environmental
microbiology

Environmental Microbiology (2015) 17(12), 5100–5108 doi:10.1111/1462-2920.13036

Closing the gaps on the viral photosystem-I *psaDCAB* gene organization

Sheila Roitman,[1†] José Flores-Uribe,[1†] Alon Philosof,[1] Ben Knowles,[2] Forest Rohwer,[2] J. Cesar Ignacio-Espinoza,[3] Matthew B. Sullivan,[4§] Francisco M. Cornejo-Castillo,[5] Pablo Sánchez,[5] Silvia G. Acinas,[5] Chris L. Dupont[6] and Oded Béjà[1*]

[1] *Faculty of Biology, Technion – Israel Institute of Technology, Haifa, Israel.*

[2] *Department of Biology, San Diego State University, San Diego, CA, USA.*

Departments of [3]Molecular and Cellular Biology and [4]Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA.

[5] *Departament of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), CSIC, Barcelona, Spain.*

[6] *Microbial and Environmental Genomics Group, J Craig Venter Institute, San Diego, CA, USA.*

Summary

Marine photosynthesis is largely driven by cyanobacteria, namely *Synechococcus* and *Prochlorococcus*. Genes encoding for photosystem (PS) I and II reaction centre proteins are found in cyanophages and are believed to increase their fitness. Two viral PSI gene arrangements are known, *psaJF→C→A→B→K→E→D* and *psaD→C→A→B*. The shared genes between these gene cassettes and their encoded proteins are distinguished by %G + C and protein sequence respectively. The data on the *psaD→C→A→B* gene organization were reported from only two partial gene cassettes coming from Global Ocean Sampling stations in the Pacific and Indian oceans. Now we have extended our search to 370 marine stations from six metagenomic projects. Genes corresponding to both PSI gene arrangements were detected in the Pacific, Indian and Atlantic oceans, confined to a strip along the equator (30°N and 30°S). In addition, we found that the predicted structure of the viral PsaA protein from the *psaD→C→A→B* organization contains a lumenal loop conserved in PsaA proteins from *Synechococcus*, but is completely absent in viral PsaA proteins from the *psaJF→C→A→B→K→E→D* gene organization and most *Prochlorococcus* strains. This may indicate a co-evolutionary scenario where cyanophages containing either of these gene organizations infect cyanobacterial ecotypes biogeographically restricted to the 30°N and 30°S equatorial strip.

Introduction

Cyanobacteria play a key role in oceanic photosynthesis and contribute to the global carbon cycle and oxygen supply (Li *et al*., 1993; Liu *et al*., 1997; Partensky *et al*., 1999). Genes encoding for photosystem-II (PSII) reaction centres (the D1 and D2 proteins encoded by the *psbA* and *psbD* genes, respectively) are found in cultured and uncultured phages that infect marine cyanobacteria (Mann *et al*., 2003; Lindell *et al*., 2004; 2005; Millard *et al*., 2004; Sullivan *et al*., 2005; 2006; Zeidner *et al*., 2005; Sharon *et al*., 2007), are expressed upon infection (Lindell *et al*., 2005; 2007; Clokie *et al*., 2006), and it was suggested that this increases phage fitness (Bragg and Chisholm, 2008; Hellweger, 2009). See Puxty and colleagues (2014) for a recent review on viral ‘photosynthesis’.

Using environmental metagenomics, uncultured cyanophages were recently found to contain gene cassettes coding for photosystem-I (PSI) reaction centres (Sharon *et al*., 2009). Two viral PSI gene organizations are currently known (Sharon *et al*., 2009; Béjà *et al*., 2012), *psaJF→C→A→B→K→E→D* and *psaD→C→A→B*. The *psaJF→C→A→B→K→E→D* cassette contains a gene fusion between the *psaJ* and *psaF* and is characterized by a low %G + C content of around 40%, while the four gene cassette, *psaD→C→A→B*, tends to have a higher %G + C content, ranging from 42% to over 50%. The fused PsaJF protein from the low %G + C cassette was hypothesized to be able to accept electrons not only from PSII (via plastocyanin or cytochrome c_6) but to also work with other electron donors like soluble cytochrome c that usually function as electron donors to cytochrome oxidase (Sharon *et al*., 2009). This was recently shown using a heterologous *Synechocystis*

Received 2 July, 2015; accepted 20 August, 2015. *For correspondence. E-mail: beja@tx.technion.ac.il; Tel. (+972) (4) 829 3961; Fax (+972) (4) 822 5153. †Equal contribution. §Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

