HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2015 October 14.

Published in final edited form as:

Nature. 2009 September 10; 461(7261): 258–262. doi:10.1038/nature08284.

All-in-one: Photosystem-I gene cassettes in marine viruses

Itai Sharon[#1,2], **Ariella Alperovitch**[#1], **Forest Rohwer**[3,4], **Matthew Haynes**[3], **Fabian Glaser**[5], **Nof Atamna-Ismaeel**[1], **Ron Y. Pinter**[2], **Frédéric Partensky**[6], **Eugene V. Koonin**[7], **Yuri I. Wolf**[7], **Nathan Nelson**[8], and **Oded Béjà**[1]

[1] Faculty of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel

[2] Faculty of Computer Science, Technion – Israel Institute of Technology, Haifa 32000, Israel

[3] Department of Biology, San Diego State University, San Diego 92182, California, USA

[4] Center for Microbial Sciences, San Diego State University, San Diego 92182, California, USA.

[5] Bioinformatics Knowledge Unit, Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

[6] CNRS and UPMC-Université Paris 6 (UMR 7144), Station Biologique, 29682 Roscoff, France

[7] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[8] Department of Biochemistry, George S. Wise Faculty of Life Sciences, Daniella Rich Institute for Structural Biology, Tel Aviv University, Tel Aviv 69978, Israel

[#] These authors contributed equally to this work.

Abstract

Cyanobacteria of the *Synechococcus* and *Prochlorococcus* genera are important contributors to photosynthetic productivity in the open oceans[1-3]. Here, using pre-existing metagenomic datasets from the global ocean sampling (GOS) expedition[4] as well as from viral biomes[5], we show the first evidence for the presence of photosystem I (PSI) genes in genomes of marine viruses that infect these marine cyanobacteria. Recently, core photosystem II (PSII) genes were identified in cyanophages; they were proposed to be functional in photosynthesis and in increasing viral fitness by supplementing the host production of these proteins[6-9]. The 7 cyanobacterial core PSI genes identified in this study, *psaA, B, C, D, E, K* and a unique *J* and *F* fusion, form a distinctive cluster in cyanophage genomes, suggestive of selection for a distinct function in virus life cycle. The existence of this PSI cluster was confirmed with overlapping and long PCR performed on environmental DNA from the Northern Line Islands. Potentially, the 7 proteins encoded by the viral genes are sufficient to form an intact monomeric PSI complex. Projection of viral predicted

Correspondence and requests for materials should be addressed to O.B. (beja@tx.technion.ac.il)..

Author contributions: A.A. envisioned the initial idea for the project; I.S. & O.B. conceived the experiments; I.S. wrote the code and analyzed the raw data, and together with F.G., R.Y.P., E.V.K., Y.I.W., N.N. & O.B. performed the bioinformatics; F.R. collected DNA & phage concentrates from the Northern Line Islands; A.A., N.A.-I. & M.H. conducted the molecular biology experiments; I.S., F.P., E.V.K., N.N. & O.B. co-wrote the paper.

The PSI sequences reported here have been deposited with GenBank under accession numbers EU926752-EU926761 (overlapping PCRs) & xxxxxxxx (long PCR).

peptides on the cyanobacterial PSI crystal structure[10] suggested that the viral-PSI components may provide a unique way of funneling reducing power from respiratory and other electron transfer chains to PSI.

Bacteriophages (viruses that infect bacteria) have the ability to manipulate the life histories and evolution of their hosts[11,12] and have developed several adaptation and defense mechanisms for efficient survival and multiplication. Most of these involve manipulation of the host DNA, as well as the incorporation, into the phage genomes, of bacterial genes that encode proteins with a potential to facilitate bacteriophage reproduction[13]. Recently, it was discovered that marine cyanophages (bacteriophages that infect cyanobacteria) carry photosynthetic genes, and it was suggested that these genes increase phage fitness[6-9,14,15]. Cyanobacterial photosynthetic membranes contain two photosystems, of which PSII mediates the transfer of electrons from water, the initial electron donor, to the plastoquinone pool whereas PSI mediates electron transfer from plastocyanin to ferredoxin, thereby generating reducing power needed for CO_2 fixation in the form of NADPH. While PSII is known to be sensitive to photodamage, PSI is considered to be a more stable complex. The PSII gene *psbA* coding for the labile D1 protein is readily detected in various cultured and environmental cyanophages infecting *Prochlorococcus* and *Synechococcus*[6,8,16-19]. In addition, other photosynthesis genes encoding the PSII D2 protein[6-8,14,17], high-light inducible proteins, pigment biosynthesis proteins (Ho1, PebA, PcyA), or the photosynthetic electron transport proteins plastocyanin (PetE) and ferredoxin (PetF) were also identified in several cyanophage genomes[8,14].

In order to assess the possible presence of other photosynthesis-related genes in viruses, we set up a designated search scheme for publicly available metagenomic data. Initially we searched for the cyanobacterial PSI gene, *psaA*. Together with PsaB the PsaA protein forms the heterodimeric core of PSI that binds the primary electron donor P700, formed by a special chlorophyll pair[10]. Using tBLASTx[20] with the cut-off expectation values of 1e-20 (see supporting online methods), different *Synechococcus* and *Prochlorococcus psaA* gene sequences were used as queries against the GOS expedition[4] dataset.

