Subject Category: Evolutionary Genetics

An antisense RNA in a lytic cyanophage links *psbA* with a gene encoding a homing endonuclease

A. D. Millard[1]*, G. Gierga[2], M. R. J. Clokie[3], D. J. Evans[1], W. R. Hess[2], D. J. Scanlan[1]

[1] Department of Biological Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom

[2] Institute of Biology III, University of Freiburg, Schänzlestraße 1, D-79104 Freiburg, Germany

[3] Department of Infection, Immunity and Inflammation, Medical Sciences Building, University of Leicester, Leicester, UK

* Corresponding author

Mailing address:

Department of Biological Sciences
University of Warwick, Gibbet Hill Road,
Coventry, CV4 7AL, United Kingdom
Telephone: +44 (0)24 76 522572
Fax: +44 (0)24 76 523701
Email: a.d.millard@warwick.ac.uk

Running title: *psbA* antisense RNA

To be submitted to ISME

Abstract

Cyanophage genomes frequently possess the *psbA* gene, encoding the D1 polypeptide of photosystem II. This protein is thought to maintain host photosynthetic capacity during infection and enhance phage fitness under high light conditions. Whilst the first documented cyanophage-encoded *psbA* gene contained a group I intron, this feature has not been widely reported since, despite a plethora of new sequences becoming available. Here, we show that in cyanophage S-PM2 this intron is spliced during the entire infection cycle. Furthermore, we report the widespread occurrence of *psbA* introns in marine metagenomic libraries, and with *psbA* often adjacent to a homing endonuclease. Bioinformatic analysis of the intergenic region between *psbA* and the adjacent homing endonuclease gene F-CphI in S-PM2 revealed the presence of an antisense RNA (asRNA) connecting these two separate genetic elements. The asRNA is co-regulated with *psbA* and F-CphI, suggesting its involvement with their expression. Analysis of scaffolds from GOS datasets shows this asRNA to be commonly associated with the 3′ end of cyanophage *psbA* genes, implying that this potential mechanism of regulating marine 'viral' photosynthesis is evolutionarily conserved. While antisense transcription is commonly found in eukaryotic and increasingly also in prokaryotic organisms, there has been no indication for asRNAs in lytic phages so far. We propose this asRNA also provides a means of preventing the formation of mobile group I introns within cyanophage *psbA* genes.

Keywords: asRNA/cyanophage/endonuclease/intron/*psbA*

Introduction

Viruses are the most abundant biological entities in the oceans, with numbers estimated to be over 10^{30} (Suttle, 2005). As important agents of microbial mortality they play critical roles in nutrient cycling and structuring microbial communities, whilst also contributing to horizontal gene transfer by mediating genetic exchange (Suttle, 2005; Suttle, 2007). Bacteriophages infecting the picocyanobacterial genera *Synechococcus* (Waterbury and Valois, 1993; Suttle and Chan, 1994; Lu *et al*., 2001; Marston and Sallee, 2003; Millard and Mann, 2006; Marston and Amrich, 2009) and *Prochlorococcus* (Sullivan *et al*., 2003) are some of the most well characterised marine viruses. Such cyanophages are widely distributed and abundant (>10^5 ml^{-1} (Suttle and Chan, 1994), with most isolates belonging to the family myoviridae (Waterbury and Valois, 1993; Suttle and Chan, 1994; Lu *et al*., 2001; Sullivan *et al*., 2003; Millard and Mann, 2006; Millard *et al*., 2009) and fewer representatives thus far known from the siphoviridae (Waterbury and Valois, 1993; Sullivan *et al*., 2009) and podoviridae (Waterbury and Valois, 1993; Suttle and Chan, 1994; Sullivan *et al*., 2003) families.

Whilst cyanophages, like several other viruses, can divert the flow of carbon through the microbial loop, they are unique in being thought to be able to directly contribute to the photosynthetic process via their possession of phage versions of the core photosystem II reaction centre polypeptides D1 and D2, encoded by *psbA* and *psbD*, respectively. Recent research from metagenomic data has shown that cyanophages also carry genes encoding complete subunits of photosystem I (Sharon *et al*., 2009). However, to date most cyanophage research has focused on PSII. During photosynthesis D1 is continually being turned over and replaced by newly synthesised D1. It is postulated that expression of the phage-encoded D1 protein provides a means to maintain photosynthesis even after host protein synthesis in infected cells is diminished, thus ensuring a source of energy for virus production (Mann *et al*., 2003; Lindell *et al*., 2004; Millard *et al*., 2004; Lindell *et al*., 2005; Clokie *et al*., 2006; Lindell *et al*., 2007). This is supported both by evidence that cyanophage *psbA* transcripts can be detected throughout the infection cycle (Lindell *et al*., 2005; Clokie *et al.*, 2006; Lindell *et al.*, 2007) and by the fact that the cyanophage D1 polypeptide increases in abundance during infection (Lindell *et al.*, 2007). Moreover, recent modelling studies suggest there is an increased fitness advantage for cyanophages possessing *psbA* particularly under high-light conditions (Bragg and Chisholm, 2008; Hellweger, 2009).

Genome sequencing (Millard *et al.*, 2004; Mann *et al.*, 2005; Sullivan *et al.*, 2005; Weigele *et al.*, 2007; Millard *et al.*, 2009) and PCR screening (Sullivan *et al.*, 2006; Wang and Chen, 2008; Marston and Amrich, 2009) efforts indicate that *psbA* is widely distributed in cyanophage isolates, whilst cyanophage-derived *psbA* transcripts can also be readily detected in the marine environment (Zeidner *et al.*, 2005; Sullivan *et al.*, 2006; Sharon *et al.*, 2007; Chenard and Suttle, 2008). Phylogenetic analysis of phage *psbA* suggests that it has been inherited from its cyanobacterial hosts on a number of occasions (Lindell *et al.*, 2004; Millard *et al.*, 2004; Zeidner *et al.*, 2005; Sullivan *et al.*, 2006), but with evidence of significant intragenic recombination between the phage and host gene (Zeidner *et al.*, 2005; Sullivan *et al.*, 2006).

An unusual feature of the first viral *psbA* gene discovered, in the cyanophage S-PM2, was that it contained a group I intron (Millard *et al.*, 2004). Whilst group I introns are common in other bacteriophage genomes, the sequencing of hundreds of other cyanophage (Sullivan *et al.*, 2006; Chenard and Suttle, 2008; Marston and Amrich, 2009) and host (Zeidner *et al.*, 2003; Sharon *et al.*, 2007) *psbA* genes has revealed only one more containing an intron (Millard *et al.*, 2004). This is likely due to the fact that the widely used reverse *psbA* PCR primer (Zeidner *et al.*, 2003) does not amplify *psbA* that contains an intron in the same position as found in S-PM2, thereby preventing detection of *psbA* genes with introns in the same position.

The origin of the *psbA* intron in S-PM2 is still unknown. Whilst introns are present in some *psbA* genes of chloroplasts (Maul *et al.*, 2002; Brouard *et al.*, 2008), they have thus far not been found in any of the cyanobacterial orthologs. The mobility of the *psbA* intron has previously been proposed to be mediated by an endonuclease in a process known as “homing” which would transfer the intron and flanking DNA containing the endonuclease into an intron-less allele of *psbA* (Millard *et al.*, 2004). The recent characterisation of a homing endonuclease situated immediately downstream of *psbA* in S-PM2, that is only able to cut intron-less copies of *psbA* supports this idea (Zeng *et al.*, 2009). Homing endonucleases are generally thought of as selfish elements that target highly specific DNA target sites of 14-40 bp in length (Jurica and Stoddard, 1999) and allow transfer of themselves, and the introns in which they often reside, to cognate sites within a population (Jurica and Stoddard, 1999). Paradoxically they can tolerate sequence variation within their target site, allowing targeting of new hosts (Jurica and Stoddard, 1999). Whilst often found within introns this is not always the case, with "intron-less homing" observed between the bacteriophages T4 and T2 (Liu *et al.*, 2003).

Despite knowledge that S-PM2 *psbA* is expressed during the lytic cycle (Clokie *et al.*, 2006) it is not known whether the intron is spliced *in vivo*, how widespread these introns are in other cyanophage genomes, or how they might have been acquired.

Another intriguing type of RNA molecule, that was discovered first in bacteriophages almost 40 years ago, are antisense RNAs (asRNAs). Such naturally occurring asRNAs were postulated first in bacteriophage λ (Spiegelman *et al.*, 1972), and only afterwards observed in bacteria (Itoh and Tomizawa, 1980; Lacatena and Cesareni, 1981) and even later in eukaryotes. More recently, it was found that expression of the photosynthetic gene *isiA* in the cyanobacterium *Synechocystis* sp. PCC6803 is regulated by the 177 nt long asRNA IsiR (Duhring *et al.*, 2006). However, asRNAs have not been described for any cyanophage gene thus far.

Materials and Methods

Culturing

Synechococcus sp. WH7803 was cultured in ASW medium (Wyman *et al.*, 1985) in 100 ml batch cultures in 250 ml conical flasks under constant illumination (5–36 μmol photons m^{-2} s^{-1}) at 25˚C. Larger volumes were grown in 0.5 l vessels under constant shaking (150 rpm). Cyanophage S-PM2 stocks and phage titre were produced as reported previously (Wilson *et al.*, 1993).

Host infection

The S-PM2 infection cycle has previously been well characterised with lysis of *Synechococcus* proceeding 9 hr post infection (Wilson *et al.*, 1996; Clokie *et al.*, 2006) .Therefore, 50 ml samples were taken prior to infection and then at 1, 3, 6, and 9 hr post infection. Phage was added at an MOI of ~5, to ensure infection of all cells. Samples were immediately centrifuged at 8000 g to pellet the samples, which were then snap frozen in 0.5 ml of RNA extraction buffer (10 mM NaAc, pH 4.5; 200 mM sucrose, 5mM EDTA) and stored at -80˚C until samples were further processed. Three biological replicates were taken.

RNA extraction

Total RNA was extracted based on a previously described method (Logemann *et al.*, 1987). Briefly, frozen samples were gently thawed in 3 vols of Z buffer (8M guanidinium hydrochloride; 50 mM β-mercaptoethanol; 20 mM EDTA) at room temperature for 30 min. Samples were extracted with the addition of ½ vol of phenol (pH 4.5) at 65°C for 30 min, followed by the addition of chloroform:isoamyl alcohol for 15 min. RNA was precipitated in 1 vol of isopropanol, followed by a wash in 70% (v/v) ethanol. RNA was treated with Turbo DNase I (Ambion) for 2 hr at 37°C, extracted with phenol:cholorform:isoamyl alcohol (25:24:1), re-precipitated with 3M NaAc and tested for DNA contamination using PCR primers gp23F/R.

