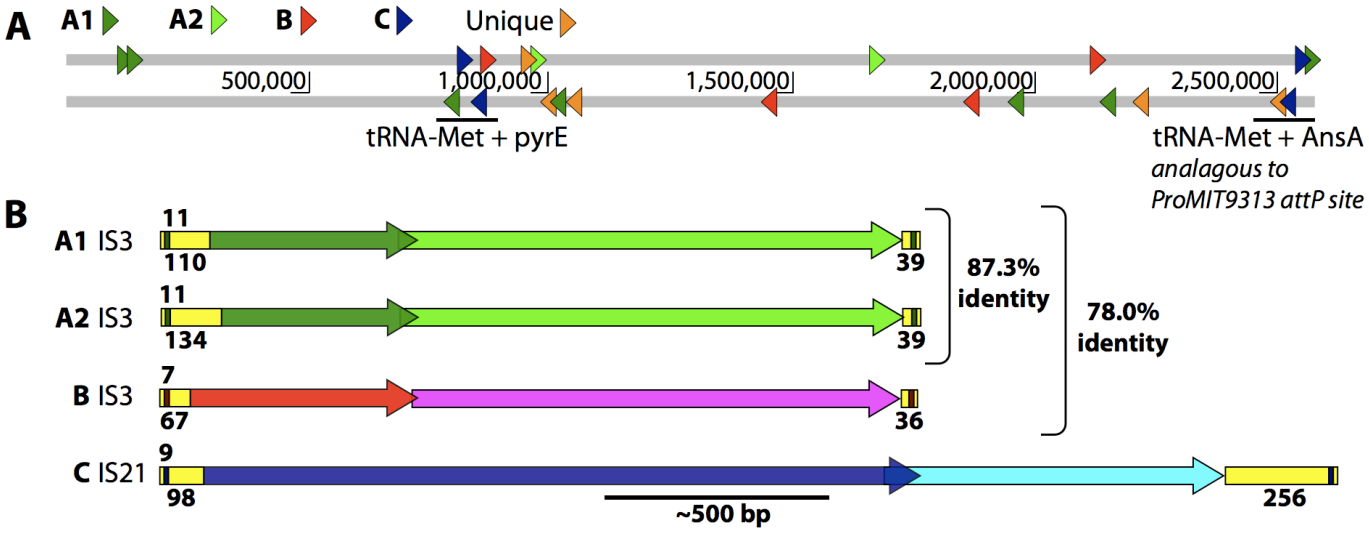


Figure 6A