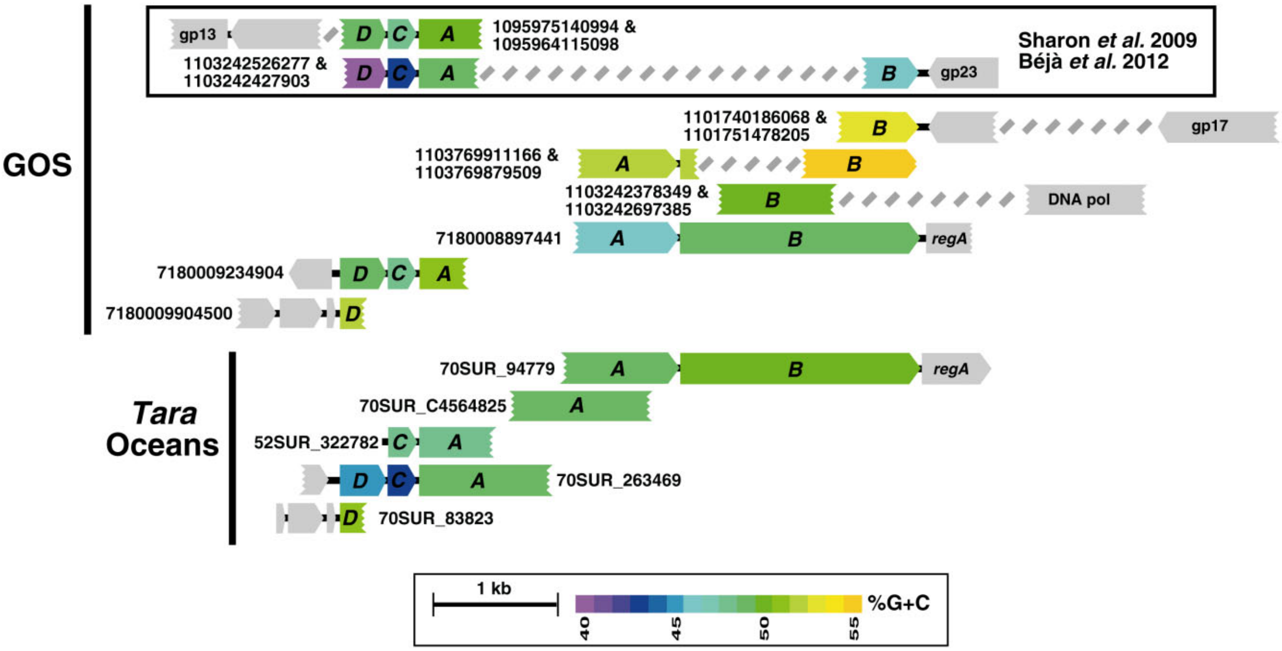

Fig. 1. Schematic gene organization of GOS and *Tara* Oceans scaffolds containing viral PSI genes from the *psaD→C→A→B* cassette. PSI genes are coloured according to their %G + C content; the calculation was performed on each gene separately. Grey boxes represent viral ORFs. Two GOS clones previously reported are boxed at the top. DNA sequences can be found in Appendix S1. gp23 – major capsid protein; gp17 – terminase large subunit; DNApol – DNA polymerase; *regA* – translation regulator. For clarity, not all detected scaffolds are shown.

system mimicking the cyanophage system (Mazor *et al*., 2014).

Our knowledge on the *psaD→C→A→B* gene cassette is based solely on two metagenomic scaffolds originating from the Global Ocean Sampling (GOS) expedition (Sharon *et al*., 2009; Béjà *et al*., 2012) (upper Fig. 1). These scaffolds contain a partial *psaD* gene, the highly conserved *psaC* gene, the beginning of *psaA*, and one clone also contains the far end of *psaB* gene. In addition, we have recently shown, using polymerase chain reaction (PCR) with primers amplifying the unique viral gene organization *psaC→psaA* (found in both gene cassettes), that *psaA* genes coming from the viral *psaD→C→A→B* cassette are diverse (Hevroni *et al*., 2015).

The goal of this study was to expand our knowledge regarding PSI genes carrying phages, with a special emphasis on phages carrying the *psaD→C→A→B* gene organization. We wanted to better understand why such two different viral PSI gene organizations exist, whether they are capable of different functions and who are the potential cyanobacterial hosts. For this 370 marine stations from six metagenomic projects were analysed for the presence of viral PSI genes. Numerous viral scaffolds were found from both PSI gene organizations, and the analysis of the scaffolds matching the *psaD→C→A→B* gene organization enabled us to close the gaps in missing parts of the gene sequences. In addition we were able to model the viral PsaA protein encoded by the *psaD→C→A→B* gene organization and to find substantial structural differences to its PsaA counterpart from the viral *psaJF→C→A→B→K→E→D* organization.

Results and discussion

To date, numerous low %G + C viral PSI cassette sequences have been identified (Sharon *et al*., 2009; Alperovitch-Lavy *et al*., 2011). In contrast, only two high %G + C viral PSI scaffolds are currently known (upper Fig. 1). To increase our understanding of the high %G + C viral PSI gene arrangement and fill up the sequence gaps in the *psaD→C→A→B* cassettes, we have examined the GOS (Rusch *et al*., 2007; Yooseph *et al*., 2007), Pacific Ocean Virome (POV) (Hurwitz and Sullivan, 2013), *Tara* Oceans (Karsenti *et al*., 2011; Brum *et al*., 2015; Sunagawa *et al*., 2015), C-MORE:BULA (Hewson *et al*., 2009), Moore Virome Project (The Gordon and Betty Moore Foundation Marine Microbial Initiative genomes), and Hawaii and Line Islands metagenomic datasets using the viral protein sequences of PsaD, PsaA and PsaB as queries.