***In vivo* splicing**

RNA was extracted and cDNA synthesized. cDNA synthesis was carried out in 20 µl reaction volumes with a 600 ng of total RNA. Each reaction contained 1 µl of 20x dNTP mix (10 mM dGTP, dCTP, dATP and dTTP), 5 µg random hexamers (VHBio, Gateshead, UK) or 2 pmole of gene specific primer, 4 µl of 5x buffer (250 mM Tris pH 8.3, 375 mM KCl, 15 mM mgCl$_2$), 2 µl of 0.1 M DTT, 200 units Superscript III™ (Invitrogen, Paisley, UK) and water to a final volume of 20 µl. The RNA, water and random hexamers were mixed, heated to 65˚C for 10 min and cooled to 4˚C in a thermal cycler, prior to the addition of 5 x buffer, DTT and superscript, heated to 50˚C for 50 min, before finally heating to 70˚C for 10 min.

The primers psbA_F and psbA_R were used to amplify PCR products of 1080 and 1291 bp in length dependent on whether splicing had occurred. PCR was carried out with 0.03 U/ml Vent DNA polymerase in 1 x buffer (20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$), 2 mM $MgCl_2$, 0.2 mM dNTPs, and 40 pM of each primer. PCR cycling conditions of 35 cycles of 94°C for 10 s, 55°C for 15 s, and 72°C for 20 sec and s final incubation at 72°C for 2 min. PCR products were sequenced in house using an ABI 3730 automated sequencer (Applied Biosystems).

Quantitative RT-PCR (qPCR)

PCR primers were designed using Primer Express (ABI, Warrington, UK). The PCR primers for the remaining genes are reported in Table S1. A variety of primer concentrations were tested and optimised to ensure amplification efficiency was within the required limits to implement relative quantification using $2^{\Delta\Delta}$CT (Livak and Schmittgen, 2001). The amplification efficiencies for target and reference primers sets were tested by ensuring the slope of the line was <0.1 when log input DNA concentration was plotted against $^{\Delta}$CT.

A no reverse transcriptase control PCR reaction was used to assess DNA contamination of experimental samples. Any sample found to be contaminated was subject to further DNAse treatment (see RNA extraction section above) and the process repeated until the control PCR reaction proved negative. cDNA synthesis was then carried out as described above, with the gene specific primer ncRNA_R used for synthesis of cDNA from the ncRNA CfrI.

PCR reactions of 1x power SYBR green mix (ABI), 150 µM forward primer, 150 µM reverse primer and 10 ng cDNA were used for amplification of 16S rRNA, *psbA* and ORF 177 (F-CphI), whilst for psbA_ncRNA 50 ng of cDNA was used per well. Thermal cycling was carried out in a 7500 sequence detector (Applied Biosystems) with an initial step of 95˚C for 10 min followed by 40 cycles of 95˚C for 30 s followed by 62˚C for 1 min and a final dissociation step. The fold change of each gene was determined using the $2^{\Delta\Delta CT}$ method (Livak and Schmittgen, 2001) using 16S rRNA as the calibrator. Absolute transcript abundance was calculated from a standard curve, whilst a dilution series of purified phage DNA was used to construct the curve.

ncRNA prediction

Sequence-dependent RNA structure within the phage genome and scaffold sequences was identified by comparing the folding free energy of the native sequence with a large number of sequence order randomised controls. In practice the scaffolds was divided into 200 nucleotide fragments for both strands of DNA, overlapping by 190 bp, each of which was randomised 999 times using a method (designated NDR; implemented in the Simmonics suite of sequence analysis programs (Simmonds and Smith, 1999), available from http://www.picornavirus.org) which retained both the nucleotide and dinucleotide composition. Sequences were stored in a MySQL database and the folding free energy for each was determined using hybrid-ss-min from the Unafold (http://dinamelt.bioinfo.rpi.edu/) suite of programs (Markham and Zuker, 2008), automated using perl scripts. For each fragment the mean folding energy difference (MFED), expressed as the percentage difference between native and the mean of the randomised sequences from the same fragment, was determined. In addition, the position of the native sequence in the distribution of energies of the randomised fragments - expressed as the Nth percentile - was calculated.

5′ Rapid Amplification of cDNA Ends (RACE)

5′ RACE experiments were conducted based on the protocol of Steglich *et al.* (2008). Briefly, RNA was treated with tobacco acid pyrophosphorylase (1 U/1 µg RNA; Epicentre, USA) for 1 h at 37°C, followed by phenol/chloroform extraction and ethanol precipitation. A synthetic RNA oligonucleotide (0.5 µl oligonucleotide [10 mM]/ 4 µg RNA; 5′-AUAUGCGCGAAUUCCUGUAGAACGAACACUAGAAGAAA-3′, Invitrogen, Germany) was ligated to RNA using T4 RNA ligase (3 U/1 µg RNA; Fermentas, Germany) for 1 h at 37 °C, followed by phenol/chloroform extraction and ethanol precipitation. Three control reactions were performed: i) omitting tobacco acid pyrophosphorylase, ii) omitting tobacco acid pyrophosphorylase and RNA oligonucleotide and iii) dephosphorylating RNA prior to ligation with calf intestine alkaline phosphatase (0.1 U/1 µg RNA; Fermentas, Germany) at 37°C for 1 h, followed by phenol/chloroform extraction and ethanol precipitation. For reverse transcription, 250 ng oligonucleotide-linked RNA per gene was incubated with 0.8 U Omniscript reverse transcriptase (Qiagen, Germany) in the provided reaction buffer, supplemented by 0.08 µM gene specific primer and 1 mM dNTPs. Incubation was carried out at 42°C for 2 h with a final inactivation step at 95°C for 5 min. All reactions were performed in the presence of 40 U Ribolock RNase Inhibitor (Fermentas, Germany). cDNA was amplified by PCR in GoTaq reaction buffer containing 1 U GoTaq polymerase (Promega, Germany), 0.2 mM dNTPs, 3.5 mM $MgCl_{2,}$ a gene-specific primer (0.2 µM), and an RNA oligonucleotide-specific primer (0.2 µM) with the following parameters: 93°C/3 min; 35 cycles of 93°C/30 s; 50°C/30 s, 72°C/45 s; followed by 72°C/5 min. Amplified PCR fragments were gel-excised and purified on Nucleospin columns (Macherey & Nagel, Germany) and then cloned into plasmid pGEM-T (Promega, Germany). After transformation into *E. coli* XL1-Blue, plasmid inserts were amplified by colony PCR, purified on Nucleospin columns and sequenced using an ABI 3130XL automatic DNA sequencer (Applied Biosystems, USA).

Northern blot analysis

RNA samples (20 µg) were denatured for 5 min at 65°C in loading buffer (Fermentas, Germany), separated on 10% (w/v) urea polyacrylamide gels at 90 V for 16 h and transferred to Hybond-N^+ nylon membranes (Amersham, Germany) by electroblotting for 1 h at 400 mA. The membranes were hybridized with single-stranded [α-32P]UTP-labelled transcripts. Following pre-hybridization in 50% (v/v) deionized formamide, 7% (w/v) SDS, 250 mM NaCl and 120 mM $Na(PO_4)$ pH 7.2 for 2 h, hybridization was performed at 62°C overnight in the same buffer. The membranes were washed in 2x SSC (3 M NaCl, 0.3 M sodium citrate, pH 7.0), 1% (w/v) SDS for 10 minutes; 1x SSC, 0.5 % (w/v) SDS for 10 min; and briefly in 0.1x SSC, 0.1% (w/v) SDS. All wash steps were performed 5°C below the hybridization temperature. Signals were detected and analyzed on a Personal Molecular Imager FX system with Quantity One software (BIO-RAD, Germany).

Bioinformatic analyses

Introns were initially identified in the GOS dataset using tblastx with the intron sequence of S-PM2 as the query sequence with an e value cut-off of $<10^{-3}$. Intron insertion sites were determined manually by identifying the point where the translated *psbA* sequences resulted in a premature stop codon or did not align with other highly conserved PsbA sequences. The intron sequence was then extracted. Using a custom perl script full length scaffolds were blasted against a custom BLAST database containing uniprot 100 and all publically accessible cyanobacterial and viral genomes (as available in October 2008) in order to identify any other genes on the scaffolds. Again, a cut-off value of $<10^{-3}$ was used to identify genes fragments. The same approach was used to identify homologues of F-CphI and genes adjacent to it.

Results

Confirmation of intron splicing *in vivo*

To determine if the intron found within the S-PM2 *psbA* is spliced *in vivo* during infection, non-quantitative reverse transcriptase-PCR was used to amplify the *psbA* transcript from RNA extracted at 1, 3 and 9 hr post-infection. Two PCR products were observed, one of 1291 bp and the other 1080 bp in length (Figure 1). The 1291 bp product corresponds to the un-spliced transcript (pre-mRNA) and the smaller product to the spliced transcript. Comparison of the sequence of the smaller product with the *psbA* gene sequence demonstrated that splicing occurred between codons 334 and 335 as previously predicted (Millard *et al.*, 2004).

Searching metagenomic data for intron sequences

Following confirmation of splicing of the *psbA* intron, we searched metagenomic libraries for the presence of other introns similar to that found in S-PM2. The global ocean survey (GOS) dataset (Rusch *et al.*, 2007) was searched using BLAST available through CAMERA (Seshadri *et al.*, 2007). 16 scaffold sequences were identified as having introns that were similar to that of S-PM2 based on sequence identity (Table 1). All intron sequences were localised within *psbA* genes and varied in length from 212-818 nt. With the exception of JCVI_SCAF_10096627024160 the introns were not found to contain ORFs. Six different intron insertion sites (IIS) were found, located throughout the length of the *psbA* gene. The most common IIS found in the 16 scaffolds (8/16) was located after codon 334 (Figure 2) which is the same position as reported in S-PM2 and S-RSM88 (Millard *et al.*, 2004). A single intron is positioned nearby, after codon 338 (Figure 2). 8/16 of these introns were very similar in sequence to that of S-PM2 with percentage nucleotide identities ranging from 62-92%. Additionally, there was conservation of the conserved paired helices (Figure S1).

Intriguingly three introns had IIS after codon 60, which is the same IIS as *psbA* intron 1 in the chloroplast genome of *Oedogonium cardiacum* (Brouard *et al.*, 2008) and *Chlamydomonas reinhardtii* (Maul *et al.*, 2002), with a further two introns located after codon 252, which is very close to the IIS of intron 4 in *C. reinhardtii* which is inserted after codon 254 (Figure 2). The remaining introns were localised in IIS sites that have not been previously documented in *psbA* genes. The introns with an IIS close to or matching that of *C. reinhardtii* and *O. cardiacum* were substantially smaller than the introns found in these two chloroplast genomes. However, they did retain regions of sequence conservation at the 5′ end of the intron, with 52-59% nucleotide identity to the chloroplast introns (Figure S2). Although these introns were detected based on their sequence similarity to S-PM2 there is a small possibility that some chloroplast introns are present in the GOS dataset. However, most would be excluded by the <0.8 μm pore sized filter used in the collection of GOS samples, that would preclude collection of chloroplast containing eukaryotes.

To ascertain the origin of these intron-containing *psbA* genes we searched for the cyanophage/cyanobacterial-specific PsbA motif R/KETTXXXSQ/H (Sharon *et al.*, 2007). This motif was found in all *psbA* sequences (Table 1), whenever the sequence fragment was long enough to encompass this region, suggesting the identified *psbA* genes are all of cyanobacterial or cyanophage origin and not from chloroplasts.

