The ISME Journal (2007) 1, 492–501

www.nature.com/ismej

ORIGINAL ARTICLE

Viral photosynthetic reaction center genes and transcripts in the marine environment

Itai Sharon[1,2,9], Shani Tzahor[1,3,9], Shannon Williamson[4,9], Michael Shmoish[5], Dikla Man-Aharonovich[1], Douglas B Rusch[4], Shibu Yooseph[4], Gil Zeidner[1], Susan S Golden[6], Shannon R Mackey[6], Noam Adir[7], Uri Weingart[8], David Horn[8], J Craig Venter[4], Yael Mandel-Gutfreund[1] and Oded Béjà[1]

[1]*Faculty of Biology, Technion—Israel Institute of Technology, Haifa, Israel;* [2]*Faculty of Computer Science, Technion—Israel Institute of Technology, Haifa, Israel;* [3]*Inter-Departmental Program for Biotechnology, Technion—Israel Institute of Technology, Haifa, Israel;* [4]*J Craig Venter Institute, Rockville, MD, USA;* [5]*The Lorry I Lokey Interdisciplinary Center for Life Sciences and Engineering, Technion—Israel Institute of Technology, Haifa, Israel;* [6]*Center for Research on Biological Clocks, Texas A&M University, College Station, TX, USA;* [7]*Schulich Faculty of Chemistry, Technion—Israel Institute of Technology, Haifa, Israel and* [8]*School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel*

Cyanobacteria of the genera *Synechococcus* and *Prochlorococcus* are important contributors to photosynthetic productivity in the open ocean. The discovery of genes (*psbA*, *psbD*) that encode key photosystem II proteins (D1, D2) in the genomes of phages that infect these cyanobacteria suggests new paradigms for the regulation, function and evolution of photosynthesis in the vast pelagic ecosystem. Reports on the prevalence and expression of phage photosynthesis genes, and evolutionary data showing a potential recombination of phage and host genes, suggest a model in which phage photosynthesis genes help support photosynthetic activity in their hosts during the infection process. Here, using metagenomic data in natural ocean samples, we show that about 60% of the *psbA* genes in surface water along the global ocean sampling transect are of phage origin, and that the phage genes are undergoing an independent selection for distinct D1 proteins. Furthermore, we show that different viral *psbA* genes are expressed in the environment.
The ISME Journal (2007) **1,** 492–501; doi:10.1038/ismej.2007.67; published online 9 August 2007
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: cyanobacteria; photosynthesis; *psbA*; virus; cyanophage; D1

Introduction

Synechococcus and *Prochlorococcus* are abundant in the marine environment and constitute the main prokaryotic component of picophytoplankton. Together, they contribute almost 50% of primary production in oligotrophic regions of the ocean (Waterbury *et al.*, 1986; Li, 1995; Partensky *et al.*, 1999). Their photosynthetic membrane contains two reaction centers, of which photosystem II (PSII) mediates the transfer of electrons and protons from water, the terminal electron donor, to the plastoquinone pool. The D1 and D2 proteins form the reaction center dimer of PSII that binds the primary electron donors and acceptors. Oxygenic photosynthesis genes were recently found in the genomes of cyanophages from two different viral families (*Myoviridae* and *Podoviridae*) (Mann *et al.*, 2003, 2005; Lindell *et al.*, 2004; Millard *et al.*, 2004; Sullivan *et al.*, 2005). A variety of viral photosynthesis genes were identified: for example, *psbA*, encoding the D1 protein; *psbD*, encoding the D2 protein; *hli* genes, encoding the HLIPs (high-light inducible proteins); and the *petE* and *petF* genes, encoding the photosynthetic electron transport proteins plastocyanin and ferredoxin, respectively (for recent review, see Clokie and Mann, 2006). Furthermore, the viral *psbA* and *psbD* genes have been readily detected in recent marine metagenomic projects (Venter *et al.*, 2004; Angly *et al.*, 2006; DeLong *et al.*, 2006). Analyses of *psbA* nucleotide sequences from environmental samples (Zeidner *et al.*, 2005) or viral culture collections (Sullivan *et al.*, 2006) show a significant separation between viral genes and those of their potential *Synechococcus* and *Prochlorococcus* hosts, suggesting that an exchange and reshuffling of *psbA* genes has occurred between *Synechococcus* and *Prochlorococcus* species via phage intermediates. To assay the

Correspondence: O Béjà, Faulty of Biology, Technion—Israel Institute of Technology, Haifa 32000, Israel.
E-mail: beja@tx.technion.ac.il
[9]These authors contributed equally to this work.
Received 9 May 2007 and accepted 3 July 2007; published online 9 August 2007

prevalence of phage photosynthesis genes and their relationship with cyanobacterial photosynthesis genes, we conducted a global survey of *psbA* genes from diverse oceanic environments.

Methods

GOS sample collection and library construction

A detailed description of the sampling sites included in this study is discussed in Rusch *et al.* (2007). Briefly, surface water samples (~200 l) were collected and analyzed using the S/V Sorcerer II between 8 August 2003 and 2 March 2004 from sites along a horizontal transect, starting from the North Atlantic near Halifax, Nova Scotia, and ending in the Equatorial Pacific. The microorganisms collected from the water were size-fractionated by serial filtration through 3.0, 0.8 and 0.1 μm membrane filters (Supor membrane disc filter, Pall Life Sciences, Ann Arbor, MI, USA). Filters were vacuum-sealed with 5 ml of sucrose lysis buffer (20 mM EDTA, 400 mM NaCl, 0.75 M Sucrose, 50 mM Tris-HCl, pH 8.0) and frozen to −20°C. Only data from the 0.8 to 0.1 size fraction filters are reported here. Detailed methods describing DNA isolation, library construction, template preparation, automated cycle sequencing and metagenomic assembly can be found in Rusch *et al.* (2007) and Venter *et al.* (2004).

Protein clustering and site quantity estimates

D1 proteins were extracted from the global ocean sampling (GOS) metagenomic data using a sequence similarity clustering approach that was recently used to investigate global protein space by Yooseph *et al.* (2007). In brief, protein sequences produced from the first six legs of the Global Ocean Expedition carried out so far were clustered with a non-redundant set of publicly available sequences within the NCBI-nr, NCBI Prokaryotic Genomes, TIGR Gene Indices and Ensemble data sets based on pairwise sequence similarity. Clustering was based on full-length sequences, rather than domains, and incorporated length-based thresholds to address fragmentary sequences, thereby minimizing the clustering of unrelated proteins. Clusters containing shadow open reading frames (ORFs) (that is, ORFs overlapping a predicted protein at the nucleotide level on either the same or opposite strand of the predicted protein) and clusters containing sequences that did not reflect signs of selective pressure were removed from the analysis (Yooseph *et al.*, 2007).