We detected 574 *psaA*-containing GOS scaffolds. These were further screened to identify those that were likely to originate from viruses using tBLASTx against refseq_viral[21], a database that contains all known viral genomes. This procedure reduced the number of suspected scaffolds to five. The PsaA homologues encoded by these sequences showed only 65-75% identity to *Prochlorococcus* or marine *Synechococcus* PsaA proteins. On a maximum likelihood tree, four of these proteins clustered together on a well supported branch related to *Prochlorococcus* PsaA, whereas the fifth sequence (JCVI_SCAF_1096628008692) was retrieved near the base of the *Synechococcus* branch (Fig. 1). The GOS general scaffold assembly represents reads that come from different GOS sample sites or from different clones and hence are chimerical by definition. For this reason, we based our further analyses on sequences assembled from single clone reads only. Analyses of the GOS clones containing the modified *psaA* genes confirmed their viral origin (likely cyanophages from the T4-like *Myoviridae* family) as indicated by the presence, in the vicinity of *psaA*, of typical viral genes, such as *nrdA* and *B* (that encode the alpha-2 and

beta-2 subunits of viral ribonucleoside diphosphate reductase, respectively) or the T4-like neck gp13 protein gene (Fig. 2). In addition to *psaA*, these clones contained clusters of PSI genes, including *psaB, psaC* and a unique fused version of the *psaF* and *psaJ* genes (*psaJF*). An analysis of the GOS datasets with other PSI peptides as baits revealed the presence of several other PSI clusters also containing *psaE*, *psaK* and *psaD* genes (see distribution in the different GOS sites in Table 1). As in the case of PsaA, phylogenies made with these additional PSI protein sequences showed that they were all clustered apart from the homologous proteins of *Prochlorococcus* and *Synechococcus*, except PsaC and PsaD from GOS clone 1061008099984 (hereafter clone 9984; a clone used to build the previously mentioned scaffold JCVI_SCAF_1096628008692) which were retrieved closer to corresponding cyanobacterial sequences than to other viral sequences (Fig. S1). Examining the *Prochlorococcus* and *Synechococcus* genome arrangements (middle panel in Fig. 2) or gene-pairs frequency modeling[22] showed that the organization observed in most viral clones, *psaJF-C-A-B-K-E-D*, differs from that observed in these cultured cyanobacterial genomes and in most other (putatively cyanobacteria-derived) GOS sequences (Fig. 3). It is reasonable to assume that there are distinct challenges for regulating PSI expression in a phage and the unique clustering observed may suggest this. Indeed, putative σ^{70} transcriptional promoters were identified upstream to the *psaJF* and to the *psaC* genes (see online methods), suggestive of their expression in a similar stage of infection. It is worth noting that the PSI genes found on clone 9984 (represented by GOS reads 1095964115098 & 1095975140994 in Fig. 2) had a different order (*psaD-C-A)* than on other clones, consistently with their distinct positions in phylogenetic trees (Fig. S1).

To validate the viral origin of these genes and their unique cluster organization, data obtained from the GOS project were cross-referenced with recently released 454 pyrosequencing metagenomic sequences obtained from a variety of marine and non-marine viral and microbial biomes datasets[5]. This was a critical step in increasing the credibility of the results because the two approaches each introduce different biases (some of which are the result of differential DNA extraction techniques, the 454's lack of cloning bias, the abilityto sequence through regions of the genome that exhibit strong secondary structures, exact duplicated sequences, a known artefact of the pyrosequencing approach, or problems with resolving repeat regions; for discussion on pros and cons of shotgun sequencing and 454 sequencing approaches, see references[23-26]). The various viral-suspected PSI GOS clones identified were used to recruit reads from these different datasets. Marine virome fragments were readily recruited to all of the viral GOS clones regions, whereas virome or microbiome fragments coming from other environments were scarcely recruited (Table 2), with a much lower identity (Fig. 4a), further supporting a marine viral origin for the PSI clones. Overall coverage measure of viromes and microbiomes to all different GOS clones containing PSI genes (Fig. 4b) clearly points to 2 distinguished populations, one from bacteria (cyanobacteria) and one from viruses (phages). Except for clone 9984, all our identified viral clones are falling in the viral population. In addition, marine virome fragments were also recruited to regions between the photosynthesis genes, linking neighbor genes in the observed viral cassette (Fig. 4a & Table S2), an observation that supports the gene cluster organization observed on the GOS clones.

To validate the juxtaposition of the genes in the identified viral-PSI gene clusters, DNA from the Northern Line Islands marine virome[27] was used to perform 'continuous' overlapping and long PCR with primers assigned to the different genes (Table S1). The results of the 'continuous' overlapping PCR (lower panel in Fig. 2) and the amplification of a ~6.2 kb long PCR amplicon (Fig. S2 & lower panel in Fig. 2) spanning the entire PSI cassette and including the viral *nrdB* gene, show that the different genes in the cluster *nrdB-hyp-psaJF-C-A-B-K-E-D* (and also a novel arrangement *nrdB-psaJF-C*; genbank # EU926755) are physically linked and exist as one photosynthetic cluster.

In order to assess the significance of the PSI genes discovery in cyanophage genomes, we compared the gene stoichiometry of phage PSI (*psaA*) and PSII (*psbA*) genes in the different Northern Line Islands viromes. In the Kiritimati Island (Christmas Island) station, for example, about 60% of cyanophages (both podoviruses & myoviruses) contained the observed PSI gene cluster (see Table 3 for variations in the PSI cassette abundance estimates at the different Northern Line Islands stations). In comparison, using the same normalized estimates, all Kiritimati station cyanophages were predicted to contain the PSII *psbA* gene. Previous PCR measures from cultured cyanophages[17] indicated that a PSII *psbA* gene indeed occurs in the majority (88%) of cyanophage isolates, while the occurrence of cyanophage isolates containing two PSII genes (*psbA* and *psbD*) is reduced to 50%. The estimates made with cultured phages give our environmental estimates a confidence and point to high temporal abundance of the PSI gene cassette in cyanophages.