Phylogenetic analysis of *psbA* sequences

To further confirm the origin of the identified *psbA* sequences phylogenetic analysis was carried out. Only sequences that were greater than 920 bp in length were used as this encompassed all regions where introns have been found to be inserted (the intron sequence itself was not included in the analysis). Sequences shorter than this were excluded from the analysis, along with *psbA* fragments used in previous phylogenetic analyses (Sullivan *et al.*, 2006; Chenard and Suttle, 2008), that did not encompass the most common IIS due to the PCR primers used.

Phylogenetic analysis of *psbA* genes was essentially congruent with 16S rRNA phylogenies with eukaryotic algae clearly separate from cyanobacteria. Discrete clades of both high-light (HL)-adapted and low-light (LL)-adapted *Prochlorococcus* strains were discernible and were, in turn, distinct from *Synechococcus* (Figure 3). Phage isolates infecting *Prochlorococcus* formed a sister group to HL-adapted *Prochlorococcus* strains as has been previously reported (Sullivan *et al.*, 2006). Phage isolates infecting *Synechococcus* formed a clade distinct from their *Synechococcus* hosts, whilst the intron-containing *psbA* sequences fell into two discrete clades that did not contain any cultured *Synechococcus* or *Prochlorococcus* strains, or cyanophage isolates (Figure 3). Clearly, the identified *psbA* genes that contain introns are cyanophage/cyanobacterial in origin as they do not group with eukaryotic algae. These newly identified *psbA* sequences fell into two clades, which are sister groups but clearly separated from the well defined *Synechococcus* host clade. Their closer phylogenetic proximity to *psbA* genes from phage isolates infecting *Synechococcus* suggests these sequences are of *Synechococcus* phage origin and not from their *Synechococcus* hosts (Figure 3). This is further supported by examination of both the average mol %GC content and 3rd codon mol %GC content, with the newly identified *psbA* sequences possessing values that are markedly different from the *Synechococcus* host and much closer to that observed in known *Synechococcus* phages.

From phylogenetic analysis and/or detection of the cyanobacterial-specific *psbA* motif it was possible to confirm that 7/16 introns were inserted into genes of cyanophage/cyanobacterial origin, with JCVI_SCAF_1101667044432 containing an intron inserted in the same IIS as found in the chloroplasts of algae and JCVI_SCAF_1096627024160 an intron at a unique site (Figure 2). However, for 9 scaffolds the origin of the *psbA* sequence could not be determined unequivocally as the *psbA* fragment was not long enough for phylogenetic analysis or its length did not extend to the region where the cyanobacterial-specific motif is located (Table 1). Where possible, all scaffolds were examined in further detail to identify the origin of those genes adjacent to *psbA* if any were present. JCVI_SCAF_1097207205912 and JCVI_SCAF_1096626190549 both have genes encoding homologues of the cyanophage protein F-CphI, thus suggesting these are also phage encoded copies of *psbA* (Figure S3). JCVI_SCAF_1096627024703 contains *talC* and *gnd* genes that, although found in *Synechococcus* and *Prochlorococcus* host genomes, are also known to be widespread in cyanophage genomes (Millard *et al.*, 2009). Indeed, these genes have highest sequence similarity to cyanophage encoded versions of these genes, and phylogenetic analysis confirms they are of cyanophage origin (Figure S5 and S6). Again, this suggests the associated scaffolds are also of cyanophage origin (Figure S3). Unfortunately, for the remaining scaffolds it was not possible to identify genes adjacent to *psbA* due to the limited size of the scaffold sequences.

However, given the *psbA* sequences on these scaffolds share higher sequence similarity with *psbA* from cyanophages, and IIS that are present on scaffolds of cyanophage origin, it is reasonable to assume they are also likely to be of phage origin.

Homing endonuclease

During the identification of intron sequences in the GOS dataset it became apparent that genes with similarity to the homing endonuclease (F-CphI) of S-PM2 were located adjacent to the intron-containing *psbA* genes [14] (Figure S3). Phylogenetic analysis also revealed that *psbA* containing introns also grouped with *psbA* genes that are found adjacent to a homing endonuclease (Figure 3). The arrangement of a homing endonuclease (HE) adjacent to *psbA* has been observed in the genome of S-PM2 and S-RSM88 (Millard *et al.*, 2004). It has been suggested that this arrangement reflects the independent convergence of two separate genetic elements: 1) the intron within *psbA* and 2) the HE F-ChpI downstream of *psbA*, on a common DNA target in a process termed “collaborative homing” and that this is the penultimate step in the proposed pathway for the formation of mobile group I introns (Bonocora and Shub, 2009).

In an effort to identify more intron sequences and determine if the arrangement of *psbA* adjacent to an HE-encoding gene is common, the GOS dataset was searched using the amino acid sequence of F-CphI from S-PM2 as the query. A total of 89 scaffolds were identified using BLAST as having similarity to F-CphI, thus demonstrating it is readily detected in the environment. The scaffold sequences in which HEs were found were extracted, and any genes adjacent to the HE identified. This was possible for 23 scaffold sequences (Table 2). Of the genes adjacent to F-CphI homologues, 15 were identified as *psbA* and 4 as *psbD*. These *psbA* sequences were then searched for introns, but this did not reveal any intron sequences that had not been previously detected. The common occurrence of a HE located next to *psbA* suggests this is not just a chance event, but that there is selective pressure to maintain this arrangement.

Analysis of the *psbA* ORF177 (HE) intergenic region in S-PM2

In an effort to understand localisation of the HE adjacent to *psbA*, we searched the corresponding region of the S-PM2 genome for elements that may maintain 'selective pressure' on the intergenic space between *psbA* and the HE-encoding gene. We used a bioinformatics approach that predicted the ability of the test sequence to form a stable secondary structure, a characteristic of non-coding RNAs (Backofen and Hess, 2010). We identified a possible transcript within the antisense strand of the S-PM2 genome starting at the 5′ end of ORF177 (F-ChpI) and ending within the 3′ half of *psbA*, close to the intron (Figure 4). This method also predicted a second putative transcript antisense to the 5′ end of *psbA* (Figure 4), though this prediction may merely be a reflection of the highly structured 5′ UTR of *psbA* on the sense strand.

Experimental confirmation of the asRNA

Since the bioinformatic analysis strongly suggested the presence of an asRNA in the intergenic region between *psbA* and F-ChpI, 5′ RACE was performed to test these predictions experimentally. 5′ RACE analysis generated two products that mapped to positions 136855 and 136741 of the S-PM2 genome (Figure 5A). The reason for two RACE products is unclear. A possible explanation is that the transcript is processed to form a shorter product. Northern blotting with a probe specific to this putative asRNA confirmed its expression during the infection process (Figure 5B). A ca. 225 bp product was clearly detected. This fits with the 5′ RACE mapped position of 136741 and the predicted 3′ terminator site (Figure 5B). This experimental evidence therefore confirmed the presence of the predicted asRNA and we designated this unique element **C**yanophage **F**unctional **R**NA I (CfrI).

Quantitative PCR analysis of CfrI expression

The expression of CfrI, phage *psbA* and ORF177 (encoding F-ChpI) was monitored using qPCR during the S-PM2 infection cycle. *psbA* expression peaked at 6 h post infection. This peak in expression was also common to CfrI and ORF177 (Figure 6). ORF177 (F-ChpI) showed a large increase in expression between 3-6 h (Figure 6), with the absolute number of ORF177 transcripts exceeding those of *psbA* at this time point (Figure 6). CfrI has a temporal expression pattern similar to both ORF 177 and *psbA* with a peak at 6 h. However, CfrI transcript abundance was significantly lower than that of both *psbA* and ORF177 throughout the infection cycle (Figure 6).

CfrI in other phage genomes

By aligning the sequence of CfrI identified in S-PM2 with the GOS scaffolds it was possible to identify CfrI on scaffolds JCVI_SCAF1101667164370, JCVI_SACF_1096627024160, JCVI_SCAF_1096627283123 and JCVI_SCAF_1097156666624 (Figure S4). Subsequently, by applying the same bioinformatic approach that predicted CfrI in S-PM2 on a selection of GOS scaffolds, an asRNA was predicted in 9 of the 11 scaffolds tested (Table S2). Additionally, the *psbA* region of the four other currently sequenced cyanomyoviruses, (Syn9, S-RMS4, P-SSM2 and P-SSM4) were also analysed for the presence of an asRNA. All four were predicted to encode an asRNA at the 3′ end of the *psbA* gene (Table S1). However, unlike the situation in S-PM2, none of the other cyanophages encode a homing endonuclease downstream of *psbA,* and the asRNA predicted do not appear to overlap the gene downstream of *psb*A.

Discussion

The self-splicing group I intron in S-PM2 that interrupts *psbA* has been shown to be spliced throughout the infection cycle, presumably to maintain a supply of D1 through the infection cycle. In the absence of splicing and excision of the intron it is assumed that a functional D1 polypeptide would not be formed. This conclusion is supported by the situation in *Chlamydomonas reinhardtii* where intron splicing was reduced by directed mutagenesis resulting in the loss of D1 production and consequent reduction in growth rate (Lee and Herrin, 2003). The detection of both spliced and unspliced transcripts during S-PM2 infection suggests that there may be a regulatory role for intron splicing. This would not be without precedent as light has been shown to regulate intron splicing in *Chlamydomonas reinhardtii* (Deshpande *et al*., 1997).

We have shown that introns with sequence conservation can insert in multiple positions within *psbA* sequences. Why multiple IIS are found is unclear. The best strategy for group I introns to proliferate is to locate into highly conserved DNA sequences that have an essential biological role, that are often encountered in the gene pool, and that are conserved across the biological spectrum (Raghavan and Minnick, 2009). *psb*A fulfils these criteria and therefore provides an ideal 'home' for an intron. The multiple IIS may thus merely be a reflection of the highly conserved nature of this gene and that each site meets the requirements for introns to insert into. The origin of these introns remains unknown. The fact that some introns share IIS with introns found in the chloroplasts of *C. reinhardtii* and *O. cardiacum* suggests they may have had a common origin. However, the *psbA* genes they are now located within are all of phage origin. Phylogenetic analysis suggests they are located within *psbA* genes from phage infecting *Synechococcus* rather than *Prochlorococcus,* with no evidence to suggest these introns are present in their *Synechococcus* host. This is consistent with sequencing of numerous *Synechococcus* (Dufresne *et al*., 2008; Scanlan *et al*., 2009) and *Prochlorococcus* (Kettler *et al*., 2007) genomes where no introns have been identified within *psbA* genes. This is surprising given the intragenic recombination of *psbA* that has been proposed to occur between cyanophage and their hosts (Zeidner *et al*., 2005; Sullivan *et al*., 2006). However, this may be due to the numerical bias in the GOS dataset that is dominated by sequences similar to those of *Prochlorococcus* and its infecting phage P-SSM4, with the consequent under-representation of *Synechococcus* (Williamson *et al*., 2008). Previous studies that used only PCR to amplify *psbA* genes have not reported introns within cyanobacterial *psbA* genes, or their phages (Zeidner *et al*., 2005; Sullivan *et al*., 2006; Sharon *et al*., 2007; Wang and Chen, 2008; Marston and Amrich, 2009). This may in part be due to the primers used; the widely used primers of Zeidner *et al*., (2005) span the boundary between the two most common IIS and the *psbA* coding sequence (Figure 2), thereby preventing amplification of any sequences that contain an intron at that particular IIS and so lead to their under-representation in *psbA* gene datasets. The more recent primer set of Wang & Chen (2008) would amplify the most common IIS. However, this primer set has been used to amplify <10 *psbA* genes.