Cyanobacterial D1 motif assignment

Based on the protein alignment shown in Figure 1a, marine cyanobacterial D1 protein signature motif, $^{R}/_{K}$ETTEXXSQN was extracted. This motif was used in BLAST searches against the Sargasso Sea shotgun data using Gapped BLAST and PSI-BLAST programs (Altschul *et al.*, 1997). While retrieving the BLAST results, a new motif was detected that includes histidine (H) instead of glutamine (Q), and this motif was also included in the subsequent searches. Among the motifs retrieved from the Sargasso Sea was a GLD triplet, which deviates from the EXX triplet within the search motif and is also found in the cultured cyanobacterium *Prochlorococcus marinus* str. NATL1MIT and the cyanophage P-SSM2. The cyanobacterial D1 motif was therefore set to $^{R}/_{K}$ETTXXXS$^{Q}/_{H}$. Each of the individual amino-acid sequences retrieved by the motif was

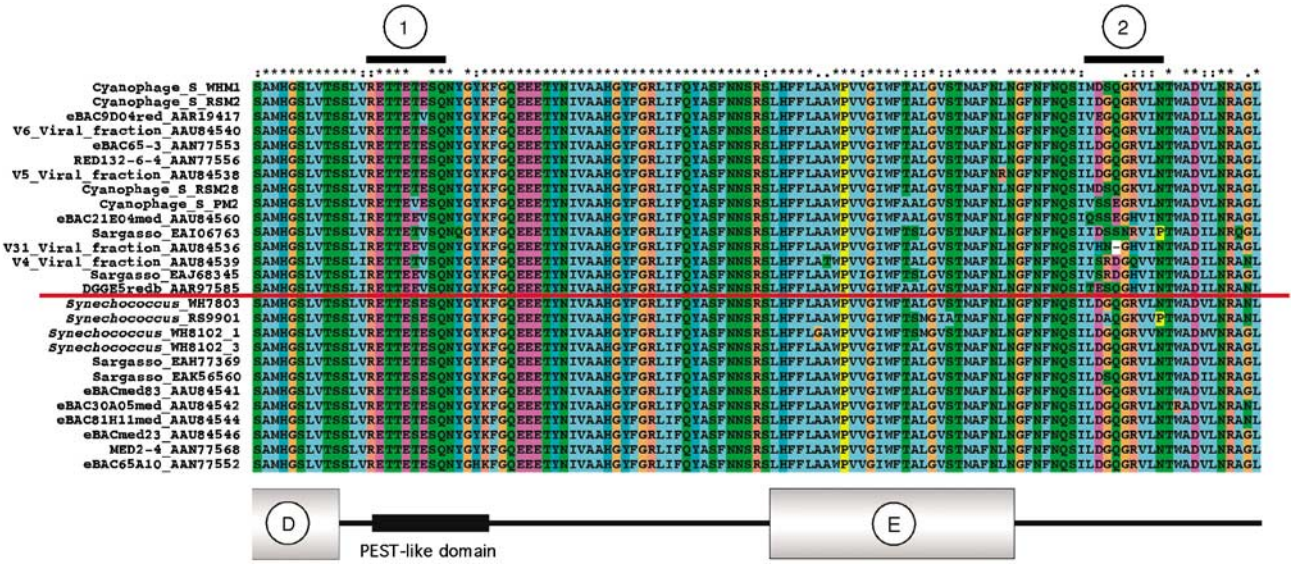

Figure 1 Multiple protein alignment of environmental *Synechococcus*-like viral and bacterial D1 proteins. The red line separates viral (top) and *Synechococcus* sequences (bottom). Thick lines above the alignment indicate variable identified regions. Transmembrane helices and the predicted PEST-like domain are indicated below the alignment. Circled numbers 1 and 2 indicate regions of variability in non-viral and viral-like sequences.

confirmed to represent a D1-like protein based on top BLAST hits to the non-redundant GenBank data set. Only D1 sequences related to *Prochlorococcus* or marine *Synechococcus* were considered in this study, while eukaryotic D1 or any other cyanobacterial D1 (only few in this GOS data set) were removed before any analyses performed in this study.

Selection criteria for taxonomic identification of scaffolds

All GOS D1-like proteins that contained the D1 protein signature motif $^{R}/_{K}$ETT$\underline{XXX}$S$^{Q}/_{H}$ were extracted and considered further. The taxonomic origin (viral or non-viral) of these proteins was predicted by analyzing the gene contents of the scaffolds on which they were identified. Only scaffolds of lengths 1 kb and longer containing at least one more gene in addition to the *psbA* gene were considered. Each gene on these scaffolds was labeled as viral or non-viral/unidentified based on a BLAST analysis against the October 2004 version of the GenBank non-redundant database with an expected cutoff value of 10. All scaffolds whose viral genes accounted for more than 75% of their overall annotated genes were classified as viral, with the rest being classified as non-viral. It is important to note that the non-viral group may contain some sequences originating from viral entities assigned in the GenBank as uncultured bacteria due to a lack of information.

The assembled GOS data of assigned viral or non-viral sequences is available in Supplementary Files S1 and S2, respectively. GOS environmental metadata, background on the GOS expedition, as well as additional analytical results may be further extracted via the Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) website (http://camera.calit2.net).

Statistics on viral signature motifs

All GOS D1 sequences containing the D1 protein signature motif $^{R}/_{K}$ETT$\underline{XXX}$S$^{Q}/_{H}$ were divided into clusters according to variable three-letter ($\underline{XXX}$) motifs: EAE, ENM and so on. To check our conjecture that a three-letter motif could discriminate between viral and non-viral clones, we performed a statistical analysis of the clones' distribution within the clusters. The analysis was performed only for the subset of D1 proteins whose origin was assigned a viral/non-viral origin. We could say that a cluster is 'virally enriched' if the proportion of clones of viral origin falling into that cluster exceeds the number expected under a random model. The degree of enrichment for a given cluster can be quantified by the hypergeometric distribution (Tavazoie *et al.*, 1999), according to which the probability of observing at least v viral clones in a cluster of size n by random chance is:

$$P = 1 - \sum_{i=0}^{v-1} \frac{\binom{V}{i}\binom{N-V}{n-i}}{\binom{N}{n}}$$

where N is the number of all 'long' clones and V is the number of viral 'long' clones. The P-value shows the significance of the enrichment v/n. Much in the same way, one could assess the opposite 'non-viral enrichment.' In Supplementary Table S3, the Hypergeometric Score field contains signed P-values, where the sign tells whether viral (plus) or non-viral (minus) clones are enriched in the given cluster. The small absolute values reflect the significance of enrichment.