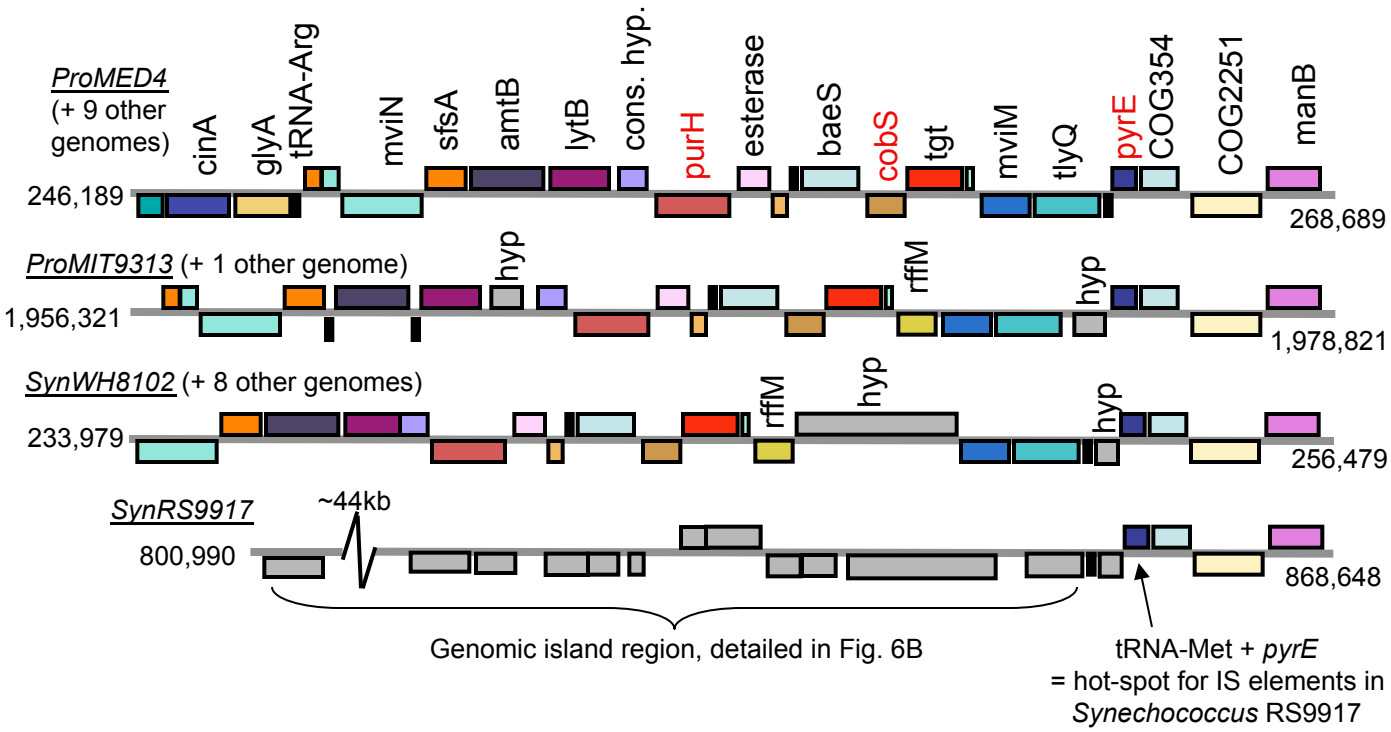
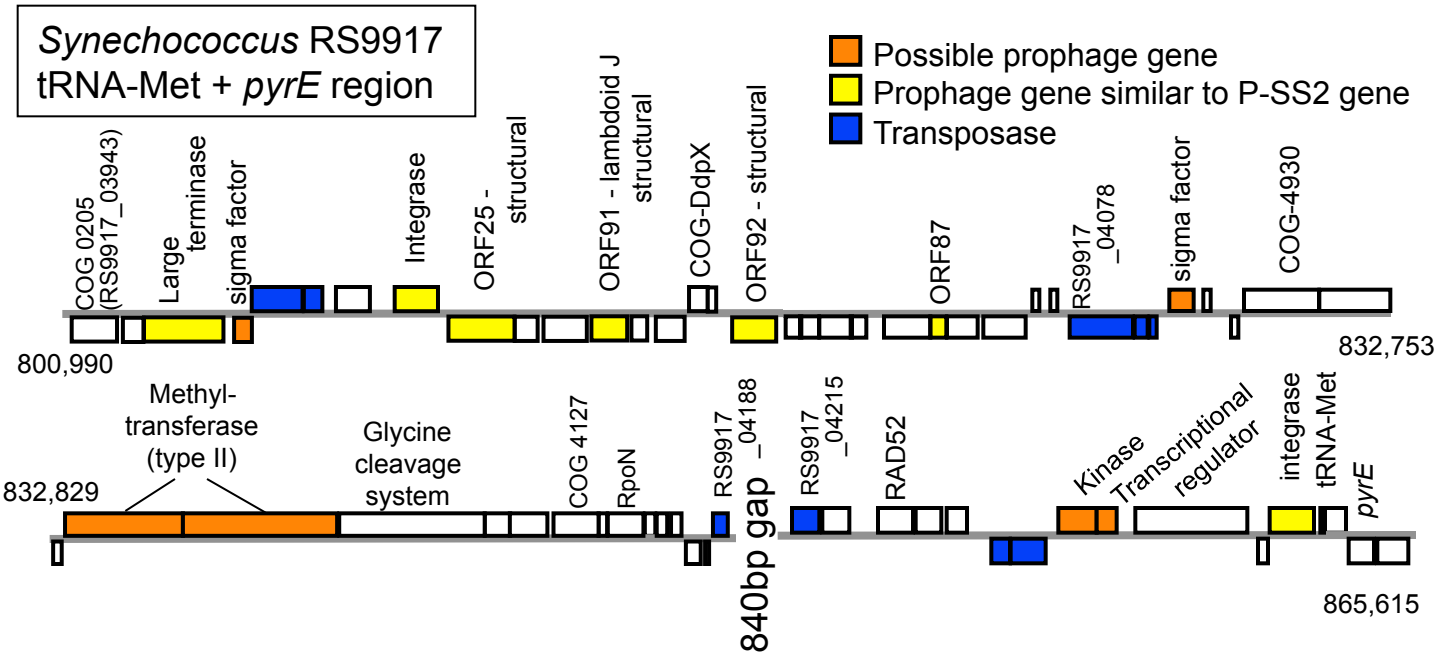