In identifying introns it became apparent that *psbA* is often localised next to a HE similar to that of F-CphI found in S-PM2. It has recently been suggested that localisation of an intron in *psbA* and the presence of a HE adjacent to *psb*A is not accidental. Indeed, it is proposed as the convergence of two genetic parasites on the same conserved region of DNA (Bonocora and Shub, 2009), with these two independent elements acting in a process of collaborative homing to proliferate within a population (Zeng *et al.*, 2009). In collaborative homing the HE targets the IIS as its cutting site, with the intron providing protection against HE nicking its own DNA, and HE providing mobility to transfer into intron-less alleles (Zeng *et al.*, 2009).

Bonocora and Shub (2009) have proposed that the HE will eventually integrate into the intron to form a mobile group I intron which is the most stable entity as the HE can never be separated from the protective function of the intron. The proposed pathway for the formation of mobile group I introns suggests both intron-less and intron-containing *psbA* genes adjacent to a F-CphI would have occurred over time. Both of these scenarios were found in this dataset supporting the proposed model of Bonocora and Shub. However, the final step of integration of F-CphI into an intron was not observed. One intron was found to contain a HE. However, this was significantly different to F-CphI showing similarity to the HE found in the *psbA* intron of *O. cardiacum.* Additionally, the gene immediately downstream of *psbA* was similar to the F-CphI found in S-PM2 (Figure S3).

The failure to detect F-CphI within an intron-containing *psbA* gene may simply be due to the relatively small sample size of the GOS dataset compared to the total gene pool that is present in the oceans. Alternatively, there might be another selective pressure that has prevented the formation of a truly mobile group I intron within cyanophage *psbA* genes. We found the expression of *psbA* in S-PM2 is consistent with previous reports (Clokie *et al.*, 2006) and fits with its proposed function of maintaining host photosynthetic function during infection (Mann *et al.*, 2003; Lindell *et al.*, 2004; Millard *et al.*, 2004; Lindell *et al.*, 2005; Lindell *et al.*, 2007; Bragg and Chisholm, 2008; Hellweger, 2009). We also measured expression of ORF177 encoding the homing endonuclease F-CphI, which was found to be co-expressed with *psbA*. As the spread of the HE into intron-less alleles of *psbA* in other cyanophages is thought to occur during a mixed infection (Zeng *et al.*, 2009), it could be rationalised that the HE would only be expressed once DNA replication has begun, when copies of phage DNA are at their most abundant to provide a substrate for insertion. Thus, the protein would not be needed until DNA has become abundant. Previous work has shown that genes involved in S-PM2 DNA replication are expressed 3 h into the infection cycle (Clokie *et al.*, 2006). DNA abundance is thus highest after this point, and prior to packaging into the protein head. Genes encoding head proteins are not expressed at maximal levels until 6 h and beyond (Clokie *et al.*, 2006). Therefore, DNA is likely to still be abundant and accessible at 6 h which would explain the large increase in expression of ORF177, and *psbA*, after 6 h. The identification of the cis-encoded asRNA, CFrI, is unprecedented in a lytic phage. *Cis*-encoded asRNAs have previously been reported in temperate phages, plasmids and bacterial chromosomes (Brantl, 2007), but not in lytic bacteriophages, or cyanophages. The target of asRNA is often the mRNA that it is complementary to, with post-transcriptional regulation of gene expression being exerted by complementary base pairing (Brantl, 2007). asRNAs which overlap in substantial parts with other genes may be an elegant way to achieve a regulatory connection between neighbouring genes. Indeed, in the cyanobacterium *Anabaena* sp. PCC7120 gene alr1690 has a long 3' overlap with *furA*, encoding a ferric uptake regulator, and controls the expression level in this way (Hernandez *et al.*, 2006). It is worth mentioning that bacterial asRNAs not only trigger degradation of their target mRNAs but can also serve as terminators of transcription (Stork *et al.*, 2007) or as signals for RNA processing, triggering the discoordination of operons (Tramonti *et al.*, 2008). In S-PM2 CfrI joins the genetic elements of *psbA* and the gene encoding F-CphI. Presumably, for the gene encoding F-CphI to become integrated into the intron it has to be removed from its current position. As HEs are normally found within intergenic regions, their inexact removal is unlikely to cause a detrimental effect (Raghavan and Minnick, 2009). However, in S-PM2 the 3′ end of *psbA* and the 5′ end of the gene encoding F-CphI are directly linked by the asRNA CfrI. Therefore, any rearrangement of the HE-encoding gene into the intron of *psbA* will cause disruption of the CfrI sequence, presumably leading to a lack of function. Thus, we propose the asRNA CfrI provides a selective pressure to maintain the current *status quo* preventing the formation of a mobile group I intron, as removal of the endonuclease-encoding gene whilst maintaining the asRNA is likely to be a rare event. Thus, we propose CfrI has prevented or slowed the evolution of the two genetic parasites of the intron and HE into a single mobile group I intron.

Whilst the function of CfrI is still unknown, the fact it is differentially expressed during the infection cycle suggests it plays a regulatory role. Given that asRNAs normally regulate the gene they are antisense to by complementary base pairing, it would be reasonable to assume it regulates *psbA* or ORF 177 (F-CphI) gene expression. However, the predicted presence of an asRNA at the 3′ end of *psbA* genes in cyanophages that lack a homologue of F-CphI downstream, suggests that this asRNA specifically regulates expression of *psbA*. Given that early and late promoter motifs have already been identified upstream of cyanophage *psbA* (Mann *et al.*, 2005) such additional regulatory capacity may be important for phage infection under particular environmental conditions e.g. high light intensities. This would be consistent with modelling studies that suggest that phage photosynthesis genes provide an increase in fitness in a manner that is correlated with irradiance (Bragg and Chisholm, 2008; Hellweger, 2009).

The archetypal example of a phage asRNA overlapping the 3′ end of a protein-gene is the 77 nt OOP asRNA of bacteriophage λ. The OOP asRNA is complementary to the 3′ end of the λ cII-repressor mRNA. Over-expression of OOP asRNA from a plasmid vector results in RNAse III dependent cleavage of cII mRNA (Krinke and Wulff, 1987). Regulation of the stress-inducible photosynthetic gene *isiA* by the asRNA IsiR in *Synechocystis* sp. PCC6803 is also consistent with this model, since accumulation of mRNA and asRNA follows inverse kinetics and is mutually exclusive (Duhring *et al.*, 2006).

In contrast, the function of CfrI is more likely to be protective as it appears to be co-ordinately expressed with *psbA*. It may act in a similar manner to some asRNAs observed in *Prochlorococcus*. In *Prochlorococcus* sp. MED4, the asRNA Yfr15 accumulates during phage infection (Steglich *et al.*, 2008). Yfr15 overlaps the 3′ end of gene PMED4_07441 (PMM0686), the most highly up-regulated host mRNA during phage infection. In contrast to this high level of expression, the vast majority of host-encoded mRNAs are rapidly degraded (Lindell *et al.*, 2007), implying that Yfr15 protects the PMED4_07441 mRNA, for example by rendering RNase E recognition sites inaccessible. In this context it is noteworthy that we detected accumulation of un-spliced *psbA* precursor transcripts, indicating slow kinetics of intron splicing. This would imply there would be a delay before exon 2 of *psb*A would – by physical occlusion by the translating ribosomes – become protected from endonuclease cleavage. This may also explain why a lower stoichiometric ratio of the asRNA relative to the mRNA may be sufficient, a hypothesis that can be tested in future experiments.

It is only recently that the role and importance of asRNAs in cyanobacterial regulation has become apparent (Steglich *et al.*, 2008; Georg *et al.*, 2009). In the cyanobacterium *Prochlorococcus*, that possesses a highly reduced genome, it is thought that *trans*-acting ncRNAs and *cis*-acting asRNAs play an important role in regulation (Steglich *et al.*, 2008). The co-evolution of virus and host, and transfer of genetic material between them, coupled with the relatively limited coding capacity of the phage genome, implies that similar genetically-conservative ncRNAs and asRNAs remain to be identified in the genomes of lytic phages.

Conclusions

The occurrence of introns inserted at multiple positions within cyanophage *psbA* genes appears to be a widespread phenomenon. These intron-containing *psbA* genes are often located adjacent to a gene encoding a homing endonuclease, seemingly the result of the co-evolution of two genetic parasites on a single conserved sequence. Within cyanophage S-PM2 these two separate genetic elements are 'joined' by an asRNA, CfrI. CfrI is the first example of an asRNA in a lytic bacteriophage. Its co-expression with *psbA* points to a role in regulation. The discovery of sequences similar to CfrI in other cyanophage scaffolds suggests asRNAs, and perhaps more generally other ncRNAs, are likely to be important in regulating cyanophage gene expression. CfrI however, also has the potentially unique property of preventing or slowing down the evolution of two genetic parasites, an intron and a HE into a single mobile group I intron.

Acknowledgments

We acknowledge Michael Zuker and Nicholas Markham for UNAFOLD. DJE was supported by grants from the Medical Research Council and the Wellcome Trust; WRH was supported by the DFG Focus program "Sensory and regulatory RNAs in Prokaryotes" SPP1258 (project HE 2544/4-1) and the BMBF - Freiburg Initiative in Systems Biology - grant 0313921; ADM was funded by Leverhulme Trust grant F/00215/AL to DJS.

Figure and Table Legends

Figure 1

In vivo splicing of a group I intron within the cyanophage S-PM2 *psbA* gene. RNA isolated from S-PM2 infected *Synechococcus* sp. WH7803 was analyzed by RT-PCR from samples taken at 1, 3 and 9 hr post infection. No reverse transcriptase controls (nrtc) were used to test for contaminating DNA in purified RNA. Genomic DNA from S-PM2 was used a positive control (Lane G). A no template sample was used as a negative control (Lane C-ve). The 1 kb and 1.5 kb size standards are marked.

Figure 2

Intron insertion sites (IIS) within *psbA*. Amino acid sequences derived from *psbA* genes identified to have introns were aligned. Due to the partial sequence of some of the *psbA* genes the IIS is reported relative to the position of the full length sequence of S-PM2. The trans-membrane domains of the D1 protein are marked by grey text. The amino acid sequences targeted by the universal *psbA* primer set (Zeidner *et al.*, 2003) are underlined. IIS are marked by arrows, with numbers corresponding to the following sequences: 1: JCVI_SCAF1101669142352, 2: JCVI_SCAF_1101667044432, 3: JCVI_SCAF_1098315327957, 4: *Chlamydomonas reinhardtii* intron 4, 5: *Oedogonium cardiacum* intron 1, 6: JCVI_SCAF_1096627024160, 7 : JCVI_SCAF_1096627666661, 8: JCVI_SCAF_1101668247417, 9: JCVI_SCAF_1101669425113, 10: *Chlamydomonas reinhardtii* intron 1, 11 : S-PM2, :12 S-RSM88, 13: JCVI_SCAF_1101667034653, 14:JCVI_SCAF_1101669070555, 15: JCVI_SCAF_1097156666624, 16: JCVI_SCAF_1096627283123, 17: JCVI_SCAF_1097207205912, 18: JCVI_SCAF_1096626190594, 19: JCVI_SCAF_1101669414852, 20: JCVI_SCAF_1101668234973, 21: JCVI_SCAF_1096627024703.