The viral PSI gene cassettes were mainly observed in GOS stations GS037 and GS047 (Table 1). We have previously noted that these two GOS stations contain a different repertoire of PSII D1 signatures[18], which are indicative for the presence of different *Prochlorococcus* genotype than in other GOS stations. Indeed, novel *Prochlorococcus* genotypes have been identified in the GOS dataset from samples associated with the South Pacific and Indian Ocean Gyres. Nevertheless, preliminary assembly of the novel *Prochlorococcus* genomes shows that their PSI genes arrangement is distinct from that of the viral cassettes presented here since the genes are scattered in the genome (personal communication, D. Rusch), as is also the case in all cultured cyanobacteria genomes described so far. The viral arrangement is also distinct from the PSI gene organization found in the newly discovered uncultured oceanic N_2-fixing UCYN-A cyanobacteria[28] that possess only PSI. It has recently been proposed[24] that different viral assemblages comprise essentially the same types of viruses, and that changes in geographical or environmental conditions allow different viral genotypes to become more or less prevalent; therefore the less abundant viruses are not lost altogether, but are merely reduced in occurrence. This would explain the observed high occurrence of the viral PSI cluster at only two GOS stations.

Although the data presented here are derived from environmental genomic datasets (non-continuous data), and therefore the lack of genes is not a proof of absence, it is notable that the PSI genes *psaI*, *psaL* and *psaM* were not found in the viral *psa* gene cassettes (*psaX* is also missing, but it is not found in *Prochlorococcus* or in marine *Synechococcus* genomes either). The *psaM* gene is naturally absent from plants[29] and its inactivation in cyanobacteria shows that it is mainly required for the formation of stable PSI trimers[30]. Similarly, targeted

inactivation of cyanobacterial *psaL* produces functional PSI complexes unable to form trimers, whereas PsaI is mostly required for stabilizing PsaL[31]. Therefore, these three proteins are mainly involved in the trimer formation of cyanobacterial PSI[32], and their potential absence from the viral clone might indicate on the formation of a monomeric PSI complex as in plants[33,34] and not a trimeric complex as in cyanobacteria[35-37] (see ref[38] for possible evolutionary scenarios that led to the formation of monomeric PSI in plants from trimeric cyanobacterial PSI). All genetic information required to form this putative minimal, monomeric PSI is clustered onto a very small cyanophage genome fragment (~5.9 kb). To our knowledge, gene clusters encoding all the components of a photosystem from an oxygenic phototroph have not been previously reported, and neither have there been reports on cyanobacterial PSI genes outside a cyanobacterial chromosome.

The existence, in the phage genome, of a unique cluster of 8 PSI genes that is not found in any of the available cyanobacterial genomes, appeared highly unexpected. Assuming that phages randomly (one by one) acquire genes from the genome of infected *Prochlorococcus* via non-homologous recombination (of course, only a tiny, and non-random minority of the acquired genes are fixed in the phage population), it is easy to estimate that there are from 10^{16} to 10^{23} possible combinations of 6 to 8 gene blocks (see the online supplement). This huge number seems to suggest that the formation of the *psaJF-C-A-B-K-E-D* cluster in the phage genome is an extremely unlikely event. However, a rough estimation of the number of recombination events occurring in the ocean during cyanophage infections suggests that, over the ~2.5 billion years of the co-existence of cyanobacteria and cyanophages, this number is within the range of ~10^{21} to 10^{25} (see the online supplement). This estimate is close to (or even exceeds) the above number of gene combinations, suggesting that the assembly of the cluster of 8 PSI genes during the evolution of the system cyanobacteria-cyanophages actually could be expected; and, given its utility, once it emerged, this cluster would be fixed and spread in the population. A corollary of these estimates is that one would expect to find, in marine phage genomes, a variety of novel clusters of functionally linked genes.

The potential structural consequences of assembling the phage proteins into the PSI complex were modeled in relation to the 2.5 Å structure of PSI from the cyanobacterium *Thermosynechococcus elongatus*[10]. We modeled the PsaJF fusion protein (where the C-terminus of PsaJ is fused to the N-terminus of PsaF) at the position of subunits J and F of PSI, using the COOT program[39]. Figure 5 shows that the viral PsaJF fusion protein fits perfectly at the position of subunits J and F in the PSI structure. The only prominent change was the absence of the N-terminus of subunit F, which is responsible for the specific binding of the natural electron donor (plastocyanin) of PSI[40-42]. In chloroplasts of green algae and plants, this part of subunit F is elongated, resulting in higher affinity of plastocyanin to the chloroplast PSI[29,42-45]. While both plastocyanin and cytochrome c_6 are capable of donating electrons to PSI[44] in *Chlamydomonas reinhardtii*, this site in higher plants is specific for plastocyanin[46]. However, the electron donation to PSI is not at all promiscuous, and several soluble cytochromes, including the respiratory cytochrome *c*, fail to donate electrons to PSI[47]. We hypothesize that the replacement of PsaJ and PsaF with the viral PsaJF fusion

protein enables electron donation through additional electron carriers, including cytochromes that usually function as electron donors to cytochrome oxidase.