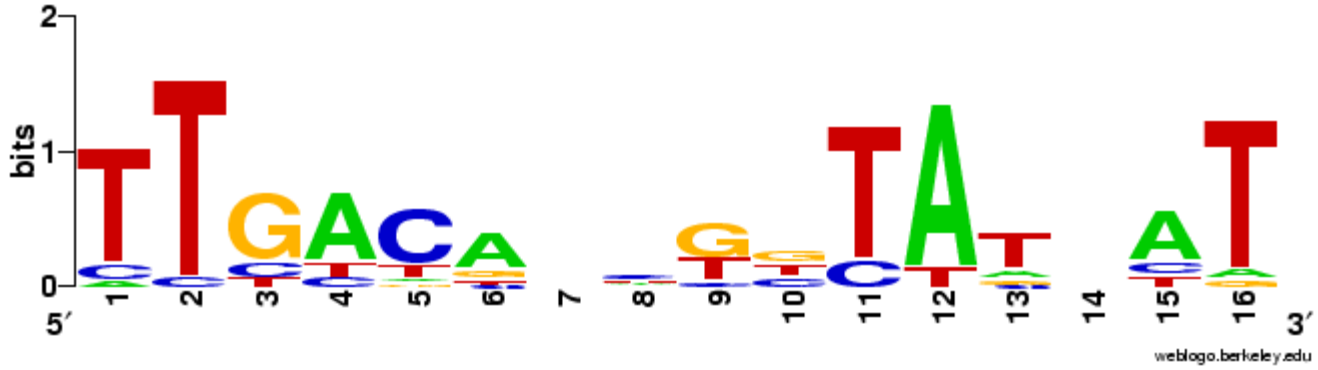