Figure 3

Phylogenetic relationships amongst *psbA* genes of cyanophages, cyanobacteria and plastids from cultures and environmental samples. Trees are based on an alignment of 925 nucleotides, clade support values are the result of 200,000 iterations and a burn-in of 25% using Mr BAYES [20]. Clade support values >90 are marked by ●, >80 and < 90 by ■ and those >70 and <90 are marked by a ○. GenBank accession numbers of *psbA* sequences used for phylogenetic analysis were as follows: *Synechococcus* (*Synechococcus* BL107: acc NA_AAT20000000, *Synechococcus* sp. WH8102: acc NC_005070, *Synechococcus* sp. WH7803: acc NC_00009481, *Synechococcus* sp. RCC307: acc NC_00009482, *Synechococcus* sp. RS9916: acc NZ_AA0A00000000, *Synechococcus* sp. CC9311: acc NC_008319); *Prochlorococcus* (*Prochlorococcus* sp. MIT9303: acc NC_008820, *Prochlorococcus* sp. MIT9319: acc NC_005071, *Prochlorococcus* sp. MED4: acc NC_005072, *Prochlorococcus* sp. MI9515: acc NC_008817, *Prochlorococcus* sp. MIT9202: acc NZ_ACDW00000000, *Prochlorococcus* sp. NAT2LA: acc NC_007335, *Prochlorococcus* sp. AS9601: acc NC_008816, *Prochlorococcus* sp. MIT9301: acc NC_009091, *Prochlorococcus* sp. MIT9211: acc NC_009976, *Prochlorococcus* sp. SS120: acc NC_00xxxx,); *Synechococcus* phage (Syn9: acc NC_008296, S-RSM4: acc CAR63316.1, S-PM2 : acc NC_006820, S-RSM88: acc AJ629075, S-RSM2: acc AJ628768 , S-WHM1:acc AJ628769 , S-RSM28: acc AJ629221, S-BM4: acc AJ628858); *Prochlorococcus* phage (P-SSM4: acc NC_006884, P-SSM2: acc NC_006883, P-SSP7: acc NC_006882); plastids (*Ostreococcus tauri*: acc NC_008289, *Oedogonium cardiacum*: acc NC_011031, *Chlamydomonas reinhardtii*: acc NC_005353, *Cyanidium caldarium*: acc NC_001840, *Guillardia theta*: acc NC_000926, *Heterosigma akashaiwo*: acc NC_010772, *Odontella sinensis*: acc NC_001713, *Phaeodactylum tricornutum*: acc NC_008588). *Arabidopsis thaliana*: acc NC_009032 was used to root the tree. Roman numerals are used to denote the different copies of *psbA* found within the genomes of *Synechococcus* and *Prochlorococcus*. The numbers in square brackets are the average mol %GC content and the 3rd base mol %GC content, respectively.

Figure 4

Prediction of an ncRNA antisense to *psbA* and ORF177. The *psbA* region of S-PM2 analysed in 200 nt windows incrementing every 10 nt. The mean folding energy (MFE) for each was calculated and compared to 1000 scrambles of the same sequence, the MFE for each window is plotted (**.**) with those windows that had a MFE above the 99th percentile of the 1000 scrambles marked (●). The position of the genes *psbA*, ORF177 (encoding F-CphI), ORF178 and ORF179 (*psbD*) are marked by arrows. The position of the previously predicted ncRNA is marked with a dotted arrow.

Figure 5

Presence of an antisense RNA linking the S-PM2 endonuclease gene with the *psbA* second exon. (**A**) Experimentally verified 5′ ends of the antisense transcript overlapping the 3′ end of the intron, exon 2 of *psbA* and the 5′ end of the endonuclease gene orf177 were mapped to positions 136855 (long transcript) and 136741 (short transcript) on the complementary strand. ORF177 was recently identified as a free standing homing endonuclease gene (HE), targeting intron-less *psbA* genes of marine cyanobacteria. The sequence elements (136526-136560, complementary strand) which are predicted to form the terminator helix for the antisense transcript are underlined. (**B**) Separation of 20 µg of total RNA from phage-infected *Synechococcus* sp. WH7803 (+) and from non-infected cells (-) on a 10% (w/v) polyacrylamide gel. The Northern hybridization (right) indicates a prominent band of approximately 225 bp and some weaker bands of higher molecular weight in the RNA from phage-infected cells but not in the RNA from control cells. The band 225 bp in size corresponds to a transcript with the second mapped 5' end (short transcript) and the predicted terminator. The blot was hybridized with a single-stranded RNA probe directed against the antisense transcript. An RNA molecular weight standard (M) is shown to the left.

Figure 6

Expression of *psbA* (▲), CfrI (●) and ORF177 (■). Plotted values are the mean of 3 independent biological replicates with error bars representing Stdev. The relative expression of each transcript is plotted in panels A, B, and C with absolute transcript abundance plotted in panel D. Cells begin to lyse after 9 h under the conditions used (see Clokie *et al*., 2006).

Table 1

Cyanophage genomes and global ocean survey scaffolds from in which introns were identified. "Fragment to short" denotes the *psbA* sequence was not long enough to cover the region where the cyanophage/cyanobacterial-specific PsbA motif R/KETTXXXSQ/H is found.

Table 2

GOS scaffolds on which homologues of the F-CphI homing endonuclease were detected.

References

Backofen R, Hess WR (2010). Computational prediction of sRNAs and their targets in bacteria. *RNA Biology* **7:** 1-10.

Bonocora RP, Shub DA (2009). A likely pathway for formation of mobile group I introns. *Curr Biol* **19:** 223-8.

Bragg JG, Chisholm SW (2008). Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS One* **3:** e3550.

Brantl S (2007). Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr Opin Microbiol* **10:** 102-9.

Brouard JS, Otis C, Lemieux C, Turmel M (2008). Chloroplast DNA sequence of the green alga *Oedogonium cardiacum* (*Chlorophyceae*): unique genome architecture, derived characters shared with the Chaetophorales and novel genes acquired through horizontal transfer. *BMC Genomics* **9:** 290.

Chenard C, Suttle CA (2008). Phylogenetic diversity of sequences of cyanophage photosynthetic gene psbA in marine and freshwaters. *Appl Environ Microbiol* **74:** 5317-24.

Clokie MR, Shan J, Bailey S, Jia Y, Krisch HM, West S *et al* (2006). Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ Microbiol* **8:** 827-35.

Deshpande NN, Bao Y, Herrin DL (1997). Evidence for light/redox-regulated splicing of *psbA* pre-

RNAs in *Chlamydomonas* chloroplasts. *RNA* **3:** 37-48.

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP *et al* (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9:** R90.

Duhring U, Axmann IM, Hess WR, Wilde A (2006). An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc Natl Acad Sci USA* **103:** 7054-8.

Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR (2009). Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol Syst Biol* **5:** 305.

Hellweger FL (2009). Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. *Environ Microbiol* **11:** 1386-94.

Hernandez JA, Muro-Pastor AM, Flores E, Bes MT, Peleato ML, Fillat MF (2006). Identification of a *furA* cis antisense RNA in the cyanobacterium *Anabaena* sp. PCC 7120. *J Mol Biol* **355:** 325-34.

Itoh T, Tomizawa J (1980). Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proc Natl Acad Sci USA* **77:** 2450-4.

Jurica MS, Stoddard BL (1999). Homing endonucleases: structure, function and evolution. *Cell Mol Life Sci* **55:** 1304-26.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** e231.

Krinke L, Wulff DL (1987). OOP RNA, produced from multicopy plasmids, inhibits lambda cII gene expression through an RNase III-dependent mechanism. *Genes Dev* **1:** 1005-13.

Lacatena RM, Cesareni G (1981). Base pairing of RNA I with its complementary sequence in the primer precursor inhibits ColE1 replication. *Nature* **294:** 623-6.

Lee J, Herrin DL (2003). Mutagenesis of a light-regulated *psbA* intron reveals the importance of efficient splicing for photosynthetic growth. *Nucleic Acids Res* **31:** 4361-72.

Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T *et al* (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449:** 83-6.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438:** 86-9.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101:** 11013-8.

Liu Q, Belle A, Shub DA, Belfort M, Edgell DR (2003). SegG endonuclease promotes marker exclusion and mediates co-conversion from a distant cleavage site. *J Mol Biol* **334:** 13-23.

Livak KJ, Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25:** 402-8.

Logemann J, Schell J, Willmitzer L (1987). Improved method for the isolation of RNA from plant tissues. *Anal Biochem* **163:** 16-20.

Lu J, Chen F, Hodson RE (2001). Distribution, isolation, host specificity, and diversity of cyanophages infecting marine *Synechococcus* spp. in river estuaries. *Appl Environ Microbiol* **67:** 3285-90.

Mann NH, Clokie MR, Millard A, Cook A, Wilson WH, Wheatley PJ *et al* (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. *J Bacteriol* **187:** 3188-200.

Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424:** 741.

Markham NR, Zuker M (2008). UNAFold: software for nucleic acid folding and hybridization. In: Keith JM (ed). Bioinformatics, Volume II. Structure, Functions and Applications. Humana Press. pp 3-31.

Marston MF, Amrich CG (2009). Recombination and microdiversity in coastal marine cyanophages. *Environ Microbiol* **11:** 2893-903.

Marston MF, Sallee JL (2003). Genetic diversity and temporal variation in the cyanophage community infecting marine *Synechococcus* species in Rhode Island's coastal waters. *Appl Environ Microbiol* **69:** 4639-47.

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH *et al* (2002). The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell* **14:** 2659-79.

Millard A, Clokie MR, Shub DA, Mann NH (2004). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* **101:** 11007-12.

Millard A, Mann NH (2006). A temporal and spatial investigation of cyanophage abundance in the Gulf of Aqaba, Red Sea. *J Mar Biol Assocn UK* **86:** 507-515.

Millard AD, Zwirglmaier K, Downey MJ, Mann NH, Scanlan DJ (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:** 2370-87.

Raghavan R, Minnick MF (2009). Group I introns and inteins: disparate origins but convergent parasitic strategies. *J Bacteriol* **191:** 6193-202.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73:** 249-99.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5:** e75.

Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N *et al* (2009). Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461:** 258-62.

Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB *et al* (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J* **1:** 492-501.

Simmonds P, Smith DB (1999). Structural constraints on RNA virus evolution. *J Virol* **73:** 5787-94.

Spiegelman WG, Reichardt LF, Yaniv M, Heinemann SF, Kaiser AD, Eisen H (1972). Bidirectional transcription and the regulation of Phage lambda repressor synthesis. *Proc Natl Acad Sci USA* **69:** 3156-60.