Stations showing the presence of viral PSI genes (marked as red stations in Fig. 2) were mainly confined between 30°N and 30°S. It is important to remark that despite the differences between projects regarding sampling protocols, sequencing techniques, data analysis,

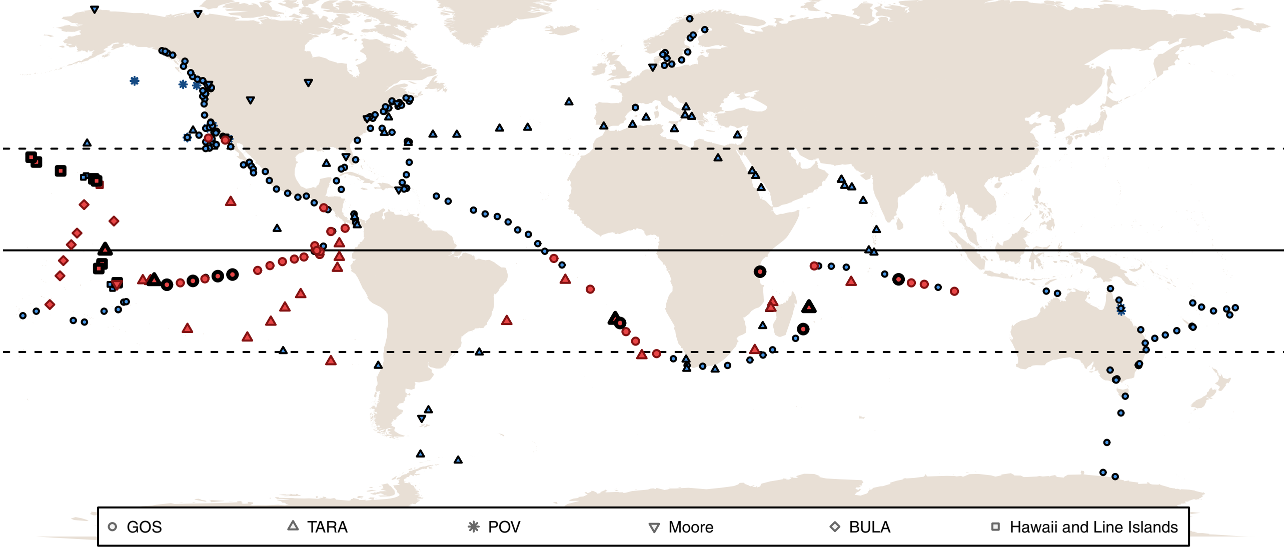

Fig. 2. Map of stations analysed for the presence of viral PSI genes. Blue symbols indicate stations where viral PSI was not detected. Red symbols indicate stations where at least one viral PSI scaffold or read were found; bold red stations were positive for high %G + C viral PSI presence. Stations are indicated as circles for GOS; up triangles, *Tara* Oceans; down triangles, Moore Virome Project; diamonds, C-MORE:BULA; asterisks, POV and squares, Hawaii and Line Islands. The equator is shown as a solid line, while latitudes 30°N and 30°S are shown as dashed lines.

etc., our results show that viral PSI gene cassettes are widespread. The presence of scaffolds matching the *psaD→C→A→B* cassette was observed in the Pacific and Indian oceans, and for the first time also detected in the Atlantic Ocean. To the same extent, it is worth noting the existence of different picocyanobacterial clades or ecotypes that occupy distinct environmental conditions (see Scanlan *et al.*, 2009, for a review). For instance, within the *Synechococcus* subcluster 5.1 (Dufresne *et al.*, 2008; Scanlan *et al.*, 2009), to which most marine *Synechococcus* belong, just a few clades such as clade II and III (Scanlan *et al.*, 2009) and clades CRD1 and CRD2 (Sohm *et al.*, 2015) have been reported between 30°N and 30°S. On the contrary, clades I or IV are found in coastal and/or temperate mesotrophic open ocean waters largely above 30°N and below 30°S. The same geographical restrictions are valid for *Prochlorococcus* clades such as the HLII clade occupying strongly stratified surface waters between 30°N and 30°S, or the contrary case clade HLI living in more weakly stratified surface waters, particularly between 35° and 48°N and 35° and 40°S, just to mention a few examples (Scanlan *et al.*, 2009). Therefore, it seems quite logical to assume that the distribution of cyanophages containing PSI-viral genes would fit with the distribution of their cyanobacterial hosts, and in this particular case we suggest that cyanophages containing PSI-viral genes are restricted to infect picocyanobacterial ecotypes living in the belt defined between 30°N and 30°S.

The newly discovered scaffolds from the high %G + C gene organization allowed us, for the first time, to construct PsaA, PsaD and PsaB phylogenetic trees containing more than one viral high %G + C entity. As previously observed (Béjà *et al.*, 2012), partial sequences of viral proteins from the high %G + C gene organization cluster together, within the marine *Synechococcus* clade, while partial proteins from the low %G + C gene organization group cluster separately forming their own clade. Having longer sequences originated from environmental scaffolds made it possible to construct full-length PsaB and PsaD phylogenetic trees (Fig. 3), based on all 756 and 193 amino acids positions of PsaB and PsaD respectively. These trees show a different topology to the previously reported, with both low and high %G + C gene organizations forming monophyletic clades outside of the *Prochlorococcus* and *Synechococcus* clades.