PsbA phylogenetic DNA tree construction

GOS D1 amino-acid sequences were aligned using a modified version of CLUSTALw. Only columns that had at least 50% non-gaps were kept. From the resulting modified alignment, all sequences that had more than 25% gaps were removed. This resulting alignment was then used to construct a distance matrix using protdist (PHYLIP (Felsenstein, 2005)), and afterwards a neighbor-joining tree was constructed using a modified version of neighbor (PHYLIP) that did not allow negative branch lengths. One hundred bootstrap replicates were also constructed and the consensus tree was computed using PHYLIP. For the DNA tree, the DNA alignment was inferred from the protein alignment, and a similar filtering of columns and sequences was done as for proteins. The DNADIST program in PHYLIP was then used to compute a distance matrix. A neighbor-joining tree was constructed using the same process described above. One hundred bootstrap replicates were also constructed and their consensus tree was computed using PHYLIP.

Nucleotide frequencies analysis and principal components analysis

For each *psbA* sequence, frequencies of all possible oligonucleotides (four mononucleotides, 16 dinucleotides, 64 trinucleotides and 256 tetranucleotides) were counted. Oligonucleotide counts representing each sequence were normalized to the maximum length of the sequence. The normalized matrix including all sequences was used as input for the principal components analysis (Jolliffe, 1986). In the current study, the nucleotide frequencies were used as variables and transformed to new uncorrelated variables, namely the principal components. The data set was projected into a three-dimensional (3D) space defined by the first three principal components. Principal components analysis was conducted with the GNU R statistical software (http://www.gnu.org/software/r/R.html). *PsbA* sequences used for oligonucleotides frequencies analyses were from the GOS data and from cultured

Synechococcus WH8101, WH8020, CC9902, BL107, BL3, RS9901, RSS9907, CC9605, WH8017, WH8018, WH8102, WH8103, WH8012, WH8109 and WH7803; cultured *Synechococcus* myoviruses Syn30, SSHM2, SSSM5, SRSM2, SPM2, SRSM88, Syn1, SBM4, Syn19, Syn33, SRSM28, SWHM1, Syn10 and Syn9; environmental *Synechococcus* podovirus *e*BAC9D04; cultured *Prochlorococcus* MIT9515, MIT9302, MIT9312, MIT9116 and MED4; cultured *Prochlorococcus* myovirus PSSM10, PRSM4, PSSM9, PSSM3, PSSM2, PRSM5, PSSM4 and PSSM8; and cultured *Prochlorococcus* podovirus PSSP6, PSSP5, PSSP7, PGSP1 and PSSP3.

Cell collection and RNA extraction
Mediterranean Sea samples (32.10°N, 34.29°E) were collected in April 2005. Five to ten liters were prefiltered through a GF/A glass-fiber filter (approximate particle size $<1.6\,\mu$m) to remove larger eukaryotic phytoplankton cells and collected on a $0.2\,\mu$m sterivex filter. Nucleic acids were extracted from the samples according to Massana *et al.* (1997). DNA was removed using RNase-Free DNase I (Ambion, Cambridgeshire, UK) for 30 min at 37°C. DNase was inactivated by heat denaturation at 75°C for 10 min. For reverse transcription PCR (RT-PCR), total RNA (100–300 ng) was reverse-transcribed with *psbA* general degenerate reverse primer 331-MHER NAHNFP-340 (5′-GGRAARTTRTGNGCRTTNCKYT CRTGC-3′) using PowerScript Reverse Transcriptase (Clontech, Palo-Alto, CA, USA) according to the manufacturer's instructions. Reaction mixtures were incubated at 42°C for 1 h. Primers used to amplify *psbA* cDNA from environmental samples were 58-VDIDGIREP-66 (5′-GTNGAYATHGAYGGNATHMGN GARCC-3′) and 331-MHERNAHNFP-340 (Zeidner *et al.*, 2003). PCR amplification was carried out in a total volume of $25\,\mu$l containing 10 ng of template DNA, $200\,\mu$M dNTPs, 1.5 mM $MgCl_2$, $0.2\,\mu$M primers and 2.5 U of BIO-X-ACT DNA polymerase (Bioline, London, UK). The amplification conditions comprised steps at 92°C for 4 min, 45 cycles of 92°C for 1 min, 45°C for 1 min and 68°C for 1 min. PCR-amplified *psbA* was ligated into the pGEM cloning vector (Promega, Madison, WI, USA). All of the tests for the presence of contaminating DNA in the RNA samples or in the reagents were carried out according to Man-Aharonovich *et al.* (2007). To show that the RT-PCR did not generate chimeric sequences that might have provided a false v-*psbA* signature triplet in an otherwise host-derived gene, we performed GC plot analyses on the cDNA sequences (Supplementary Figure S2). No chimeras were observed (at least for the *Synechococcus*-like sequences) since no amplicons-containing chimeras between *Synechococcus* (high GC) and their viruses (moderate GC content) were observed. The mosaic nature of *Synechococcus* viral *psbA* sequences was reported previously by Zeidner *et al.* (2005) and Sullivan *et al.* (2006) but in the case of artificial chimeras obtained during PCR, the chimeras are not expected to have such an advanced mosaic appearance but rather to show 50/50 or 25/75 ratios. We therefore believe that no chimeras were produced in this preliminary experiment. In addition, control amplifications of known *psbA* sequences (*Synechococcus* WH7803 and BAC9D04) mixed in a 1:1 ratio produced no chimeras ($n=26$).

Accession numbers
The D1 cDNA sequences reported in this study were deposited in the GenBank with accession numbers DQ401466–DQ401504.