Steglich C, Futschik ME, Lindell D, Voss B, Chisholm SW, Hess WR (2008). The challenge of regulation in a minimal photoautotroph: non-coding RNAs in *Prochlorococcus*. *PLoS Genet* **4:** e1000173.

Stork M, Di Lorenzo M, Welch TJ, Crosa JH (2007). Transcription termination within the iron transport-biosynthesis operon of *Vibrio anguillarum* requires an antisense RNA. *J Bacteriol* **189:** 3479-88.

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3:** e144.

Sullivan MB, Krastins B, Hughes JL, Kelly L, Chase M, Sarracino D *et al* (2009). The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. *Environ Microbiol* **11:** 2935-51.

Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* **4:** e234.

Sullivan MB, Waterbury JB, Chisholm SW (2003). Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424:** 1047-51.

Suttle CA (2005). Viruses in the sea. *Nature* **437:** 356-61.

Suttle CA (2007). Marine viruses - major players in the global ecosystem. *Nat Rev Microbiol* **5:** 801-12.

Suttle CA, Chan AM (1994). Dynamics and distribution of cyanophages and their effect on marine *Synechococcus* spp. *Appl Environ Microbiol* **60:** 3167-3174.

Tramonti A, De Canio M, De Biase D (2008). GadX/GadW-dependent regulation of the *Escherichia coli* acid fitness island: transcriptional control at the *gadY-gadW* divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. *Mol Microbiol* **70:** 965-82.

Wang K, Chen F (2008). Prevalence of highly host-specific cyanophages in the estuarine environment. *Environ Microbiol* **10:** 300-12.

Waterbury JB, Valois FW (1993). Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl Environ Microbiol* **59:** 3393-3399.

Weigele PR, Pope WH, Pedulla ML, Houtz JM, Smith AL, Conway JF *et al* (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol* **9:** 1675-95.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI *et al* (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3:** e1456.

Wilson WH, Carr NG, Mann NH (1996). The effect of phosphate status on the kinetics of cyanophage infection in the oceanic cyanobacterium *Synechococcus* sp WH7803. *J Phycol* **32:** 506-516.

Wilson WH, Joint IR, Carr NG, Mann NH (1993). Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. Strain WH7803. *Appl Environ Microbiol* **59:** 3736-3743.

Wyman M, Gregory RP, Carr NG (1985). Novel role for phycoerythrin in a marine cyanobacterium, *Synechococcus* strain DC2. *Science* **230:** 818-820.

Zeidner G, Bielawski JP, Shmoish M, Scanlan DJ, Sabehi G, Beja O (2005). Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ Microbiol* **7:** 1505-13.

Zeidner G, Preston CM, Delong EF, Massana R, Post AF, Scanlan DJ *et al* (2003). Molecular diversity among marine picophytoplankton as revealed by *psbA* analyses. *Environ Microbiol* **5:** 212-6.

Zeng Q, Bonocora RP, Shub DA (2009). A free-standing homing endonuclease targets an intron insertion site in the *psbA* gene of cyanophages. *Curr Biol* **19:** 218-22.

Supplementary Information

Supplementary Figure & Table Legends

Figure S1

Sequence alignment of introns inserted after codon 334 and 338. Introns from scaffolds JCVI_SCAF_1096626190549 JCVI_SCAF_1096627283123, JCVI_SCAF_1097156666624, JCVI_SCAF_1097207205912, JCVI_SCAF_1101667034653, JCVI_SCAF_1101669070555, JCVI_SCAF_1096627024703 were aligned using LocARNA [1] with the introns of cyanophages S-PM2 and S-RSM88. The conserved stem structures as previously identified in the S-PM2 intron [2] are marked as P1, P3, P4,P5, P6, P6A, P7, P7.1, P8 and P9.

Figure S2

Sequence alignment of introns from scaffolds JCVI_SCAF_11098315327957 and JCVI_SCAF_1101667044432 with the introns from C. *reinhardtii* and *O. cardiacum.* There was conservation in sequence identity at the 5' end of the intron. Sequences were aligned with locARNA [1].

Figure S3

Gene order in S-PM2 and scaffolds from the GOS dataset. A) The scaffolds JCVI_SCAF1097207205912 and JCVI_SCAF1096626190549 were found to have a gene order similar to that of S-PM2 with a homologue of F-CphI adjacent to *psbA*, whereas for JCVI_SCAF1096627024703 *psbA* is adjacent to genes encoding transaldolase and 6-phosphogluconate dehydrogenase, respectively. Both of these genes are known to be widespread in cyanophage genomes, with both genes having greatest similarity to cyanophage copies of these genes, rather than *Synechococcus* of *Prochlorococcus* host copies. B) Gene order in S-PM2 compared to the GOS scaffold sequences JCVI_SCAF_1096627283123, JCVI_SCAF_1096627674162 , JCVI_SCAF_1096627019931, JCVI_SCAF_1096626856934 and JCVI_SCAF_1096627024160. The gene encoding the homing endonuclease F-CphI was often found to be downstream of *psbA* in JCVI_SCAF_1096627283123, JCVI_SCAF_1096627674162, JCVI_SCAF_1096626856934 and JCVI_SCAF_1096627024160, as was observed in the genome of S-PM2. In S-PM2, downstream of F-CphI, is a hypothetical protein followed by *psbD*. Similar hypothetical proteins were not observed in any of the GOS sequences. However, other phage associated genes were found, such as *hli, gnd* and *psbD.* Variation in the genes found downstream of F-CphI suggests there may be rearrangement of the genome in other cyanophages. This could possibly be caused by a mobile element such as a homing endonuclease. Intriguingly, for JCVI_SCAF_1096627024160 a homologue of F-CphI was found downstream of *psbA*, whilst a second homing endonuclease with similarity to the endonuclease found in intron of *O. cardiacum* was found within the intron of the *psbA* gene itself.

Figure S4

Sequence alignment of regions similar to the asRNA CfrI in the phage S-RSM88 and scaffolds JCVI_SCAF_1101667164370, JCVI_SCAF_1096627024160, JCVI_SCAF1096627283123 JCV_SCAF_1097156666624. The sequence of CfrI is reported 5′ to 3′. The blue line shows the region in which the asRNA CfrI overlaps with the endonuclease-encoding gene on the sense strand. The green line represents the intergenic space between the endonuclease-encoding gene and *psbA* on the sense strand. The black line is the region CfrI overlaps with the 3′ end of the *psbA* gene on the sense strand. The stop codon of *psbA* and the start codon of the endonuclease-encoding gene on sense strand are underlined.

Figure S5

Phylogenetic tree based on an alignment of 193 amino acid residues of the translated *gnd* gene. MrBayes was used to reconstruct the phylogeny with the Dayhoff model for amino acid substitution used for 200,000 iterations, with a burn-in after 20%. *Escherichia coli* was used as an outgroup.

JCVI_SCAF1096627024703was clearly found to cluster with the phage sequences S-RSM4 and Syn9, and is clearly separate from the *Synechococcus* host clade. This suggests the scaffold is of phage origin.

Figure S6

Phylogenetic tree based on an alignment of 132 amino acid residues of the translated *talC* gene. MrBayes was used to reconstruct the phylogeny with the Dayhoff model of amino acid substitution used for 300,000 iterations with a burn-in after 20%. *Escherichia coli* was used as an outgroup. A clade containing all *Synechococcus* was observed, alongside a discrete *Prochlorococcus* clade. The GOS sequences were found to group in a separate clade with other cyanophage sequences, including those from S-RSM4, Syn9, S-RSM2, P-SSM2 and P-SSM4. This suggests these GOS sequences are of phage origin.

Table S1

Primers and probes used in this study

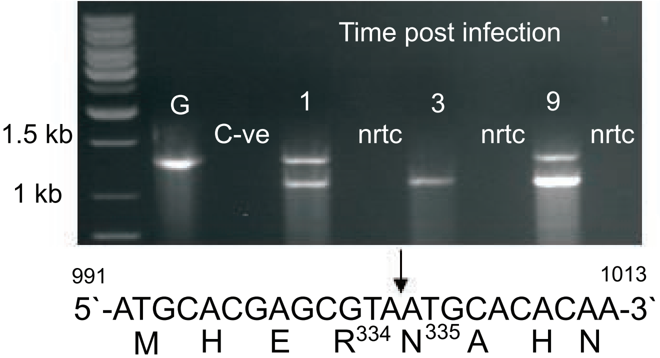
Table S2

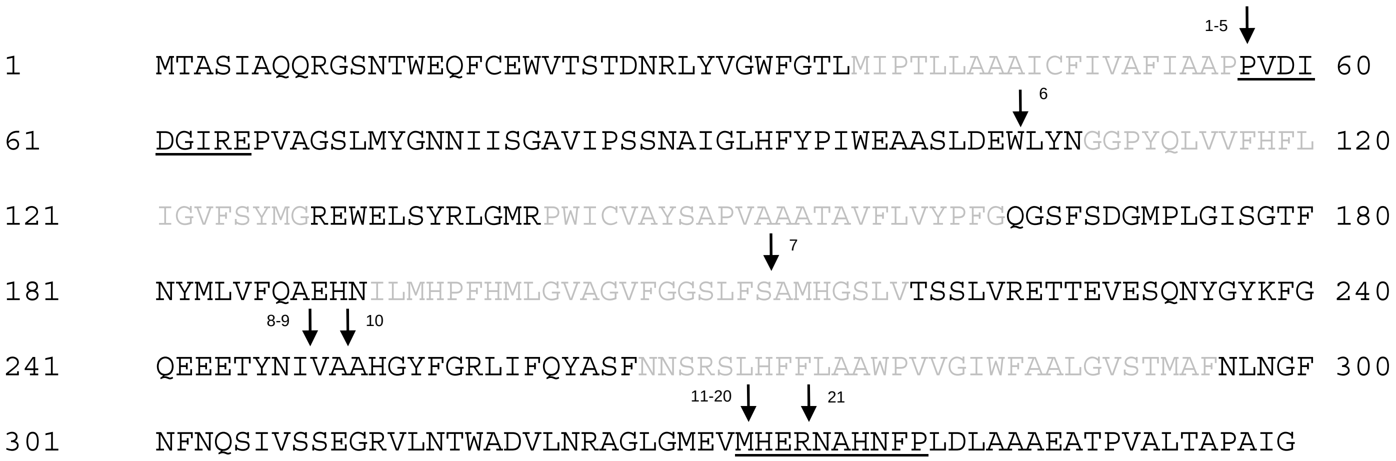
GOS scaffolds and phage genomes tested for the prediction of an asRNA. The gene coordinates of *psbA*, and the gene immediately downstream of *psbA* are listed. The predicted start and stop position of any predicted ncRNA are also listed. Predictions were based on a window size of 200 nt with an overlap of 190 nt, 1000 scrambles of each window were made. A predicted ncRNA was decided based on the actual sequence being above the 99th percentile of 1000 scrambles. Testing of the listed phage and scaffolds required ~ 30 million folding calculations.

1. Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F., and Backofen, R. (2007). Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering.