Some of the new high %G + C scaffolds from GOS and Tara Oceans contained several viral genes other than PSI genes. These genes (e.g. DNA polymerase or *regA* genes) resemble genes from cyanomyophages (T4-like phages; see Fig. S1), suggesting a possible *Myoviridae* origin for these scaffolds. In all viral scaffolds where sequences reaching beyond the photosynthetic gene arrangement borders were available, there were always viral genes constraining the arrangement (Fig. 1). This supports the notion that there are no other neighbouring photosynthetic genes accompanying the *psaD→C→A→B* gene arrangement. It is important to remark that the

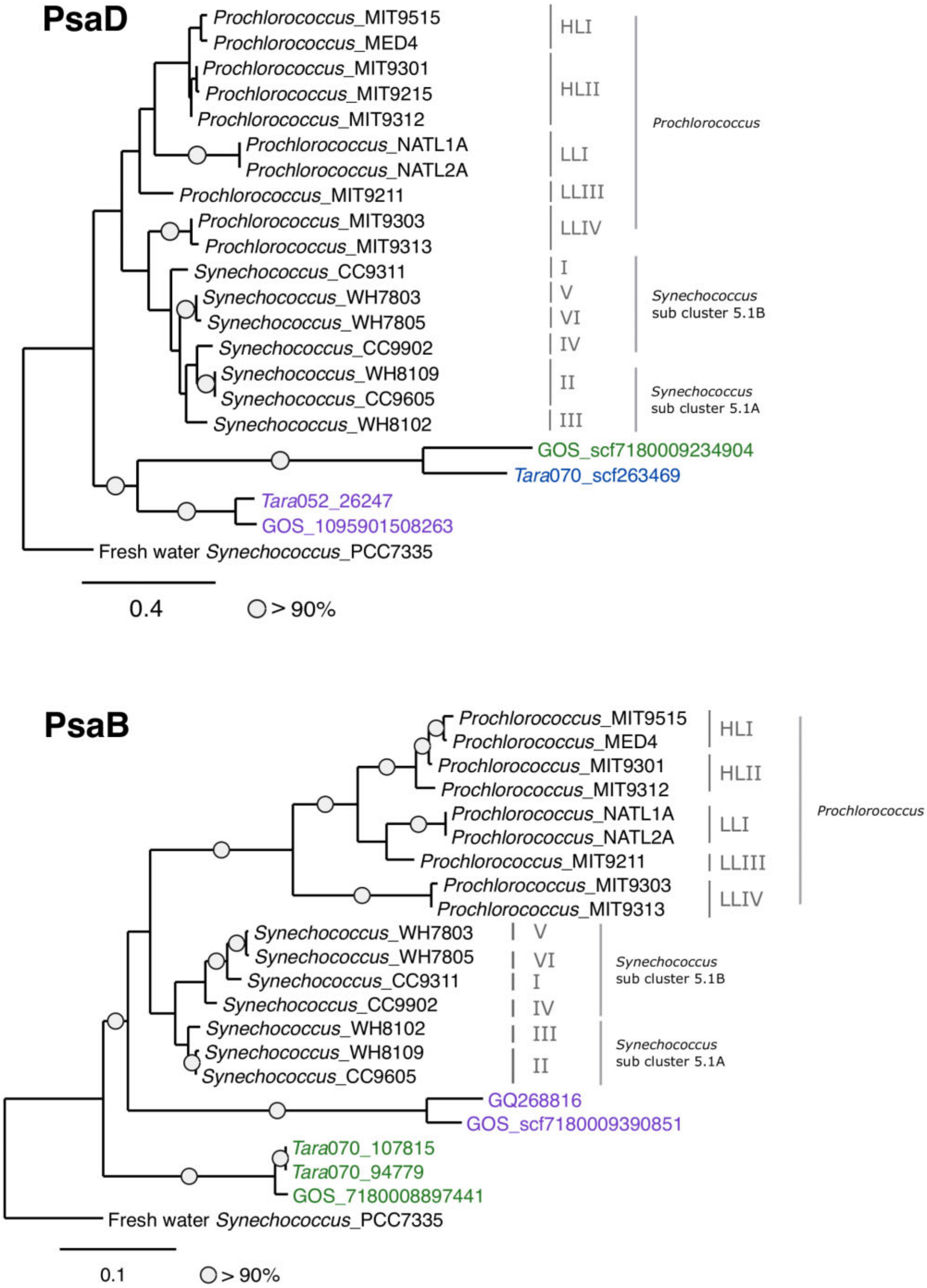

Fig. 3. Maximum likelihood phylogenetic trees of (A) PsaD – based on 193 amino acids positions, and (B) PsaB – based on 756 amino acids positions. Circles represent bootstrap values higher than 90%. Phage name colours represent %G + C classification according to the colour index in Fig. 1, purple stands for low %G + C sequences, green and blue for high %G + C. The scale bar indicates the average number of amino acid substitutions per site.

retrieved viral *psaD* sequences were always flanked by a non-photosynthetic viral open reading frame (ORF), either upstream or downstream in the high %G + C and low %G + C gene organizations respectively. The same occurs downstream to the high %G + C *psaB* and upstream the *psaJF* (which can be found only in low %G + C sequences). Furthermore, the sequences retrieved consistently match the *psaJF*→*C*→*A*→*B*→*K*→*E*→*D* or *psaD*→*C*→*A*→*B* gene organizations, accordingly to their %G + C content, which might indicate that viral PSI genes are found solely in the two previously described cassettes. However, as metagenomic data are fragmented by nature, we cannot rule out the presence of standalone PSI photosynthetic genes (e.g. the *psaJ* gene;