ProMED4
(+ 9 other genomes)
cinA
glyA
tRNA-Arg
mviN
sfsA
amtB
lytB
cons. hyp.
purH
esterase
baeS
cobS
tgt
mviM
tlyQ
pyrE
COG354
COG2251
manB
246,189
268,689
ProMIT9313 (+ 1 other genome)
hyp
rffM
hyp
1,956,321
1,978,821
SynWH8102 (+ 8 other genomes)
rffM
hyp
hyp
233,979
256,479
SynRS9917
~44kb
800,990
868,648
Genomic island region, detailed in Fig. 6B
tRNA-Met + pyrE
= hot-spot for IS elements in
Synechococcus RS9917

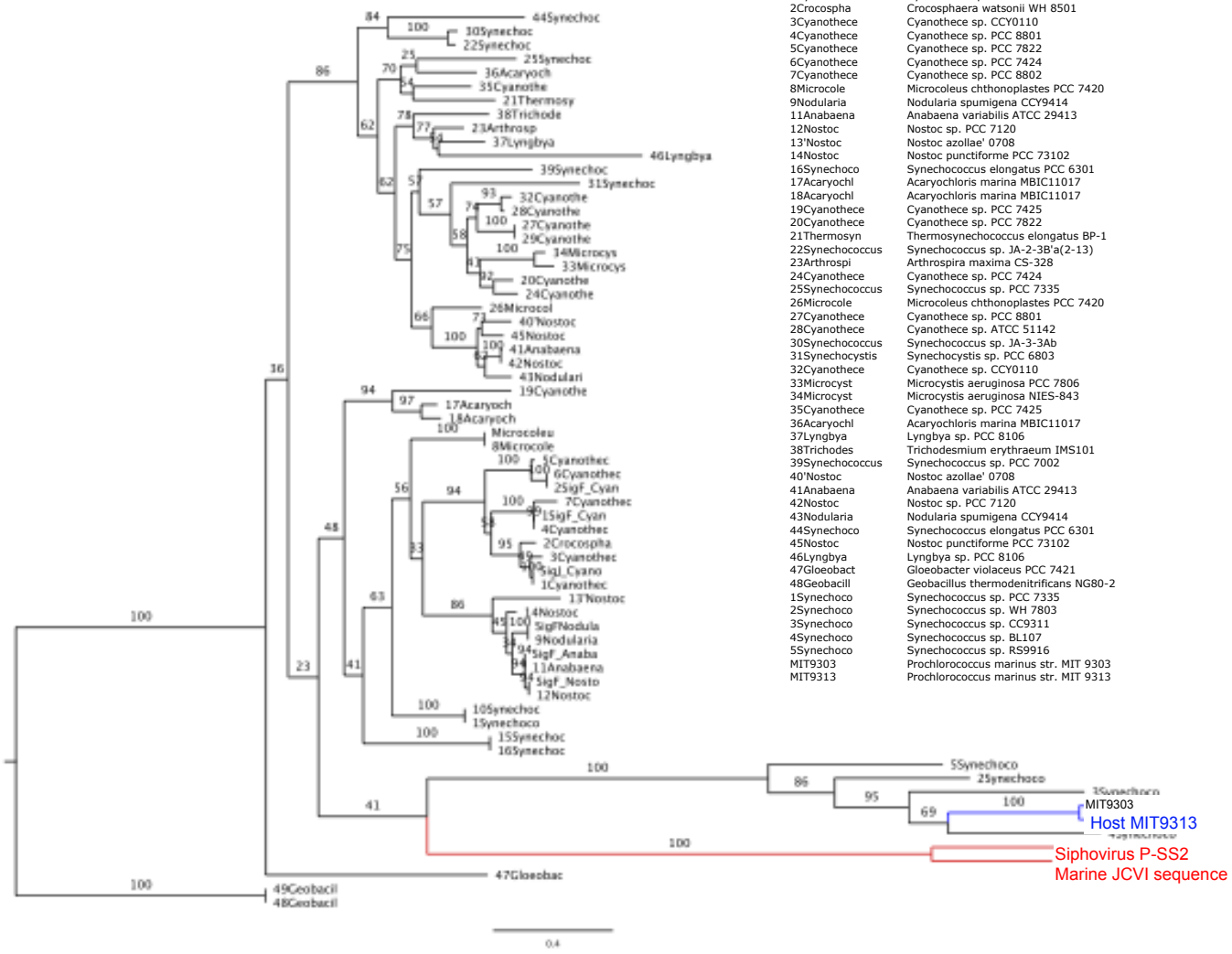
Figure 6B


*Synechococcus* RS9917
tRNA-Met + *pyrE* region
Possible prophage gene
Prophage gene similar to P-SS2 gene
Transposase
COG 0205 (RS9917_03943)
Large terminase
sigma factor
Integrase
ORF25 - structural
ORF91 - lambdoid J structural
COG-DdpX
ORF92 - structural
ORF87
RS9917_04078
sigma factor
COG-4930
800,990
832,753
Methyl-transferase (type II)
Glycine cleavage system
COG 4127
RpoN
RS9917_04188
840bp gap
RS9917_04215
RAD52
Kinase
Transcriptional regulator
integrase
tRNA-Met
*pyrE*
832,829
865,615