Results and discussion

We analyzed aligned sequences of deduced marine Mediterranean and Red Seas environmental *Synechococcus*-like D1 proteins and determined that two regions in the viral-like D1 (v-D1) proteins (Figure 1) deviate from the classical cyanobacterial sequence whereas the rest of the protein is generally conserved. The first region of variability lies within the PEST-like domain located in the loop between transmembrane helices D and E (region 1 in Figure 1). This loop has been implicated as the site of initial cleavage in the D1 protein that initiates protein turnover (Greenberg *et al.*, 1987). D1 is notorious for its rapid turnover and the requirement for its replacement to enable the sustained functioning of PSII (Chow and Aro, 2005). The second region of v-D1, which differs from cyanobacterial sequences, is a stretch of variable residues (region 2 in Figure 1) that follows the end of transmembrane helix E and precedes the C-terminal residues of mature cyanobacterial D1; the mature end is formed by the cleavage of the last 16 carboxy-terminal residues from the nascent protein and participates in binding the Mn^{2+} ions of the oxygen-evolving complex (Nixon *et al.*, 1992; Debus, 2001). The crystal structure of PSII from *Thermosynechococcus elongatus* (Ferreira *et al.*, 2004) suggests that this loop comes into contact with the PsbEF (cyt. b_{559}), PsbJ and PsbV (cyt. C-550) proteins, and is likely important for complex assembly (Suorsa *et al.*, 2004). Thus, potential v-D1 proteins represent a unique form of this critical photosynthetic reaction center protein, which may effect the stability and/or assembly of PSII (for a detailed review, see Adir *et al.*, 2003 and Edelman and Mattoo, 2006).

v-D1 sequences in the global ocean sampling expedition project
We tested the generality of the novel v-D1 proteins from this initial environmental Mediterranean and Red Seas data set by exploring the GOS expedition shotgun sequencing project (Rusch *et al.*, 2007; Yooseph *et al.*, 2007), which also includes within the Sargasso metagenome (Venter *et al.*, 2004). The

GOS project produced a total of 7.7 million random sequence reads, yielding approximately 6.3 Gb of assembled environmental DNA sequences from such diverse environments as the North Atlantic Ocean, the Panama Canal and the South Pacific Gyre (Rusch *et al.*, 2007). This data set was already mined for the presence of anoxygenic photosynthesis genes ('non-cyanobacterial') in search for diversity and biogeography of anoxygenic phototrophic bacteria (Yutin *et al.*, 2007).

First, by using protein alignments and BLAST searches (Altschul *et al.*, 1997) against the GenBank data set, we extracted a general marine cyanobacterial D1 protein signature motif ($^{R}/_{K}$ETT$\underline{\text{XXX}}$S$^{Q}/_{H}$) that maps to the variable loop located between helices D and E (region 1 in Figure 1). Eight triplet signatures were identified at this stage: ETE, ETV, EVE, EEV, ESE, EDE, ENM and EQE. We then identified 1407 *Synechococcus*-like and *Prochlorococcus*-like D1 sequences in the GOS data set using a clustering algorithm (see Methods). Of these, 848 contained the conserved cyanobacterial D1 search motif; the remaining D1 sequences were truncated to different extents and were not considered further. Twenty-two triplet signatures, not previously identified in the cyanobacterial D1 search motif $^{R}/_{K}$ETT$\underline{\text{XXX}}$S$^{Q}/_{H}$, were identified in the GOS data set (END, ENE, EDV, EDI, EAE, EDM, EEE, EQV, ENV, EEI, EIE, ETI, ELE, ETK, VDE, GLV, GLT, GLE, GLD, DNE, DEV and DDV). Of the 848 D1 scaffolds that contain the searched motif, 570 carry at least one ORF in addition to *psbA*. Each of these 570 scaffolds were assigned a 'viral-like' or 'non-viral' identity by the predicted products of ORFs that neighbor *psbA* (see Methods; assigned-scaffold sequences could be found in FASTA Supplementary Files S1 and S2). Based on the annotation of neighboring proteins, 62.1% of these scaffolds (354) were classified as viral-like and the rest (216) were classified as non-viral (bacterial and unassigned). A 'sanity-check' of the assignment was achieved by analyzing GC content plots (Zeidner *et al.*, 2005) of all *psbA* sequences assigned to the viral/non-viral categories (Supplementary Figure S1). Sequences in the viral assignment show a generally homogenous appearance of plots with an absence of high-GC containing DNA sequences, whereas the non-viral assigned sequences show a greater diversity of plots that can be sorted to high- versus low-GC containing DNA sequences. The absence of high-GC plots from the viral file points toward the absence of 'contaminating' high-GC-containing marine *Synechococcus* sequences in this file, and hence increases the reliability of our automated sorting. It is important to note that only data from the 0.8 to 0.1 size fraction filters (see Methods for details) are reported here. An estimation of the exact quantity of different cyanobacterial groups and phages could not be made since a fraction of marine *Synechococcus* cells might be trapped on the 0.8 μm filter, and free viruses may be retained on the 0.1 μm filters when they become clogged. In addition, the exact identity of the viral entities could not be assigned; they could be prophages or free phage particles, or could emanate from phage particles within bacterial cells about to be lysed. Nevertheless, when individual stations (see Supplementary Table S1 for a description of all sampling locations; a more comprehensive description of all sampling locations and their accompanying physio-chemical parameters can be found in Table 1 of Rusch *et al.*, 2007) from the GOS project were examined, some contained only viral-*psbA* (v-*psbA*) or only non-viral *psbA* sequences, whereas most exhibited variable v-*psbA*/non-viral *psbA* ratios of 3–80% (Supplementary Table S2). The different v-D1 ratios observed in this GOS snapshot-view probably reflect complex viral-cyanobacterial dynamics that exist in time and space during host infection.

A comparison between the $^{R}/_{K}$ETT$\underline{\text{XXX}}$S$^{Q}/_{H}$ motifs and the viral-assigned GOS scaffolds (Figure 2 and Supplementary Table S3) indicated that several statistically significant triplet signatures (EQE, ENE, EEV, EEE, EDV, EVE, EQV and EDE) are preferentially found in v-D1 proteins. These results also increase the credibility of our binning method for assigning 'viral-like' versus 'non-viral' notations, as the identified motifs were found almost exclusively in the 'viral-like' file (statistically significant based on hypergeometric distribution; see Methods and Supplementary Table S3). Interestingly, some of these viral-specific motifs, as well as others in other regions, were also identified independently by the unsupervised *de novo* motif extraction (MEX) algorithm (Solan *et al.*, 2005; Kunik *et al.*, 2007; a list of different viral-specific D1 peptides identified by the MEX algorithm is presented in Supplementary File S3).

V-*psbA* sequences could also be separated from the bacterial *psbA* genes based on nucleotide linguistics (Mrazek and Karlin, 2007; Figures 3a and b); we searched for the distribution of the predicted v-D1 triplet signatures (EQE, ENE, EEV, EEE, EDV, EVE, EQV and EDE) in the different groups separated by the principal components analysis (see Methods). These v-D1 motifs could also be confirmed based on the principal components analysis separation, as they were preferentially found on viral-assigned *psbA*s (Figures 3c and d).