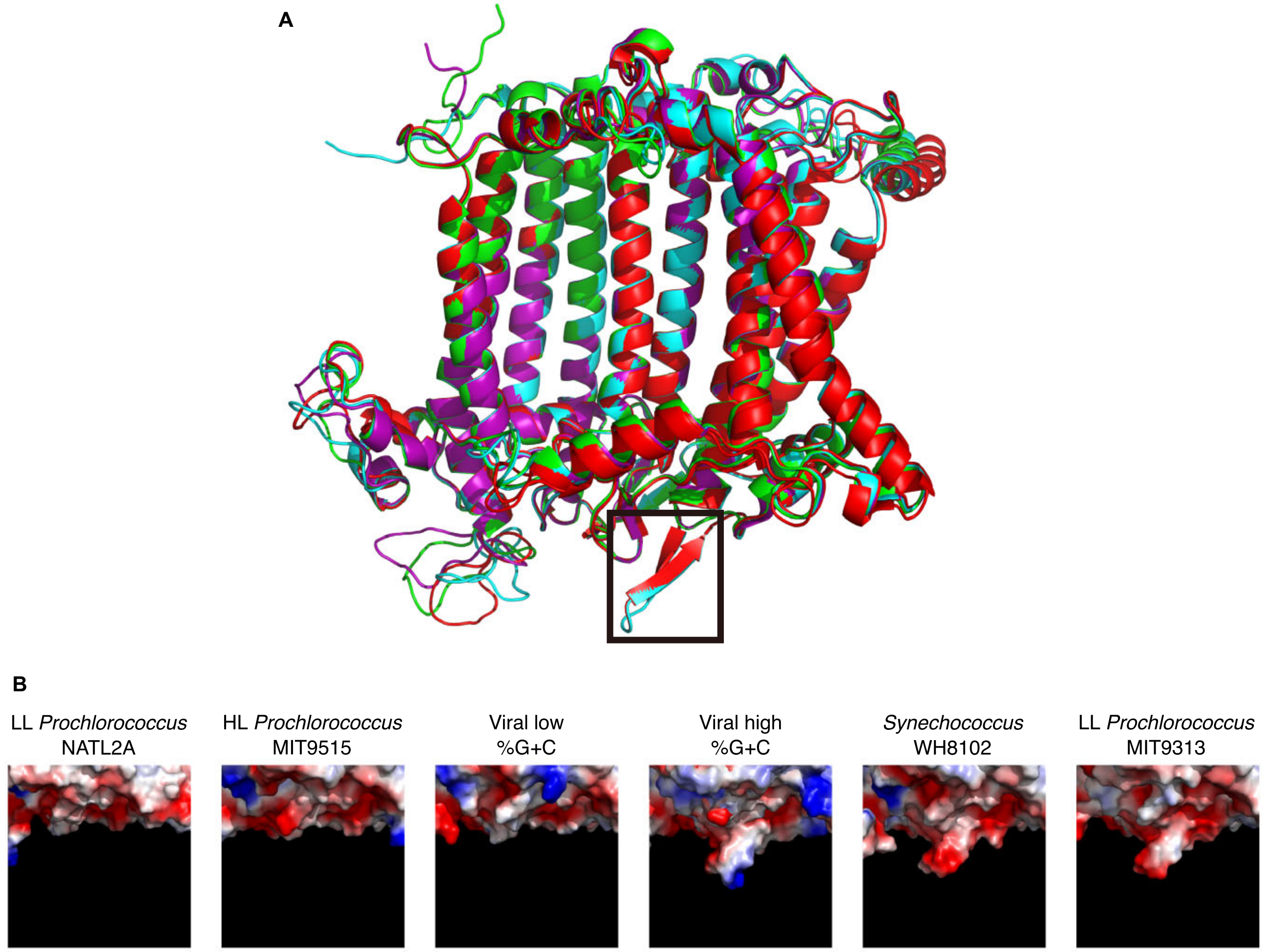
A
B
LL *Prochlorococcus* NATL2A
HL *Prochlorococcus* MIT9515
Viral low %G+C
Viral high %G+C
Synechococcus WH8102
LL *Prochlorococcus* MIT9313

Fig. 4. Structure modelling of PsaA proteins from cyanobacteria and cyanophages. (A) PsaA from *Synechococcus* (in cyan), HL *Prochlorococcus* (green), low %G + C viral (purple) and from the reconstructed high %G + C viral (red). The loop missing in PsaA from *Prochlorococcus* (except in LL *Prochlorococcus* MIT9313 and MIT9303) and the low %G + C viral, but present in *Synechococcus*, LL *Prochlorococcus* MIT9313 and MIT9303, and in the high %G + C viral PsaA is boxed. (B) Electrostatic potential of the lumenal side of PsaA proteins boxed in panel A. Red and blue indicate negative and positive potentials respectively. The loop sequences alignment can be found in Fig. S2.

Sharon *et al*., 2011) or other yet-unknown photosynthetic cassettes in the genomes of these cyanophages.