Suppl. Fig. 1


bits
2
1
0
5′
3′
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
weblogo.berkeley.edu

# Suppl. Fig. 2




| SeqID | Strain name |
|---|---|
| 1Cyanothece | Cyanothece sp. ATCC 51142 |
| 2Crocospha | Crocosphaera watsonii WH 8501 |
| 3Cyanothece | Cyanothece sp. CCY0110 |
| 4Cyanothece | Cyanothece sp. PCC 8801 |
| 5Cyanothece | Cyanothece sp. PCC 7822 |
| 6Cyanothece | Cyanothece sp. PCC 7424 |
| 7Cyanothece | Cyanothece sp. PCC 8802 |
| 8Microcole | Microcoleus chthonoplastes PCC 7420 |
| 9Nodularia | Nodularia spumigena CCY9414 |
| 11Anabaena | Anabaena variabilis ATCC 29413 |
| 12Nostoc | Nostoc sp. PCC 7120 |
| 13'Nostoc | Nostoc azollae' 0708 |
| 14Nostoc | Nostoc punctiforme PCC 73102 |
| 16Synechoco | Synechococcus elongatus PCC 6301 |
| 17Acaryochl | Acaryochloris marina MBIC11017 |
| 18Acaryochl | Acaryochloris marina MBIC11017 |
| 19Cyanothece | Cyanothece sp. PCC 7425 |
| 20Cyanothece | Cyanothece sp. PCC 7822 |
| 21Thermosyn | Thermosynechococcus elongatus BP-1 |
| 22Synechococcus | Synechococcus sp. JA-2-3B'a(2-13) |
| 23Arthrospi | Arthrospira maxima CS-328 |
| 24Cyanothece | Cyanothece sp. PCC 7424 |
| 25Synechococcus | Synechococcus sp. PCC 7335 |
| 26Microcole | Microcoleus chthonoplastes PCC 7420 |
| 27Cyanothece | Cyanothece sp. PCC 8801 |
| 28Cyanothece | Cyanothece sp. ATCC 51142 |
| 30Synechococcus | Synechococcus sp. JA-3-3Ab |
| 31Synechocystis | Synechocystis sp. PCC 6803 |
| 32Cyanothece | Cyanothece sp. CCY0110 |
| 33Microcyst | Microcystis aeruginosa PCC 7806 |
| 34Microcyst | Microcystis aeruginosa NIES-843 |
| 35Cyanothece | Cyanothece sp. PCC 7425 |
| 36Acaryochl | Acaryochloris marina MBIC11017 |
| 37Lyngbya | Lyngbya sp. PCC 8106 |
| 38Trichodes | Trichodesmium erythraeum IMS101 |
| 39Synechococcus | Synechococcus sp. PCC 7002 |
| 40'Nostoc | Nostoc azollae' 0708 |
| 41Anabaena | Anabaena variabilis ATCC 29413 |
| 42Nostoc | Nostoc sp. PCC 7120 |
| 43Nodularia | Nodularia spumigena CCY9414 |
| 44Synechoco | Synechococcus elongatus PCC 6301 |
| 45Nostoc | Nostoc punctiforme PCC 73102 |
| 46Lyngbya | Lyngbya sp. PCC 8106 |
| 47Gloeobact | Gloeobacter violaceus PCC 7421 |
| 48Geobacill | Geobacillus thermodenitrificans NG80-2 |
| 1Synechoco | Synechococcus sp. PCC 7335 |
| 2Synechoco | Synechococcus sp. WH 7803 |
| 3Synechoco | Synechococcus sp. CC9311 |
| 4Synechoco | Synechococcus sp. BL107 |
| 5Synechoco | Synechococcus sp. RS9916 |
| MIT9303 | Prochlorococcus marinus str. MIT 9303 |
| MIT9313 | Prochlorococcus marinus str. MIT 9313 |

# Suppl. Fig. 3