The mechanistic consequence of a less selective electron donation to PSI might be the possibility of sharing reducing power generated by the respiratory chain with the photosynthetic electron transport chain. A similar phenomenon, called chloro-respiration, that was evidenced in both cyanobacteria and chloroplasts, was attributed to the plastid terminal plastoquinone oxidase (PTOX)[48]. The electron mediator in this process is plastoquinone, which shuttles between the respiratory-like chain and the chloroplast b_6f complex[48]. In our hypothesis, after phage infection and the incorporation of the phage gene products into PSI, the function of electron mediation would be carried out by a soluble cytochrome. In addition, the phage may be boosting the amount of PSI in order to lead the infected cyanobacterial cells toward a cyclic photosynthesis for the generation of ATP in expense for the production of reducing power for CO_2 fixation. Interestingly, PSI levels have been shown to be notably low in both oceanic *Synechococcus*[49] and in *Prochlorococcus*[50], possibly as a result to adaptation to low iron levels, and it was recently proposed that (at least in the first genus) there might be a compensatory mechanism, involving alternative electron flow to O_2[ref49].

Although we have not proved that phage PSI gene products participate in host photosynthesis or that they are connected to electron mediation within the host cells, it appears most likely that the phage proteins do perform these functions. First, the maintenance of eight PSI genes in phage genomes obviously requires strong selection, so by inference, these genes are likely to be important for phage reproduction; second, the presence of all genes necessary to build a monomeric PSI complex is suggestive of a photosynthetic function of these proteins; finally, the observed modifications are in an area implicated in binding the PSI electron donor. Obviously, the next stages in testing the 'alternative reducing power' hypothesis are to study the behavior of this phage-host system using phages in which the PSI genes have been inactivated (when such a phage is discovered) or to use heterologous cyanobacterial systems in which viral PSI genes would replace the host's native ones.

The phage PSI gene fusion *psaJF* described here is the first example of a phage gene innovation that involves structural membrane proteins. Modification toward a new function of existing cyanobacterial proteins by their phages was recently demonstrated for the divergent phage PebA homolog[15] (now renamed PebS [phycoerythrobilin synthase]). The phage PebS single-handedly catalyzes a reaction for which uninfected host cells require two consecutive enzymes, PebA and PebB. Considering these findings and the above calculations that suggest a high likelihood of gene cluster formation in phage genomes, the oceanic virome is expected to be an almost unlimited source of naturally bioengineered gene cassettes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank M. Rosenberg for technical support, D. Rusch, J. Zehr and S. Bench for sharing genomic data, D. Lindell and R. Sorek for encouragement and discussions and U. Pick for the comments on cyclic photosynthesis. This work was supported in part by grants 1203/06 (O.B.) and 356/06 (N.N.) from the Israel Science Foundation, by the Henry Taub Award for Academic Excellence and the Technion V.P.R. Fund-Henri Gutwirth Promotion of Research Fund (O.B.).

References

1. Li WKW, Dickie PM, Harrison WG, Irwin BD. Biomass and production of bacteria and phytoplankton during the spring bloom in the western North Atlantic Ocean. Deep-Sea Res. 1993; 40:307–327.
2. Liu H, Nolla HA, Campbell L. *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquat Microb Ecol. 1997; 12:39–47.
3. Partensky F, Hess WR, Vaulot D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev. 1999; 63:106–127. [PubMed: 10066832]
4. Rusch DB, et al. The Sorcerer II Global Ocean Sampling expedition: I, The northwest Atlantic through the eastern tropical Pacific. PLoS Biol. 2007; 5:e77. [PubMed: 17355176]
5. Dinsdale EA, et al. Functional metagenomic profiling of nine biomes. Nature. 2008; 452:629–632. [PubMed: 18337718]
6. Mann NH, Cook A, Millard A, Bailey S, Clokie M. Bacterial photosynthesis genes in a virus. Nature. 2003; 424:741. [PubMed: 12917674]
7. Millard A, Clokie MRJ, Shub DA, Mann NH. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. Proc Natl Acad Sci U S A. 2004; 101:11007–11012. [PubMed: 15263091]
8. Lindell D, et al. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A. 2004; 101:11013–11018. [PubMed: 15256601]
9. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. Photosynthesis genes in marine viruses yield proteins during host infection. Nature. 2005; 438:86–89. [PubMed: 16222247]
10. Jordan P, et al. Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. Nature. 2001; 411:909–917. [PubMed: 11418848]
11. Suttle CA. Marine viruses - major players in the global ecosystem. Nature Rev Microbiol. 2007; 5:801–812. [PubMed: 17853907]
12. Rohwer F, Thurber RV. Viruses manipulate the marine environment. Nature. 2009; 459:207–212. [PubMed: 19444207]
13. Brown NF, Wickham ME, Coombes BK, Finlay BB. Crossing the line: selection and evolution of virulence traits. PLoS Pathog. 2006; 2:e42. [PubMed: 16733541]
14. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol. 2005; 3:e144. [PubMed: 15828858]
15. Dammeyer T, Bagby SC, Sullivan MB, Chisholm SW, Frankenberg-Dinkel N. Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. Curr Biol. 2008; 18:442–448. [PubMed: 18356052]
16. Zeidner G, et al. Potential photosynthesis gene recombination between *Prochlorococcus* & *Synechococcus* via viral intermediates. Environ Microbiol. 2005; 7:1505–1513. [PubMed: 16156724]
17. Sullivan MB, et al. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. 2006; 4:e234. [PubMed: 16802857]
18. Sharon I, et al. Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J. 2007; 1:492–501. [PubMed: 18043651]
19. Tzahor S, et al. A supervised learning approach for taxonomic classification of core-photosystem-II genes and transcripts in the marine environment. BMC Genomics. 2009; 10:229. [PubMed: 19445709]

20. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
21. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35(Database issue):D61–65. [PubMed: 17130148]
22. Kagan J, Sharon I, Béjà O, Kuhn J. The tryptophan pathway genes of the Sargasso Sea metagenome: new operon structures and the prevalence of nonoperon organization. Genome Biol. 2008; 9:R20. [PubMed: 18221558]
23. Goldberg SM, et al. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A. 2006; 103:11240–11245. [PubMed: 16840556]
24. Angly F, et al. The marine viromes of four oceanic regions. PLoS Biol. 2006; 4:e368. [PubMed: 17090214]
25. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nature Biotechnol. 2008; 26:1117–1124. [PubMed: 18846085]
26. Harismendy O, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome biology. 2009; 10:R32. [PubMed: 19327155]
27. Dinsdale EA, et al. Microbial ecology of four coral atolls in the Northern Line Islands. PLoS ONE. 2008; 3:e1584. [PubMed: 18301735]
28. Zehr JP, et al. Globally distributed uncultivated oceanic N_2-Fixing cyanobacteria lack oxygenic photosystem II. Science. 2008; 322:1110–1112. [PubMed: 19008448]
29. Ben-Shem A, Frolow F, Nelson N. The crystal structure of plant photosystem I. Nature. 2003; 426:630–635. [PubMed: 14668855]
30. Naithani S, Hou JM, Chitnis PR. Targeted inactivation of the *psaK1*, *psaK2* and *psaM* genes encoding subunits of Photosystem I in the cyanobacterium *Synechocystis* sp. PCC 6803. Photosyn Res. 2000; 63:225–236. [PubMed: 16228433]
31. Xu Q, et al. Mutational analysis of photosystem I polypeptides in the cyanobacterium *Synechocystis* sp, PCC 6803 - Targeted inactivation of *psaI* reveals the function of PsaI in the structural organization of PsaL. J Biol Chem. 1995; 270:16243–16250. [PubMed: 7608190]
32. Xu W, Tang H, Wang Y, Chitnis PR. Proteins of the cyanobacterial photosystem I. Biochim Biophys Acta. 2001; 1507:32–40. [PubMed: 11687206]
33. Kouril R, van Oosterwijk N, Yakushevska AE, Boekema EJ. Photosystem I: a search for green plant trimers. Photochem Photobiol Sci. 2005; 4:1091–1094. [PubMed: 16307127]
34. Nechushtai R, Nelson N. Photosystem I reaction centers from *Chlamydomonas* and higher plant chloroplasts. J Bioenerg Biomembr. 1981; 13:295–306. [PubMed: 6277884]
35. Boekema EJ, et al. Evidence for a trimeric organization of the photosystem I complex from the thermophilic cyanobacterium *Synechococcus* sp. FEBS Lett. 1987; 217:283–286.
36. Bibby TS, Mary I, Nield J, Partensky F, Barber J. Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. Nature. 2003; 424:1051–1054. [PubMed: 12944966]
37. Bibby TS, Nield J, Partensky F, Barber J. Oxyphotobacteria: Antenna ring around photosystem I. Nature. 2001; 413:590. [PubMed: 11595938]
38. Ben-Shem A, Frolow F, Nelson N. Evolution of photosystem I - from symmetry through pseudo-symmetry to asymmetry. FEBS Lett. 2004; 564:274–280. [PubMed: 15111109]
39. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr. 2004; D60:2126–2132.
40. Wynn RM, Malkin R. Interaction of plastocyanin with photosystem I: a chemical cross-linking study of the polypeptide that binds plastocyanin. Biochemistry. 1988; 27:5863–5869. [PubMed: 3056515]
41. Farah J, Rappaport F, Choquet Y, Joliot P, Rochaix JD. Isolation of a psaF-deficient mutant of *Chlamydomonas reinhardtii*: efficient interaction of plastocyanin with the photosystem I reaction center is mediated by the PsaF subunit. EMBO J. 1995; 14:4976–4984. [PubMed: 7588626]

42. Hippler M, Drepper F, Farah J, Rochaix JD. Fast electron transfer from cytochrome c6 and plastocyanin to photosystem I of *Chlamydomonas reinhardtii* requires PsaF. Biochemistry. 1997; 36:6343–6349. [PubMed: 9174349]
43. Sommer F, Drepper F, Haehnel W, Hippler M. J Biol Chem. 2006; 281:35097–35103. [PubMed: 16984911]
44. Nelson N, Yocum C. Structure and function of photosystems I and II. Annu Rev Plant Biol. 2006; 57:521–565. [PubMed: 16669773]
45. Amunts A, Drory O, Nelson N. The structure of a plant photosystem I supercomplex at 3.4 Å resolution. Nature. 2007; 447:58–63. [PubMed: 17476261]
46. Merchant S, Sawaya MR. The light reactions: a guide to recent acquisitions for the picture gallery. Plant Cell. 2005; 17:648–663. [PubMed: 15746074]
47. Kerfeld CA, Krogmann DW. Photosynthetic cytochromes *c* in cyanobacteria, algae and plants. Annu Rev Plant Physiol Plant Mol Biol. 1998; 49:397–425. [PubMed: 15012240]
48. Rumeau D, Peltier G, Cournac L. Chlororespiration and cyclic electron flow around PSI during photosynthesis and plant stress response. Plant Cell Environ. 2007; 30:1041–1051. [PubMed: 17661746]
49. Bailey S, et al. Alternative photosynthetic electron flow to oxygen in marine *Synechococcus*. Biochim Biophys Acta. 2008; 1777:269–276. [PubMed: 18241667]
50. Partensky F, LaRoche J, Wyman K, Falkowski PG. The divinyl-chlorophyll *a/b*-protein complexes of two strains of the oxyphototrophic marine prokaryote *Prochlorococcus* characterization and response to changes in growth irradiance. Photosynth Res. 1997; 51:209–222.
51. Markowitz VM, et al. The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res. 2008; 36(Database issue):D528–533. [PubMed: 17933782]
52. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52:696–704. [PubMed: 14530136]
53. Hordijk W, Gascuel O. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics. 2005; 21:4338–4347. [PubMed: 16234323]