The principal components analysis could further divide the viral *psbA*s into two groups, the myoviruses and the podoviruses. Interestingly, our analyses predict that some v-*psbA* belong to yet uncultured *Synechococcus* podoviruses since environmental BAC clone BAC9D04 (a predicted podovirus based on neighboring proteins (Zeidner *et al.*, 2005; Sullivan *et al.*, 2006) falls within these GOS *psbAs*. While both *Prochlorococcus* myoviruses and podoviruses-containing *psbA* sequences are frequently found, only *Synechococcus* myoviruses containing *psbA* sequences were reported previously (Millard *et al.*, 2004; Sullivan *et al.*, 2006).

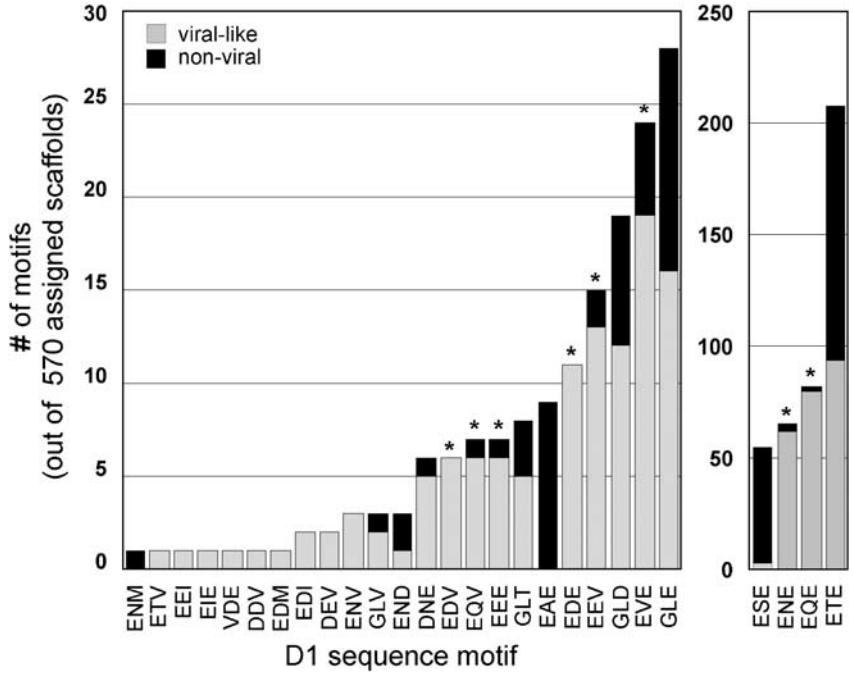

Figure 2 Occurrence of different D1 $^{R}/_{K}$ETTXXXS$^{Q}/_{H}$ motifs in viral-like and non-viral scaffold sequences (out of total number of 570 assigned scaffolds) from the GOS database. Asterisks denote statistically significant viral motifs (based on hypergeometric distribution values; see Supplementary Table S3). GOS, global ocean sampling.

In addition, several *Synechococcus* podoviruses that do not contain v-*psbA* genes have been reported (P60 (Chen and Lu, 2002), and syn12 and syn5 (Sullivan *et al*., 2006)). Our results stress the need for further efforts to isolate *Synechococcus* podoviruses that contain *psbA* sequences.

Prevalence of v-D1 triplet motifs in different parts of the world's oceans

The D1 motifs in the GOS metagenomic data set were mainly dominated by four triplet motifs (Figure 2 and Supplementary Table S3). Motif ETE (shared by both viruses and bacteria) appeared in more than 40% of the cases, viral motifs ENE and EQE occurred in 16% and 17% of the motifs, respectively, and bacterial motif ESE was present in 10% of the cases.

A comparison of motif distribution at different stations (Figure 4) revealed some interesting trends. While most stations showed a somewhat similar pattern dominated by the ETE motif, hypersaline station GOS33, fresh water station GOS20, Pacific warm seep station GOS30 as well as Pacific open ocean stations GOS37 and GOS47 and Pacific coastal stations GOS31 and GOS35 showed a different patterning. Station GOS33 had more occurrences of the ESE and EAE motifs (both non-viral motifs) than almost all other samples while the ETE motif was completely missing. The ETE motif was also missing in station GOS20, which had many ENE and ESE motifs. Warm seep station GOS30 and open ocean stations GOS37 and GOS47 were dominated with the GLE motif and coastal stations GOS31 and GOS35 mainly with the GLD motif. The clustering of North Atlantic stations GOS2, GOS3 and GOS7 was also interesting, containing the EQV motif in addition to the ETE motif and almost lacking all other motifs. The clustering itself seems to carry geographical characteristics; with a few exceptions, major clusters were created for tropical stations from both sides of the Panama Canal, Pacific Ocean stations, tropical Galapagos stations, temperate northern stations and Sargasso stations. Stations GOS20 and GOS33 were clustered separately from the rest of the stations, probably due to their unique environmental characteristics. Similar analyses performed on the genomic data set of scaffolds from the microbial community at the Pacific ALOHA station (DeLong *et al*., 2006) gave similar results where the same motifs dominating the GOS data set were also dominating the ALOHA data set and the majority (69%) of ALOHA D1 sequences were confirmed to be of viral origin (these analyses as well as analyses of the virome metagenome data set (Angly *et al*., 2006) are presented in Supplementary File S4).

Environmental v-psbA transcripts

To test if any *psbA* transcripts in the marine environment are of viral origin, we sought to refine the D1 search sequence to use it as a reliable identifier of the lineage of a given *psbA* transcript (region 1 in Figure 1). A comparison of variations in