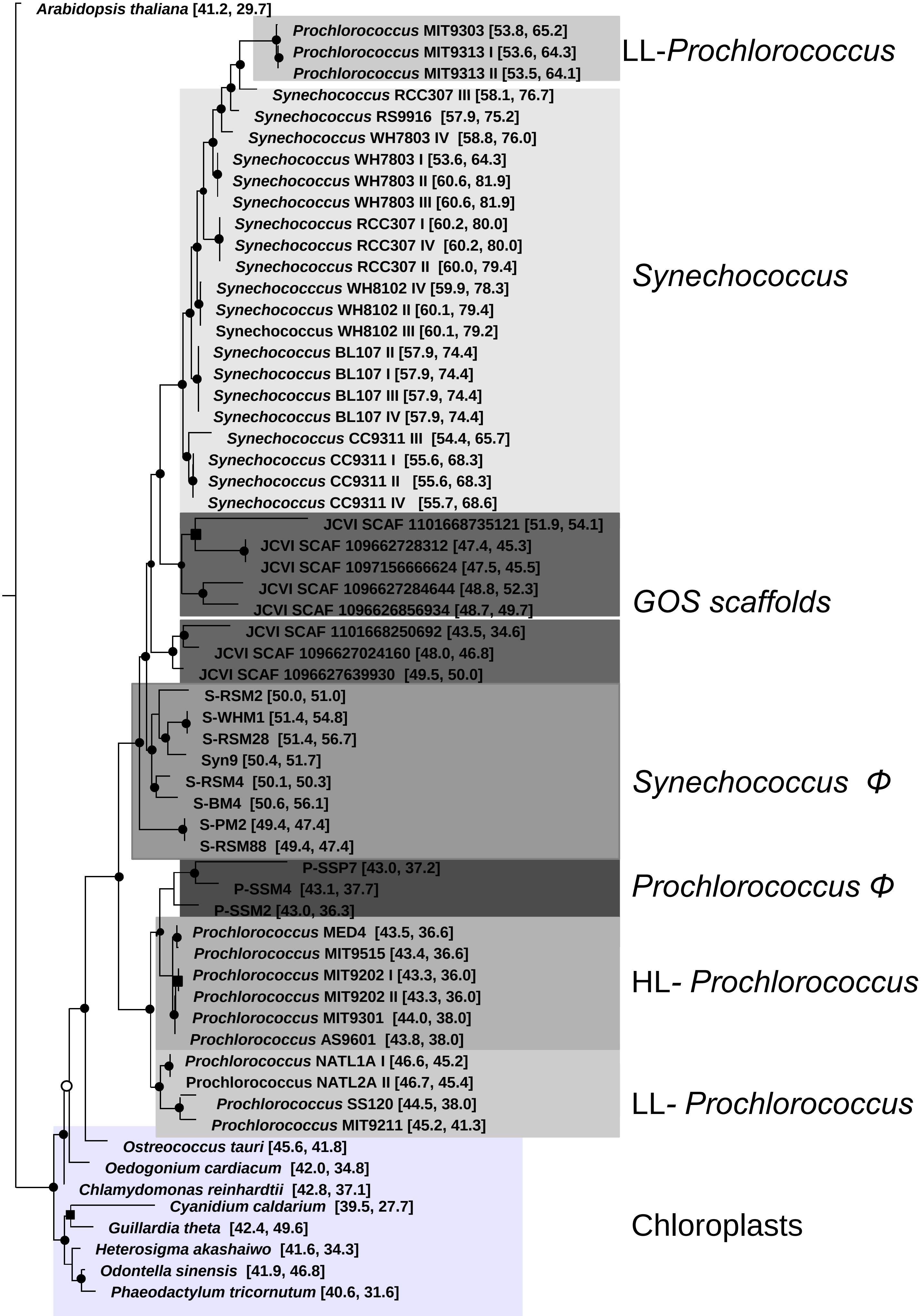
PLoS Comput Biol **3:** e65.

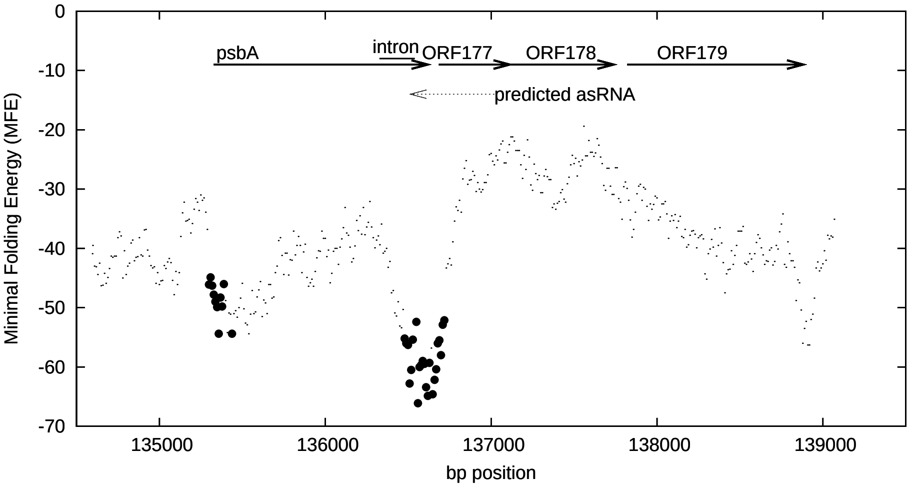
2. Millard, A., Clokie, M.R., Shub, D.A., and Mann, N.H. (2004). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* **101:** 11007-11012.

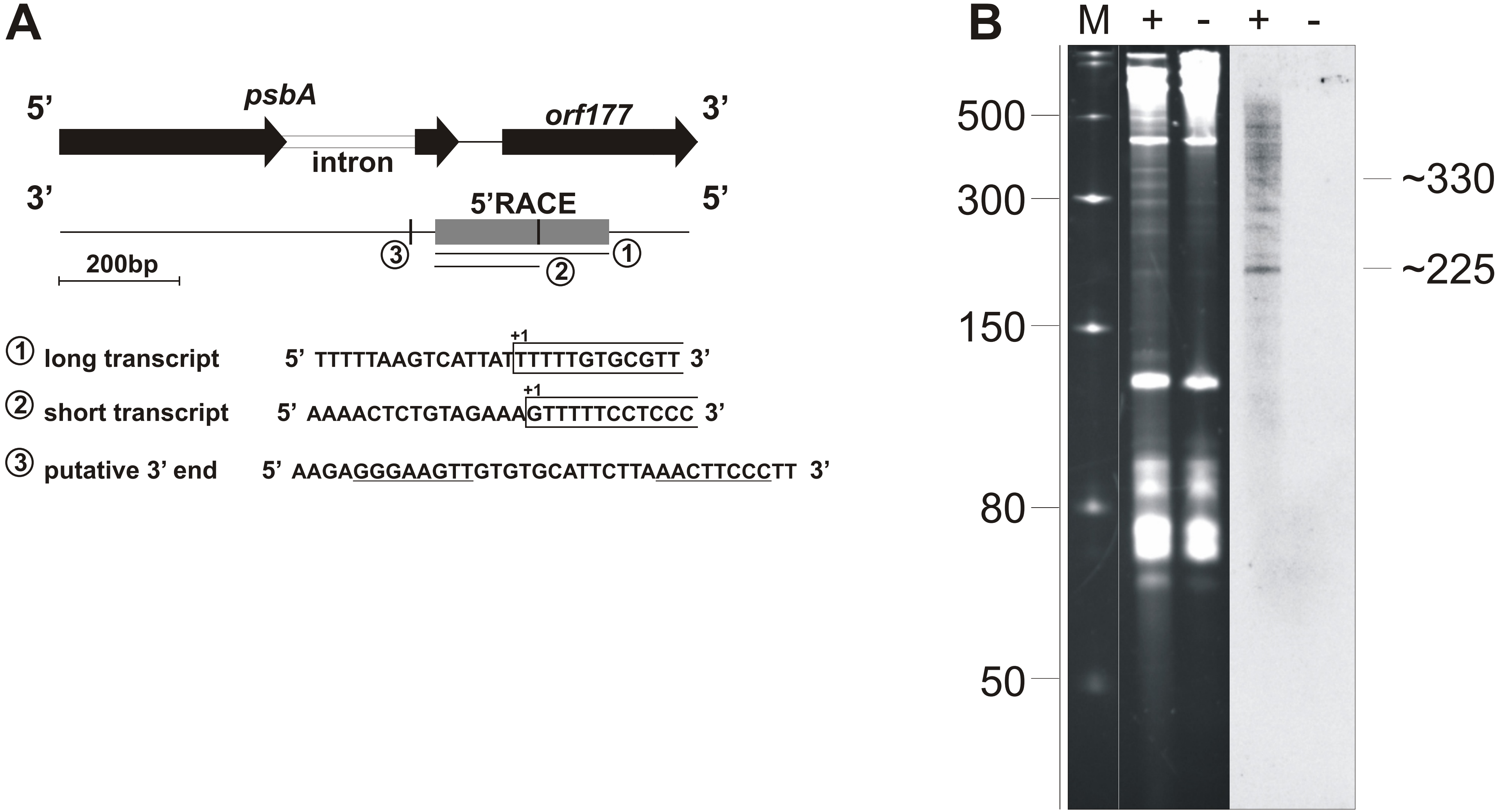

Time post infection
G
C-ve
1
nrtc
3
nrtc
9
nrtc
1.5 kb
1 kb
991
1013
5`-ATGCACGAGCGTAATGCACACAA-3`
M H E R334 N335 A H N

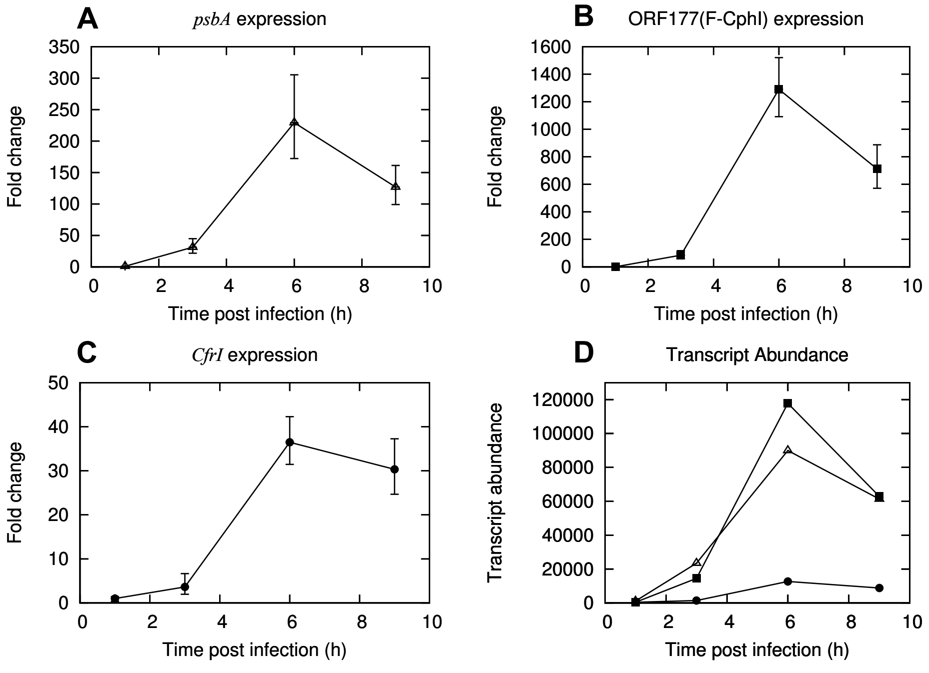


```
1     MTASIAQQRGSNTWEQFCEWVTSTDNRLYVGWFGTLMIPTLLAAAICFIVAFIAAPPVDI  60

61    DGIREPVAGSLMYGNNIISGAVIPSSNAIGLHFYPIWEAASLDEWLYNGGPYQLVVFHFL 120

121   IGVFSYMGREWELSYRLGMRPWICVAYSAPVAAATAVFLVYPFGQGSFSDGMPLGISGTF 180

181   NYMLVFQAEHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLVRETTEVESQNYGYKFG 240

241   QEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVVGIWFAALGVSTMAFNLNGF 300

301   NFNQSIVSSEGRVLNTWADVLNRAGLMEVMHERNAHNFPLDLAAAEATPVALTAPAIG
```