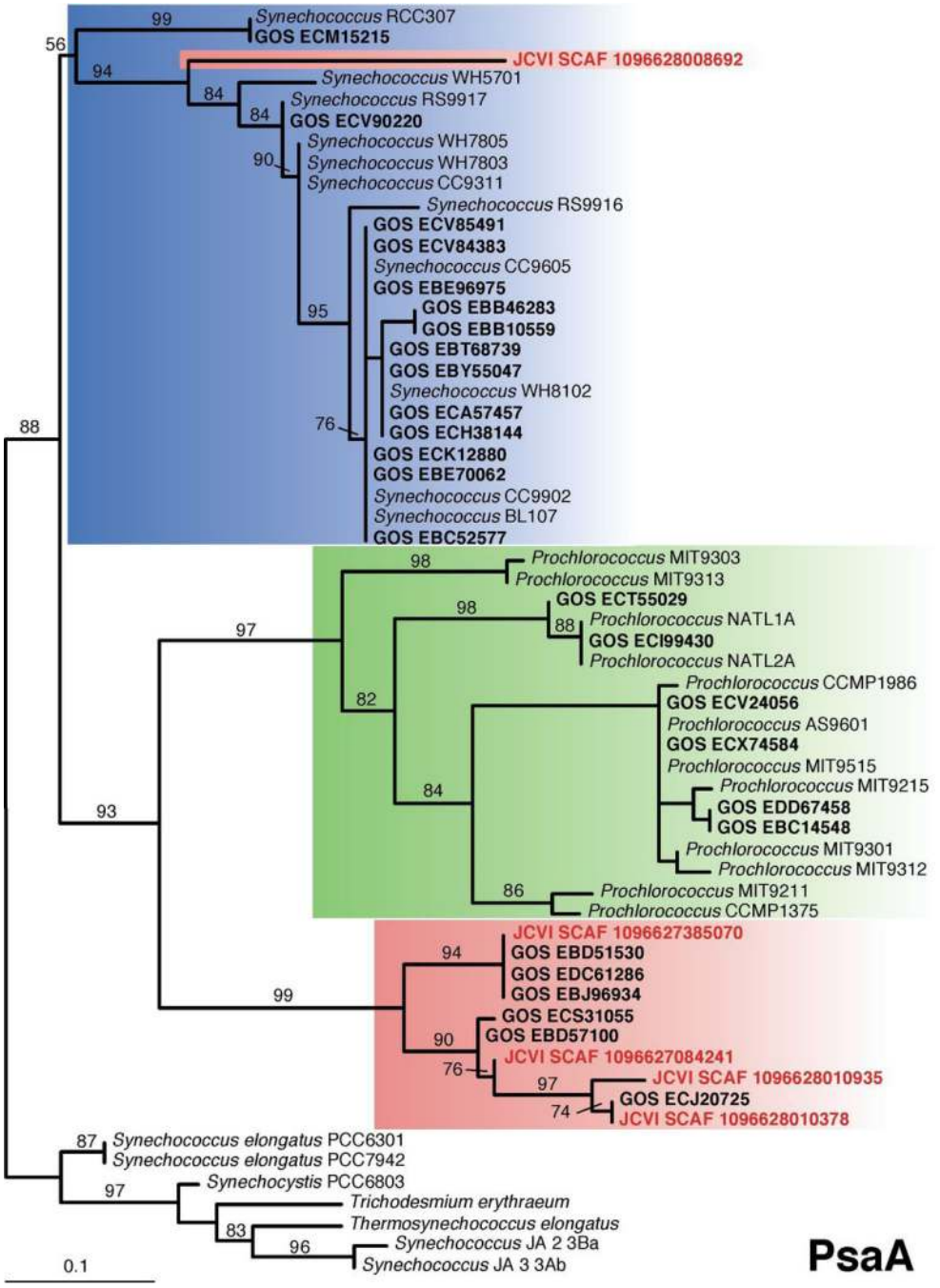


Figure 1. A maximum likelihood phylogenetic tree of *psaA*-deduced amino acid sequences obtained from the GOS expedition

PsaA sequences from the 27 fully sequenced and annotated *Synechococcus* and *Prochlorococcus* genomes retrieved from the IMG system[51] are shown. For clarity, the tree shows only a subset of the 583 partial PsaA sequences found in the GOS dataset. The tree is based on an alignment of 94 shared amino acids. The tree was created using the program PhyML[52], with nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR)[53] topological search, and a rapid approximate likelihood-ratio test (aLRT) to calculate branch support percentage in all analyses to provide confidence estimation for the inferred

topologies. Bootstrap values exceeding 50% are indicated above the branches. Sequences from the GOS expedition are shown in boldface (protein GenBank accession numbers are shown), and sequences from the original five scaffolds obtained in this study are indicated in red. *Synechococcus* cluster background is colored with blue, *Prochlorococcus* with green, and viral in red.