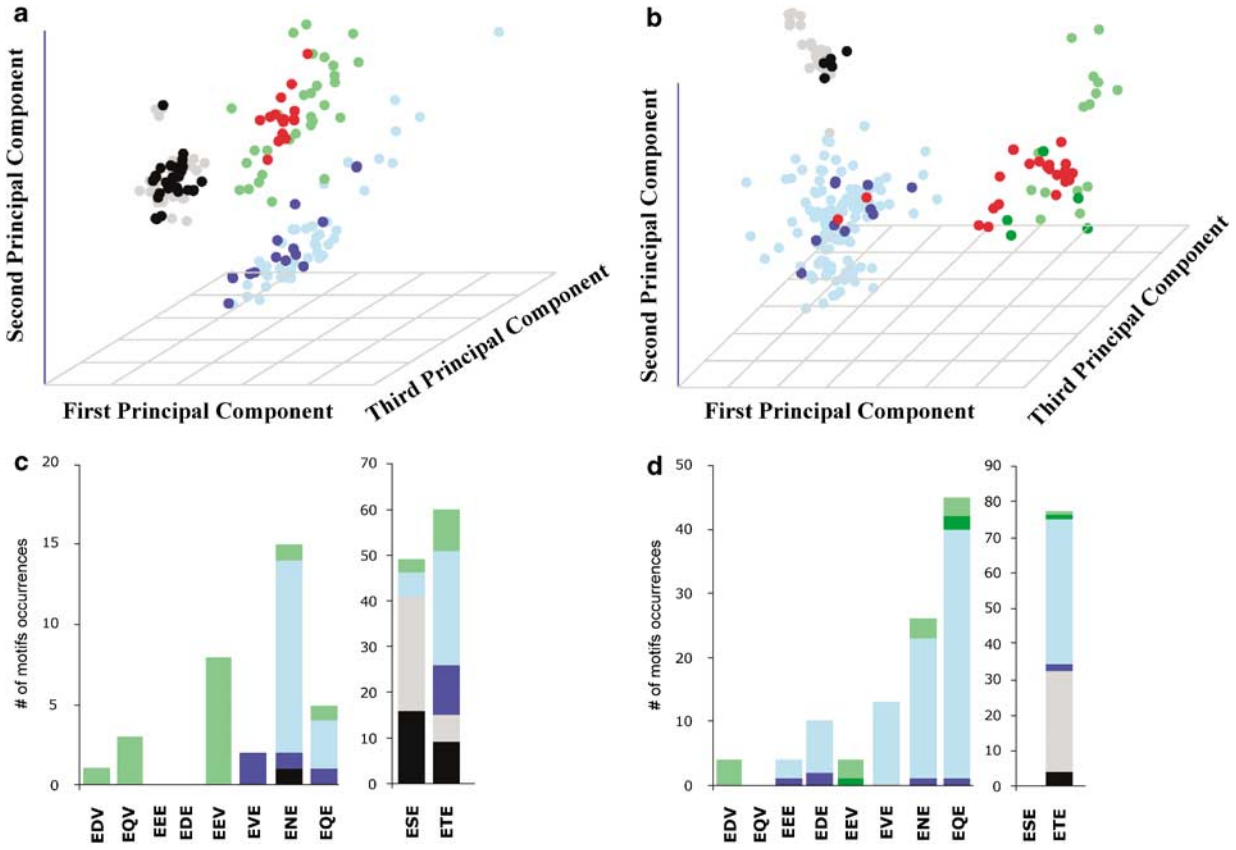

Figure 3 Principal components analysis of *psbA* sequences from cultured myovirus, podovirus, *Synechococcus*, *Prochlorococcus* and environmental GOS sequences. (**a**) Sequences from cultured *Synechococcus* (black) and *Synechococcus* myovirus (blue); corresponding environmental GOS sequences are presented in light colors (gray, light blue, light green for *Synechococcus*, myovirus and podovirus, respectively), environmental cDNA are shown in red. (**b**) Sequences from cultured *Prochlorococcus* (black), *Prochlorococcus* myovirus (blue) and *Prochlorococcus* podovirus (green); the corresponding environmental GOS sequences are in light colors (gray, light blue and light green for *Prochlorococcus*, myovirus and podovirus, respectively), environmental cDNA are shown in red. The graphs show the data projected onto the first three principal components obtained by the principal components analysis. Normalized frequencies of oligonucleotides (four mononucleotides, 16 dinucleotides, 64 trinucleotides and 256 tetranucleotides) were used as variables for principal components analysis input. Group colors were assigned based on a *psbA* DNA phylogenetic tree (see Methods). Exact 3D coordinates for each sequence are shown in Supplementary Tables S4 and S5. (**c**) Occurrence of different viral D1 $^{R}/_{K}$ETTXXXS$^{Q}/_{H}$ motifs in the different *Synechococcus*-like *psbA* groups, colored as in (**a**) non-viral motifs ETE and ESE are shown as negative controls. (**d**) Occurrence of different viral D1 $^{R}/_{K}$ETTXXXS$^{Q}/_{H}$ motifs in the different *Prochlorococcus*-like *psbA* groups, colored as in (**b**) non-viral motifs ETE and ESE are shown as negative controls. GOS, global ocean sampling.

the search motif from the GOS data set with recently released D1 protein sequences from both marine cyanophage isolates and their *Prochlorococcus* and *Synechococcus* hosts (Sullivan *et al.*, 2006) showed that EQE and EVE signatures in the variable positions of the D1 search sequence are reliable indicators of products derived from v-*psbA* for *Synechococcus*-like sequences, and EQE, EDV, EDE, EEV and ENE for *Prochlorococcus*-like sequences. We then used RT-PCR to identify *psbA* messenger RNAs directly from seawaters and evaluated their sequences for refined viral signatures. Transcripts from both *Prochlorococcus*- and *Synechococcus*-like *psbA* genes were detected, and different triplet motifs were identified, including motifs specifically predictive of viral sequences (ENE, EQE, EDV and EEV) and some that may be encoded by genes from either source (Figures 3e, f and 5). One of the triplets associated with v-D1 in this study, ENE, is found in *Synechococcus* CC9311, *Anabaena variabilis*, *Nostoc* PCC7120l, plants and many algae. However, top standard DNA–DNA BLAST (BLASTn) hits with all cDNA containing this motif (cDNA1, cDNA6 and cDNA20) and DNA linguistic analyses (red dots in Figures 3a and b) clearly points to a viral origin. Taken together with the recent finding of a viral-specific D1 peptide expressed during infection of its *Prochlorococcus* host (Lindell *et al.*, 2005), our results indicate that v-*psbA* genes are actively expressed in the marine environment and represent a gene pool that encodes for diverse v-D1 proteins with the potential to contribute to photosynthetic complexes.