Based on data from the environmental scaffolds, we were able to assemble a full-length viral PsaA protein sequence from the high %G + C family by overlapping partial sequences. Structure prediction of the assembled viral PsaA was then compared with that of PsaA proteins from *Synechococcus*, *Prochlorococcus* and with the viral low %G + C. As shown in Fig. 4A, the overall structure of the four proteins is conserved except for a small loop (boxed in Fig. 4A) facing outside the membrane. This loop is in close proximity to the hydrophobic binding site of plastocyanin/cytochrome c_6 (Sommer *et al*., 2004; Mazor *et al*., 2012), therefore potentially influencing the electron transfer between the electron donor and P700 in PSI. To further confirm the existence of this loop, we designed a set of degenerate primers based on a PsaA protein's sequence alignment and used them to amplify the gene from viral concentrates collected from the Line Islands. Positive overlapping *psaA* PCR products were successfully retrieved (GenBank #s KP411049-KP411210), and their %G + C content was similar to viral low and high %G + C groups. This loop is found in the viral high %G +C PsaA, in *Synechococcus*, and also in low light adapted (LL) *Prochlorococcus* MIT9313 and MIT9303, and is missing in PsaA proteins of other LL *Prochlorococcus*, high light adapted (HL) *Prochlorococcus*, and in the viral low %G+C version. Interestingly, the high %G+C viral version of the loop is different from the marine cyanobacterial loop, containing a conserved arginine residue (Fig. S2); the viral loop is therefore positively charged compared with the cyanobacterial versions

(Fig. 4B). Interestingly, viral encoded plastocyanin (PetE) has a lower isoelectric point (and thus potentially being negatively charged) compared with the host plastocyanin (Puxty *et al*., 2014). We suggest that the viral PetE version may have a higher affinity for the viral version of PsaA, possibly because of the lumenal loop of the phage's PsaA. It is important to note that no *petE* genes were found on any of the viral PSI metagenomic scaffolds retrieved in this study. Alternatively, the change in the viral PsaA loop may perhaps have a similar function as the fused PsaJF protein found in the viral *psaJF→C→A→B→K→E→D* cassette, namely being promiscuous for its electron donors and being able to accept electrons from electron donors other than plastocyanin or cytochrome c_6 (Mazor *et al*., 2014).

We suspect that cyanophages carrying high %G + C PSI genes infect hosts with similar protein structures, leading to the assumption that phages carrying the *psaD→C→A→B* gene organization might infect cyanobacteria which have the lumenal loop in PsaA and have a similar geographical distribution. The hosts could potentially be *Prochlorococcus* MIT9313 and MIT9303 [belonging to an LL *Prochlorococcus* clade (clade LLIV)], which is considered to be closely related to *Synechococcus* and is a clade widely distributed within the 40°N to 35°S latitudinal range, largely restricted to the deep euphotic zone (Scanlan *et al*., 2009; Biller *et al*., 2015), or *Synechococcus* from clades CRD1 and CRD2 which are present in similar latitudes, 40°N to 30°S (Sohm *et al*., 2015).

The lumenal loop is missing in PsaA protein versions originating from the viral *psaJF→C→A→B→K→E→D* cassette. This could indicate that cyanophages containing the *psaJF→C→A→B→K→E→D* gene organization might infect different cyanobacterial hosts as compared with phages containing the *psaD→C→A→B* gene organization. Likewise, these cyanophages would only infect cyanobacterial species or ecotypes latitudinally restricted to the strip defined between 30°N and 30°S. However, we cannot rule out the possibility that these two kinds of phages infect the same hosts but perform differently to achieve a similar outcome, namely PsaJF and the lumenal loop in PsaA could perform similar functions regarding the docking of electron donors to PSI.

Our PsaA modelling suggests that the PsaA protein from the *psaD→C→A→B* gene organization might function differently from PsaA versions of the potential hosts and the other viral gene organization, presenting a new kind of PSI complex. Evolutionary studies regarding PSI proposed that a minimal complex composed by the PsaA, PsaB, PsaC and PsaD proteins could theoretically form a functional reaction centre (Nelson, 2011). Therefore, characterizing phages with the *psaD→C→A→B* gene set might shed light on PSI evolution and lead to a better understanding of PSI light reactions, as this might be the only extant case of a minimal, functional PSI that comprised only four subunits.

Experimental procedures

Metagenomic data analysis

Microbial and viral metagenomic datasets were downloaded from CAMERA (Seshadri *et al*., 2007), iMicrobe database (http://imicrobe.us) or MG-RAST (Meyer *et al*., 2008).

Microbial metagenomes from the GOS expedition project (Venter *et al*., 2004; Rusch *et al*., 2007), Hawaii and Line Islands, Biogeochemistry of the Upper Ocean: Latitudinal Assessment (C-MORE:BULA) project (Hewson *et al*., 2009), *Tara* Oceans expedition (Sunagawa *et al*., 2015), and viral metagenomes from the POV project (Hurwitz and Sullivan, 2013), Moore Virome project, and *Tara* Oceans expedition virome (Brum *et al*., 2015) were analysed using BLAST v2.2.28$^+$ tools (see Table S1 for the metagenomic datasets).

First a collection of amino acid sequences from low and high %G + C viral PSI genes *psaA* and *psaB* (Table S2) was used as query for a TBLASTN search (e-value 0.1) against the metagenomes. Contigs and reads matching PSI proteins and their paired-end mates were further screened using BLASTX (e-value 10e-10) against the NCBI non-redundant (nr) protein database to identify those that were likely to have a viral origin according to the top score hit of taxonomy assignment and presence of viral genes on the contig or read mate.