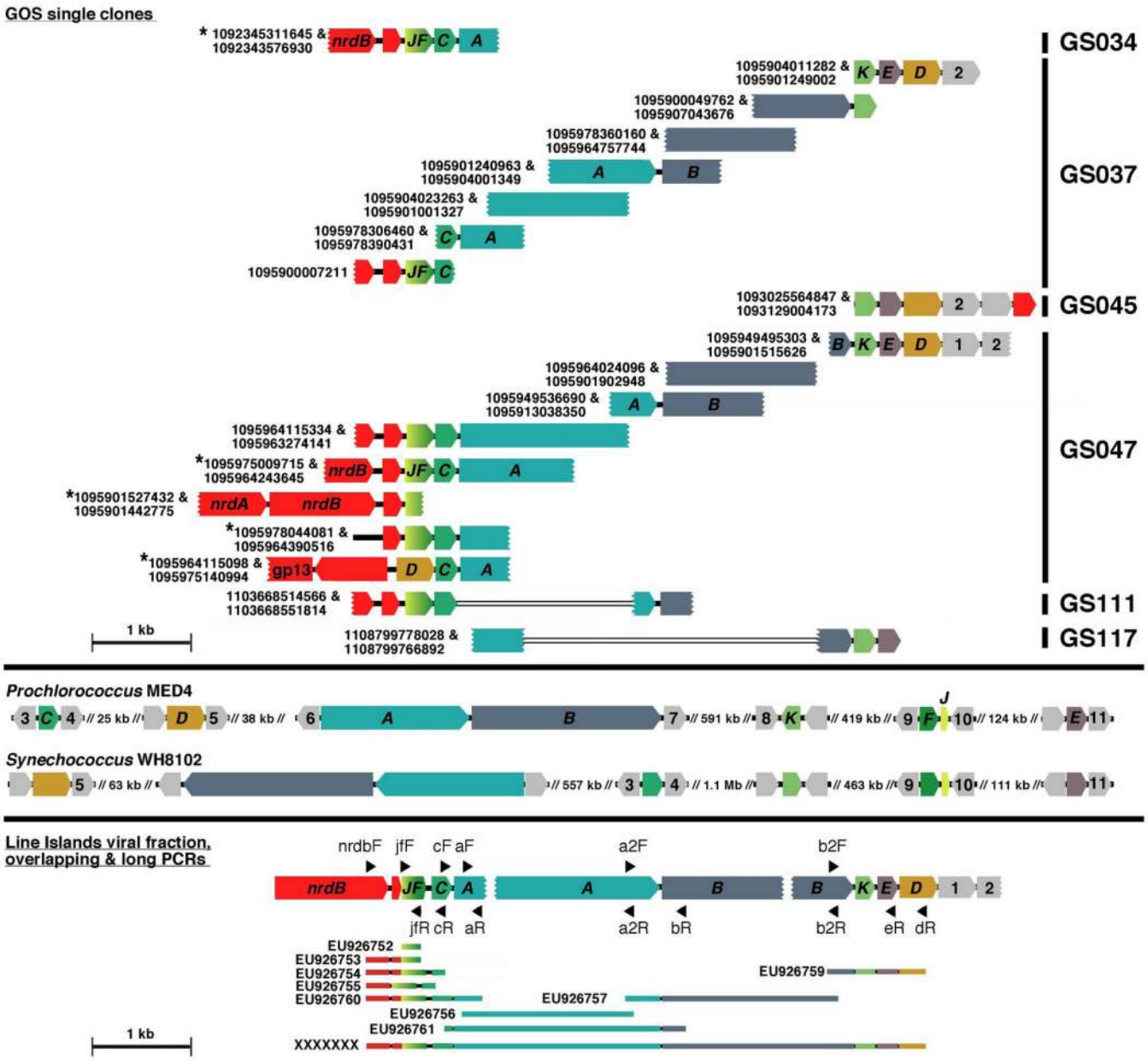


Figure 2. Schematic physical maps of viral-suspected clones & environmental PCR products containing PSI genes

Schematic physical map of GOS viral-suspected physical clones containing PSI genes (upper panel), *Prochlorococcus* and *Synechococcus* genomes (middle panel) and Northern Line Islands overlapping PCR products (lower panel). Predicted ORFs are depicted as colored arrows indicating the direction of the transcription and are not drawn to scale. Red arrows represent ORFs with predicted viral origin and grey arrows represent unknown ORFs. *JF, C, D, A, B, K* and *E* markings denote the fused gene *psaJF* and the genes *psaC*, *psaD*, *psaA*, *psaB*, *psaK* and *psaE*, respectively. Numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11 denote a conserved membrane protein, NADH plastoquinone oxidoreductase, acyl carrier protein, D-fructose-6-phosphate amidotransferase, anthranilate synthase component I and

chorismate binding enzyme, bi-functional CbiH protein, glycosyl transferase, 1-deoxy-D-xylulose-5-phosphate synthase, o-sialoglycoprotein endopeptidase, guanylate kinase, and formamidopyrimidine-DNA glycolase, respectively. Accession numbers are shown on the left (JCVI_READ_#). The five GOS clone scaffolds originally identified using the *psaA* gene are marked with asterisks and GOS station # are shown on the right. For reasons of clarity, not all clones are shown, and a full list and sequence data of the different clones may be found in Fi`le_S1.fasta. Gaps shown in Indian Ocean GOS clones (reads 1103668514566 and 1103668551814, 1108799778028 and 1108799766892) are the result of regions in the clones that were not covered by the end reads due to the large size of these clones (5 kb). Northern Line Islands overlapping PCR products GenBank accession numbers are shown to the left of each schematic PCR product. Primer positions on GOS clones are indicated by triangles, while thick colored lines denote PCR products. Northern Line Islands gene organizations are shown below.