Potential ecological implications

Phylogenetic analysis indicates that both phage and host *psbA* genes are under purifying selection

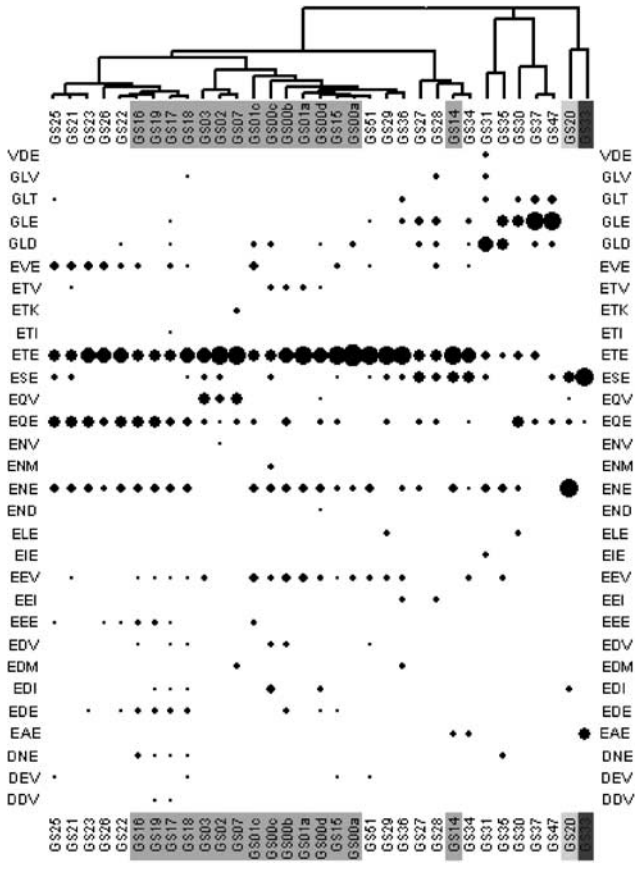

Figure 4 Presence and abundance of various D1 $^{R}/_{K}$ETT$\underline{\text{XXX}}$S$^{Q}/_{H}$ motifs in different GOS stations. The relative quantity of various motifs (rows) in each station (columns) is represented by the area of the corresponding spot (if any). Relative quantity is based on the share of each motif from the total number of motifs in a given station. Stations were clustered based on the fingerprint of motifs in the different stations, using hierarchical clustering with the Pearson correlation coefficient as the similarity measure. Station numbering is according to Rusch *et al.* (2007). Atlantic Ocean station names are labeled with a gray background, hypersaline station GOS33 is labeled with a dark gray background and freshwater station GOS20 is marked with light gray.

(Zeidner *et al.*, 2005; Sullivan *et al.*, 2006), and that the two classes of genes strongly favor different signatures in the searched motif (Figure 2). If v-D1 and host D1 are fully functional in the PSII reaction center, why would there be a selection for different variants in the two lineages, particularly when there is evidence of recombination at the nucleotide level over evolutionary time? Sullivan *et al.* (2006) suggested that v-D1 sequences of *Synechococcus*-like phages are more similar to Form II of D1, which is known in freshwater strains to be a high-light and stress-induced variant, than to 'housekeeping' Form I encoded by the *Synechococcus* species (Schaefer and Golden, 1989; Clarke *et al.*, 1993). We propose that like host Form II, v-D1 is less susceptible to photodamage and turnover than the host D1, a feature that could be important for 'life support' during the short period of phage morphogenesis. Conversely, host Form I is more adaptive to long-term sustainable metabolism in live cells where protein turnover is not a severe limitation. This implies that the catalytic role of marine viruses in the oceanic biochemical cycle is more complex than simply channeling the movement of nutrients from microorganisms to the dissolved and particulate organic matter pools via cell lysis (for a review see Suttle, 2005). If cyanophages do participate (in the form of prophages integrated into bacterial genomes (Brussow and Hendrix, 2002), as linear plasmid prophages (Casjens *et al.*, 2004) or as lytic phages), to a certain extent, directly in photosynthesis via shunting their modified D1 proteins into cyanobacterial photosynthetic apparatuses, then they also have a direct role in primary production. In addition, if these 'photosynthetic' viruses maintain photosynthesis during infection of cyanobacterial host cells in the environment (as shown in cultures by Lindell *et al.*, 2005; Clokie *et al.*, 2006) then there is a net gain of several hours of light energy, which would otherwise be lost. The fact that some viruses invest in modifying the D1 proteins suggest an adaptive role for this function, perhaps facilitating adaptation to harsh light conditions or modifying the D1 role for their selfish benefits (a more stable D1 that is functional only for the short time of infection). It is too early to draw global conclusions as the extent of this phenomenon in oceanic waters is not yet known, nevertheless, the fact that most marine cyanophage isolates carry the *psbA* gene (Sullivan *et al.*, 2006) and that certain viruses select for modified D1 proteins, point to an important role for viral 'photosynthesis'.

Conclusions

Taking all of the observations presented here, we can state that phage photosynthesis genes are distinct, abundant and diverse, and are expressed in the marine environment. We therefore hypothesize that (1) cyanobacterial phage genomes have a significant long-term effect on the evolution of marine photosynthesis, as well as on the short-term response to environmental changes; (2) phage-encoded proteins may play a direct role in determining the level of photosynthetic productivity in oceans (oxygen evolution and carbon fixation), thus redefining the roles of cyanophages in the overall biosphere.

Acknowledgements

We thank E DeLong, M Sullivan, D Lindell, S Chisholm and F Rohwer for the ability to use data before publication, and the captain and crew of the RV Shikmona for their expert assistance at sea. We also thank the governments of Canada, Mexico, Honduras, Costa Rica, Panama, Ecuador and French Polynesia/France for facilitating sampling activities. All sequencing data collected from waters of the

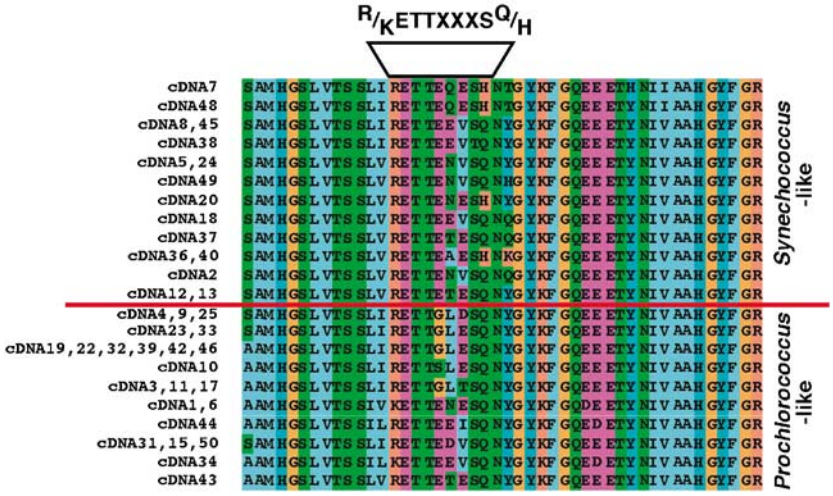

Figure 5 Multiple protein alignment of environmental cDNA-translated D1 protein sequences. The red line separates the *Synechococcus*-like (top) and the *Prochlorococcus*-like sequences (bottom). The consensus sequence for variable region 1 is shown enlarged above the alignment. Redundant protein sequences vary at the nucleotide level and were removed for clarity, while redundant protein names are shown together; see Supplementary Figure S2 for GC plots of the same cDNA sequences.

above countries remain part of the genetic patrimony of the country from which they were obtained. This work was supported by a grant from the Israel Science Foundation, a grant from the Israeli Ministry of Science and Technology, an EMBO YIP award (OB), a COBI grant from the Israeli Ministry of Science and Technology (MS), grants from the US Department of Energy Office of Science's Biological and Environmental Research Program grant and the Gordon and Betty Moore Foundation (J Craig Venter Institute).