psaA amplification and cloning

Degenerate primers were designed against a PsaA protein multiple sequence alignment of a wide variety of organisms, including eukaryotes, prokaryotes and viral PsaA proteins obtained from GenBank (Primers TTTW[I/V]W_fwd, ACNACNACNTGGRTNTGGAA; HHIHAF_rev, RAANGC RTGDATRTGRTG; MPPY[P/A]Y_fwd, ATGCCNCCNTA YSCNTA; TTW[A/S]FF_rev, RAARAANBWCCANGTNGT; with 512, 192, 256 and 1536 degeneracy respectively). Two PCR reactions (Reaction B: TTTW[I/V]W_fwd – HHIHAF_rev; Reaction L: MPPY[P/A]Y_fwd – TTW[A/S]FF_rev) were performed directly on viral concentrates from the Pacific Southern Line Islands [collected in April 2009 and in November 2013 from Caroline island (Millennium Island) and in October 2013 from Vostok Island]. Viral concentrates were prepared according to Haas and colleagues (2014). The PCR reactions B and L were performed using BIO-X-ACT^{TM} Short mix (Bioline, London, UK), in a total volume of 30 μl containing 1 μl of phage concentrate as template, OptiBuffer (1×), 2 μM primers (each), 0.8 mM dNTPs, 2 mM $MgCl_2$ and 2.4 U BIO-X-ACT^{TM} Short DNA polymerase. The PCR conditions were the following: Reaction B – 95°C, 5 min; 40 cycles of 95°C, 30 s; 53°C, 30 s, 72°C, 100 s; and Reaction L – 95°C, 5 min; 40 cycles of 95°C, 30 s; 50°C, 30 s, 72°C, 70 s. Reaction L PCR was also performed using the Tiangen 2× Taq PCR MasterMix (Tiangen Biotech, Beijing), in a total volume of 25 μl containing 1 μl phage concentrate as template, 1.2 μM primers (each) and Master Mix (1×). The PCR amplification conditions were as previously described. The PCR products

(Reaction B – 1500 bp approximately; Reaction L – 1100 bp approximately) were cloned using the PCRII-TOPO vector (Invitrogen, San Diego, CA) according to the manufacturer's specifications and sequenced using Sanger sequencing (Macrogen Europe, Amsterdam, NL). Sequences retrieved were checked against published *psaC→A* viral sequences using k-mers analysis in the overlapping region between the amplicons (Hevroni *et al*., 2015).

Phylogenetic tree construction and analysis

PsaB and PsaD sequences from the *Tara* Oceans and GOS projects were obtained by translating the scaffold DNA sequence according to the correct open reading frame and aligned along with sequences from *Prochlorococcus* and *Synechococcus* (retrieved from GenBank). Multiple sequence alignments were constructed using ClustalX v2.1 (Larkin *et al*., 2007). Maximum likelihood phylogenetic trees were constructed using the phylogeny.fr pipeline (Dereeper *et al*., 2008), which included PhyML v3.0 (Guindon *et al*., 2010) and the WAG substitution model for amino acids (Whelan and Goldman, 2001). One hundred bootstrap replicates were conducted for each analysis. See Appendices S2–S7 for the alignments used to construct the trees.

PsaA protein structure models

Structural models for the viral, *Prochlorococcus* and *Synechococcus* PsaA proteins were predicted and folded according to the Protein Data Bank 1JB0 record (Jordan *et al*., 2001) using the HHPred software v2.0.16 (Soding *et al*., 2005) and Modeller v9.11 (Sali *et al*., 1995). Protein models were visualized and the electrostatic potential calculated using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.). For each protein model, we performed several protein predictions and selected one representative sequence (in bold) for Fig. 4 (*Synechococcus* WH7803, **WH8102**, WH7805, WH8109, CC9902, RS9917, RS9916, CC9311, RCC307; HL-*Prochlorococcus* CCMP1986, **MIT9515**, MIT9301; LL-*Prochlorococcus* NATL2A, MIT9313, MIT9211; low %GC viral **GQ268816**; high %GC viral *psaA* sequences from Fig. 1 and five different sequences of PCR reactions B and L, among them **KP411157.1** and **KP411207.1**).

Acknowledgements

This work was supported by funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA Grant Agreement No. 317184, European Commission ERC Advanced Grant No. 321647, and Israel Science Foundation Grant No. 580/10. We thank the Tara Oceans expedition (http://oceans.taraexpeditions.org) for data sharing used in our analyses. This is contribution number 30 of the Tara Oceans. The authors declare no conflict of interest.

References

Alperovitch-Lavy, A., Sharon, I., Rohwer, F., Aro, E.M., Milo, R., Nelson, N., and Béjà, O. (2011) Reconstructing a puzzle: existence of cyanophages containing both photosystem-I & photosystem-II gene-suites inferred from oceanic metagenomic datasets. *Environ Microbiol* **13:** 24–32.

Béjà, O., Fridman, S., and Glaser, F. (2012) Viral clones from the GOS expedition with an unusual photosystem-I gene cassette organization. *ISME J* **6:** 1617–1620.

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015) *Prochlorococcus:* the structure and function of collective diversity. *Nat Rev Microbiol* **13:** 13–27.

Bragg, J.G., and Chisholm, S.W. (2008) Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS ONE* **3:** e3550.

Brum, J., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., *et al*. (2015) Patterns and ecological drivers of ocean viral communities. *Science* **348:** 1261498.

Clokie, M.R.J., Shan, J., Bailey, S., Jia, Y., Krisch, H.M., West, S., and Mann, N.H. (2006) Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ Microbiol* **8:** 827–835.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., *et al*. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36:** W465–W469.

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.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:** 307–321.

Haas, A.F., Knowles, B., Lim, Y.W., McDole Somera, T., Kelly, L.W., Hatay, M., and Rohwer, F. (2014) Unraveling the unseen players in the ocean – a field guide to water chemistry and marine microbiology. *J Vis Exp* **5:** e52131.