Arabidopsis thaliana [41.2, 29.7]
Prochlorococcus MIT9303 [53.8, 65.2]
Prochlorococcus MIT9313 I [53.6, 64.3]
Prochlorococcus MIT9313 II [53.5, 64.1]
LL-Prochlorococcus
Synechococcus RCC307 III [58.1, 76.7]
Synechococcus RS9916 [57.9, 75.2]
Synechococcus WH7803 IV [58.8, 76.0]
Synechococcus WH7803 I [53.6, 64.3]
Synechococcus WH7803 II [60.6, 81.9]
Synechococcus WH7803 III [60.6, 81.9]
Synechococcus RCC307 I [60.2, 80.0]
Synechococcus RCC307 IV [60.2, 80.0]
Synechococcus RCC307 II [60.0, 79.4]
Synechococcus WH8102 IV [59.9, 78.3]
Synechococcus WH8102 II [60.1, 79.4]
Synechococcus WH8102 III [60.1, 79.2]
Synechococcus BL107 II [57.9, 74.4]
Synechococcus BL107 I [57.9, 74.4]
Synechococcus BL107 III [57.9, 74.4]
Synechococcus BL107 IV [57.9, 74.4]
Synechococcus CC9311 III [54.4, 65.7]
Synechococcus CC9311 I [55.6, 68.3]
Synechococcus CC9311 II [55.6, 68.3]
Synechococcus CC9311 IV [55.7, 68.6]
Synechococcus
JCVI SCAF 1101668735121 [51.9, 54.1]
JCVI SCAF 109662728312 [47.4, 45.3]
JCVI SCAF 1097156666624 [47.5, 45.5]
JCVI SCAF 1096627284644 [48.8, 52.3]
JCVI SCAF 1096626856934 [48.7, 49.7]
GOS scaffolds
JCVI SCAF 1101668250692 [43.5, 34.6]
JCVI SCAF 1096627024160 [48.0, 46.8]
JCVI SCAF 1096627639930 [49.5, 50.0]
S-RSM2 [50.0, 51.0]
S-WHM1 [51.4, 54.8]
S-RSM28 [51.4, 56.7]
Syn9 [50.4, 51.7]
S-RSM4 [50.1, 50.3]
S-BM4 [50.6, 56.1]
S-PM2 [49.4, 47.4]
S-RSM88 [49.4, 47.4]
Synechococcus Φ
P-SSP7 [43.0, 37.2]
P-SSM4 [43.1, 37.7]
P-SSM2 [43.0, 36.3]
Prochlorococcus Φ
Prochlorococcus MED4 [43.5, 36.6]
Prochlorococcus MIT9515 [43.4, 36.6]
Prochlorococcus MIT9202 I [43.3, 36.0]
Prochlorococcus MIT9202 II [43.3, 36.0]
Prochlorococcus MIT9301 [44.0, 38.0]
Prochlorococcus AS9601 [43.8, 38.0]
HL- Prochlorococcus
Prochlorococcus NATL1A I [46.6, 45.2]
Prochlorococcus NATL2A II [46.7, 45.4]
Prochlorococcus SS120 [44.5, 38.0]
Prochlorococcus MIT9211 [45.2, 41.3]
LL- Prochlorococcus
Ostreococcus tauri [45.6, 41.8]
Oedogonium cardiacum [42.0, 34.8]
Chlamydomonas reinhardtii [42.8, 37.1]
Cyanidium caldarium [39.5, 27.7]
Guillardia theta [42.4, 49.6]
Heterosigma akashaiwo [41.6, 34.3]
Odontella sinensis [41.9, 46.8]
Phaeodactylum tricornutum [40.6, 31.6]
Chloroplasts


0
-10
-20
-30
-40
-50
-60
-70
Minimal Folding Energy (MFE)
135000
136000
137000
138000
139000
bp position
psbA
intron
ORF177
ORF178
ORF179
predicted asRNA


A
5'
psbA
orf177
3'
intron
3'
5'RACE
5'
③
①
②
200bp
① long transcript 5' TTTTTAAGTCATTATTTTTTGTGCGTT 3'
+1
② short transcript 5' AAAACTCTGTAGAAAGTTTTTCCTCCC 3'
+1
③ putative 3' end 5' AAGAGGGAAGTTGTGTGCATTCTTAAACTTCCCTT 3'
B
M
+
-
+
-
500
300
150
80
50
~330
~225


A
psbA expression
350
300
250
200
150
100
50
0
Fold change
0
2
4
6
8
10
Time post infection (h)
B
ORF177(F-CphI) expression
1600
1400
1200
1000
800
600
400
200
0
Fold change
0
2
4
6
8
10
Time post infection (h)
C
CfrI expression
50
40
30
20
10
0
Fold change
0
2
4
6
8
10
Time post infection (h)
D
Transcript Abundance
120000
100000
80000
60000
40000
20000
0
Transcript abundance
0
2
4
6
8
10
Time post infection (h)

	Presence of the			% ID to		
Cyanophage/Scaffold	“G/KETTXXXSQ/H”	Intron	Phylogenetic	S-PM2	Scaffold	
Sequence	motif in PsbA	Length	Classification	intron	Length	Area of Isolation (Reference)
S-PM2	✓	212	Cyanophage	100	n/a	Plymouth - United Kingdom (Wilson *et al* 1993)
S-RSM88	✓	212	Cyanophage	100	n/a	Gulf of Aqaba - Red Sea (Millard *et al* 200)
JCVI_SCAF_1101667034653	✓	212	Unknown	92	1018	GS007 - Northern Gulf of Maine (Rusch 2007)
JCVI_SCAF_1101669070555	✓	212	Unknown	92	959	GS007 - Northern Gulf of Maine (Rusch 2007)
JCVI_SCAF_1101668247417	Fragment too short	207	Unknown	73	1574	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1101669425113	Fragment too short	207	Unknown	73	1574	GS007 - Northern Gulf of Maine (Rusch 2007)
JCVI_SCAF_1101669142352	Fragment too short	263	Unknown	70	566	GS010 - Cape May, NJ - USA (Rusch 2007)
JCVI_SCAF_1101667044432	✓	263	Unknown	70	566	GS010 - Cape May, NJ - USA (Rusch 2007)
JCVI_SCAF_1098315327957	Fragment too short	259	Unknown	70	1804	MOVE858 - Chesapeake Bay, USA (Rusch 2007)
JCVI_SCAF_1101669414852	Fragment too short	241	Unknown	70	1522	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1101668234973	Fragment too short	241	Unknown	70	1522	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1097156666624	✓	204	Cyanophage	65	3436	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1096627283123	✓	204	Cyanophage	65	3437	GS002 - Gulf of Maine - Canada (Rusch 2007)
JCVI_SCAF_1097207205912	Fragment too short	204	Unknown	62	920	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1096626190549	Fragment too short	207	Unknown	62	920	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1096627024703	Fragment too short	236	Unknown	65	3221	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1096627666661	✓	164 min	Unknown		1508	GS020 - Lake Gatun -Panama (Rusch 2007)
JCVI_SCAF_1096627024160	✓	818*	Cyanophage	53	3755	GS020 - Lake Gatun -Panama (Rusch 2007)

				Presence of the		
	Scaffold	*psbA*	Intron detected in	“G/KETTXXXSQ/H”	Phylogenetic	
Scaffold Accession Number	Size (nt)	detected	*psbA* gene	motif in PsbA	Classification of *psbA*	Site of DNA isolation
JCVI_SCAF_1096627879849	1071	✓	Fragment too short	✓	not determined	GS031 - Upwelling, Fernandina Island
JCVI_SCAF_1101667164370	1026	✓	x	✓	not determined	GS020 - Lake Gatun
JCVI_SCAF_1101667171453	755	x	n/a	n/a	not determined	GS020 - Lake Gatun
JCVI_SCAF_1101667171626	732	x	n/a	n/a	n/a	GS020 - Lake Gatun
JCVI_SCAF_1101668541828	1657	✓	Fragment too short	Fragment too short	not determined	GS031 - Upwelling, Fernandina Island
JCVI_SCAF_1096626190549	920	✓	✓	Fragment too short	not determined	GS020 - Lake Gatun
JCVI_SCAF_1096626856934	3100	✓	x	✓	Cyanophage	GS003 - Browns Bank, Gulf of Maine
JCVI_SCAF_1096627283123	3436	✓	✓	✓	Cyanophage	GS002 - Gulf of Maine - Canada
JCVI_SCAF_1096627676525	1603	x	n/a	n/a	n/a	GS020 - Lake Gatun
JCVI_SCAF_1101668745121	1695	✓	x	✓	Cyanophage	GS047 - 201 miles from F. Polynesia
JCVI_SCAF_1097156666624	3436	✓	✓	✓	Cyanophage	GS020 - Lake Gatun -Panama
JCVI_SCAF_1096627021912	1947	✓	✓	✓	not determined	GS020 - Lake Gatun
JCVI_SCAF_1096627284644	2516	✓	x	✓	Cyanophage	GS002 - Gulf of Maine - Canada
JCVI_SCAF_1096627299009	3055	x	n/a	n/a	n/a	GS012 - Chesapeake Bay, MD
JCVI_SCAF_1096627313094	1528	✓	Fragment too short	✓	not determined	GS020 - Lake Gatun
JCVI_SCAF_1096627639930	1706	✓	Fragment too short	✓	Cyanophage	GS020 - Lake Gatun
JCVI_SCAF_1096627674162	1845	✓	n/a	n/a	n/a	GS020 - Lake Gatun
JCVI_SCAF_1096627675073	1700	✓	Fragment too short	✓	not determined	GS020 - Lake Gatun
JCVI_SCAF_1096627912725	1227	x	n/a	n/a	n/a	GS031 - Upwelling, Fernandina Island
JCVI_SCAF_1101668250692	1605	✓	x	x	Cyanophage	GS020 - Lake Gatun
JCVI_SCAF_1101668253187	1575	x	n/a	n/a	n/a	GS020 - Lake Gatun
JCVI_SCAF_1101668699797	1416	✓	Fragment too short	Fragment too short	not determined	GS035 - Wolf Island
JCVI_SCAF_1096627024160	3075	✓	✓	✓	Cyanophage	GS020 - Lake Gatun