References

Adir N, Zer H, Shochat S, Ohad I. (2003). Photoinhibition—a historical perspective. *Photosyn Res* **76**: 343–370.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

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

Brussow H, Hendrix RW. (2002). Phage genomics: small is beautiful. *Cell* **108**: 13–16.

Casjens SR, Gilcrease EB, Huang WM, Bunny KL, Pedulla ML, Ford ME *et al.* (2004). The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. *J Bacteriol* **186**: 1818–1832.

Chen F, Lu J. (2002). Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl Environ Microbiol* **68**: 2589–2594.

Chow WS, Aro EM. (2005). Photoinactivation and mechanisms of recovery. In: Wydrzynki T, Satoh K (eds). *Photosystem II: The Light-Driven Water:Plastoquinone Oxidoreductase. Springer, The Netherlands*. Springer: Dordrecht, The Netherlands, pp 627–648.

Clarke AK, Soitamo A, Gustafsson P, Oquist G. (1993). Rapid interchange between two distinct forms of cyanobacterial photosystem II reaction-center protein D1 in response to photoinhibition. *Proc Natl Acad Sci USA* **90**: 9973–9977.

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**: 827–835.

Clokie MRJ, Mann NH. (2006). Marine cyanophages and light. *Environ Microbiol* **8**: 2074–2082.

Debus R. (2001). Amino acid residues that modulate the properties of tyrosine Y(Z) and the manganese cluster in the water oxidizing complex of photosystem II. *Biochim Biophys Acta* **5**: 164–186.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U *et al.* (2006). Community genomics among stratified microbial assemblages in the Ocean's interior. *Science* **311**: 496–503.

Edelman M, Mattoo AK. (2006). The D1 protein: past and future perspectives. In: Demmig-Adams B, Adams WW, Mattoo AK (eds). *Photoprotection, Photoinhibition, Gene Regulation, and Environment*. Springer: Germany, pp 23–38.

Felsenstein J. (2005). PHYLIP (phylogeny inference package), version 3.6. Department of Genome Sciences, University of Washington, Seattle.

Ferreira KN, Iverson TM, Maghlaoui K, Barber J, Iwata S. (2004). Architecture of the photosynthetic oxygen-evolving center. *Science* **303**: 1831–1838.

Greenberg BM, Gaba V, Mattoo AK, Edelman M. (1987). Identification of a primary *in vivo* degradation product of the rapidly-turning-over 32 kd protein of photosystem II. *EMBO J* **6**: 2865–2869.

Jolliffe IT. (1986). *Principal Component Analysis*. Springer: New York.

Kunik V, Meroz Y, Solan Z, Sandbank B, Weingart U, Ruppin E *et al.* (2007). Functional representation of enzymes by specific peptides. *PLoS Comput Biol* (in press).

Li WKW. (1995). Composition of ultraphytoplankton in the central North Atlantic. *Mar Ecol Prog Ser* **122**: 1–8.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86–89.

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**: 11013–11018.

Man-Aharonovich D, Kress N, Bar Zeev E, Berman-Frank I, Béjà O. (2007). Molecular ecology of *nifH* genes and transcripts in Eastern Mediterranean Sea. *Environ Microbiol* (online early articles doi:10.1111/j.1462-2920.2007.01353.x).

Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, Wheatley PG *et al.* (2005). The genome of S-PM2, a 'photosynthetic' T4-type bacteriophage that infects marine *Synechococcus* strains. *J Bacteriol* **187**: 3188–3200.

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

Massana R, Murray AE, Preston CM, DeLong ED. (1997). Vertical distribution and phylogenetic characterization of marine planktonic *Archaea* in the Santa Barbara channel. *Appl Environ Microbiol* **63**: 50–56.

Millard A, Clokie MRJ, Shub DA, Mann NH. (2004). Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* **101**: 11007–11012.

Mrazek J, Karlin S. (2007). Distinctive features of large complex virus genomes and proteomes. *Proc Natl Acad Sci USA* **104**: 5127–5132.

Nixon PJ, Trost JT, Diner BA. (1992). Role of the carboxy terminus of polypeptide D1 in the assembly of a functional water-oxidizing manganese cluster in photosystem II of the cyanobacterium *Synechocystis* sp. PCC 6803 assembly requires a free carboxyl group at C-terminal position 344. *Biochemistry* **31**: 10859–10871.

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

Rusch DB, Halpern AL, Heidelberg KB, Sutton G, Williamson SJ, 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.

Schaefer MR, Golden SS. (1989). Light availability influences the ratio of two forms of D1 in cyanobacterial thylakoids. *J Biol Chem* **264**: 7412–7417.

Solan Z, Horn D, Ruppin E, Edelman S. (2005). Unsupervised learning of natural languages. *Proc Natl Acad Sci USA* **102**: 11629–11634.

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, 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.

Suorsa M, Regel RE, Paakkarinen V, Battchikova N, Herrmann RG, Aro EM. (2004). Protein assembly of photosystem II and accumulation of subcomplexes in the absence of low molecular mass subunits PsbL and PsbJ. *Eur J Biochem* **271**: 96–107.

Suttle CA. (2005). Viruses in the sea. *Nature* **437**: 356–361.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. (1999). Systematic determination of genetic network architecture. *Nat Genet* **22**: 281–285.

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

Waterbury JB, Watson SW, Valois FW, Franks DG. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can Bull Fish Aquat Sci* **214**: 71–120.

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol* **5**: e16.

Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch D *et al.* (2007). Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean sampling expedition metagenomes. *Environ Microbiol* **9**: 1464–1475.

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

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**: 212–216.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)