Hellweger, F.L. (2009) Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. *Environ Microbiol* **11:** 1386–1394.

Hevroni, G., Enav, H., Rohwer, F., and Béjà, O. (2015) Diversity of viral photosystem-I *psaA* genes. *ISME J* **9:** 1892–1898.

Hewson, I., Paerl, R.W., Tripp, H.J., Zehr, J.P., and Karl, D.M. (2009) Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. *Limnol Oceanogr* **54:** 1981–1994.

Hurwitz, B.L., and Sullivan, M.B. (2013) The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS ONE* **8:** e57355.

Jordan, P., Fromme, P., Witt, H.T., Klukas, O., Saenger, W., and Krauss, N. (2001) Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution. *Nature* **411:** 909–917.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., *et al*. (2011) A holistic approach to marine eco-systems biology. *PLoS Biol* **9:** e1001177.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947–2948.

Li, W.K.W., Zohary, T., Yacobi, Y.Z., and Wood, A.M. (1993) Ultraphytoplankton in the eastern Mediterranean Sea – towards deriving phytoplankton biomass from flow cytometric measurements of abundance, fluorescence and light scatter. *Mar Ecol Prog Ser* **102:** 79–87.

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

Lindell, D., Jaffe, J.D., 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., 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.

Liu, H., Nolla, H.A., and Campbell, L. (1997) *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. *Aquat Microb Ecol* **12:** 39–47.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Bacterial photosynthesis genes in a virus. *Nature* **424:** 741.

Mazor, Y., Greenberg, I., Toporik, H., Béjà, O., and Nelson, N. (2012) The evolution of PSI in light of phage encoded reaction centers. *Philos Trans R Soc B Biol Sci* **367:** 3400–3405.

Mazor, Y., Nataf, D., Toporik, H., and Nelson, N. (2014) Crystal structures of virus-like photosystem I complexes from the mesophilic cyanobacterium *Synechocystis* PCC 6803. *eLife* **3:** e01496.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al*. (2008) The Metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386.

Millard, A., Clokie, M.R.J., 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.

Nelson, N. (2011) Photosystems and global effects of oxygenic photosynthesis. *Biochim Biophys Acta* **1807:** 856–863.

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

Puxty, R.J., Millard, A.D., Evans, D.J., and Scanlan, D.J. (2014) Shedding new light on viral photosynthesis. *Photosynth Res* doi:10.1007/s11120-014-0057-x.

Rusch, D.B., Halpern, A.L., Heidelberg, K.B., Sutton, G., Williamson, S.J., Yooseph, S., *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: I, the Northwest Atlantic through the eastern tropical Pacific. *PLoS Biol* **5:** e77.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins* **23:** 318–326.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., *et al*. (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73:** 249–299.

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

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

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:** 258–262.

Sharon, I., Battchikova, N., Aro, E.-M., Giglione, C., Meinnel, T., Glaser, F., *et al*. (2011) Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J* **5:** 1178–1190.

Soding, J., Biegert, A., and Lupas, A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33:** W244–W248.

Sohm, J.A., Ahlgren, N.A., Thomson, Z.J., Williams, C., Moffett, J.W., Saito, M.A., *et al*. (2015) Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* doi:10.1038/ismej.2015.115.

Sommer, F., Drepper, F., Haehnel, W., and Hippler, M. (2004) The hydrophobic recognition site formed by residues PsaA-Trp^{651} and PsaB-Trp^{627} of photosystem I in *Chlamydomonas reinhardtii* confers distinct selectivity for binding of plastocyanin and cytochrome c_6. *J Biol Chem* **279:** 20009–20017.

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

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

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., *et al*. (2015) Structure and function of the global ocean microbiome. *Science* **348:** 1261359.

Venter, J.C., Remington, K., Heidelberg, J., Halpern, A.L., Rusch, D., Eisen, J.A., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.

Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18:** 691–699.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., *et al*. (2007) The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol* **5:** e16.

Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., and Béjà, O. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* & *Synechococcus* via viral intermediates. *Environ Microbiol* **7:** 1505–1513.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Maximum likelihood phylogenetic trees of (A) DNApol, (B) gp17, (C) gp23 and (D) RegA. Circles represent bootstrap values higher than 90%. Phage sequences retrieved in this study are coloured in red. The scale bar indicates the average number of amino acid substitutions per site.

Fig. S2. Multiple sequence alignment of the loop area in partial PsaA proteins. The arginine conserved in high %G + C viral sequences is marked in blue. Conserved negative amino acids are coloured in red. Names of the viral sequences represent reads/scaffolds or PCR products retrieved in this study (except for GQ268816).

Table S1. Metagenomic datasets analysed.

Table S2. Sequences used as query for the TBLASTN analysis.

Appendix S1. DNA sequences from Fig. 1.

Appendix S2. Protein alignment used to construct the PsaD phylogenetic tree (Fig. 3A).

Appendix S3. Protein alignment used to construct the PsaB phylogenetic tree (Fig. 3B).

Appendix S4. Protein alignment used to construct the DNApol phylogenetic tree (Fig. S1A).

Appendix S5. Protein alignment used to construct the gp17 phylogenetic tree (Fig. S1B).

Appendix S6. Protein alignment used to construct the gp23 phylogenetic tree (Fig. S1C).

Appendix S7. Protein alignment used to construct the RegA phylogenetic tree (Fig. S1D).