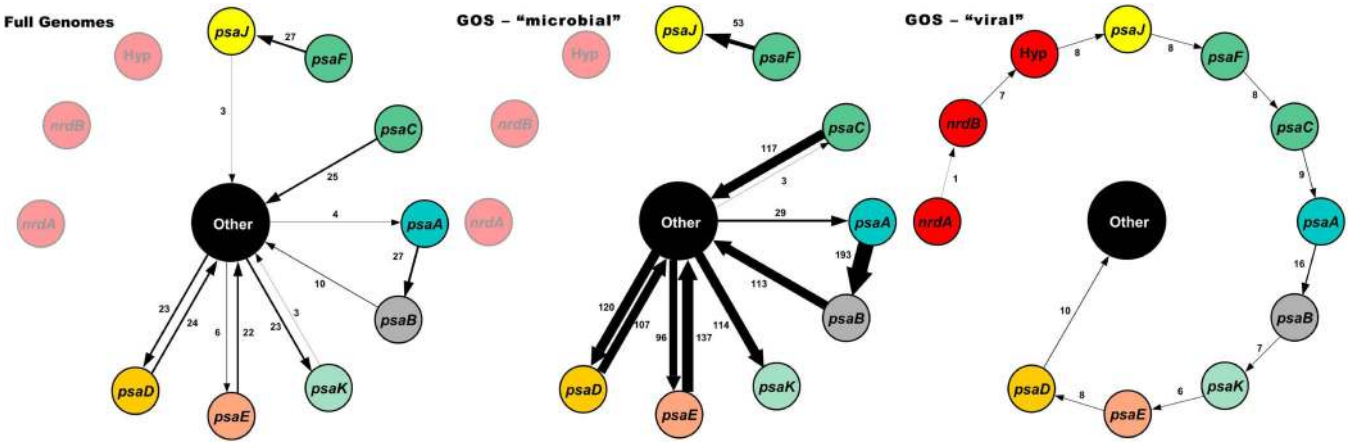
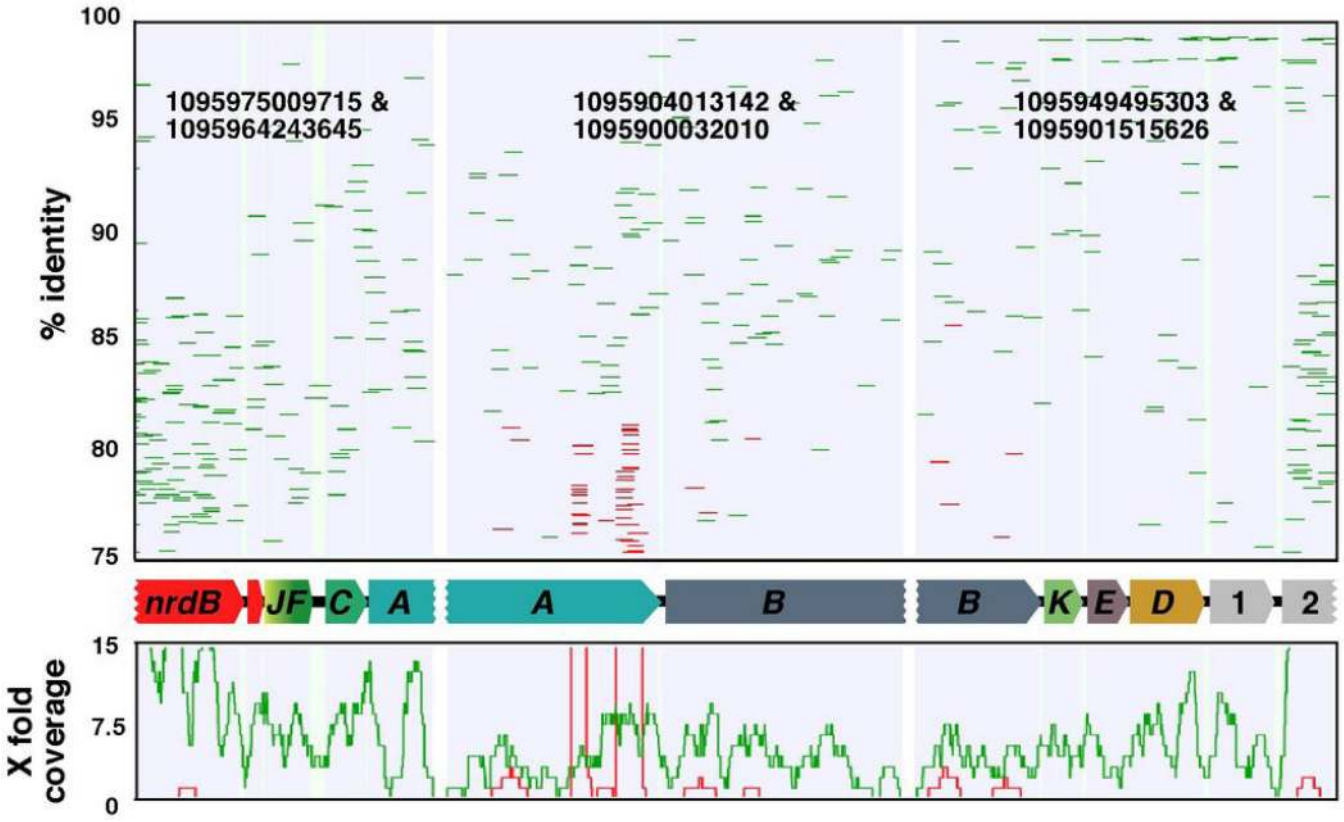


Figure 3. Distribution of neighboring genes involving at least one PSI gene

Each arrow connects neighboring genes, and its thickness represents the number of pairs found in *Synechococcus* and *Prochlorococcus* genomes [27 genomes retrieved from the integrated microbial genomes (IMG) system[51], (left gene-circle)], microbial sequences from the GOS metagenome (middle gene-circle), and viral sequences from the GOS metagenome (right gene-circle). Note the uninterrupted clustering of PSI genes in phage genomes that contrasts the scattered arrangement of these genes in cyanobacterial genomes (in both cultures and GOS). Gene connections observed only once are not shown.



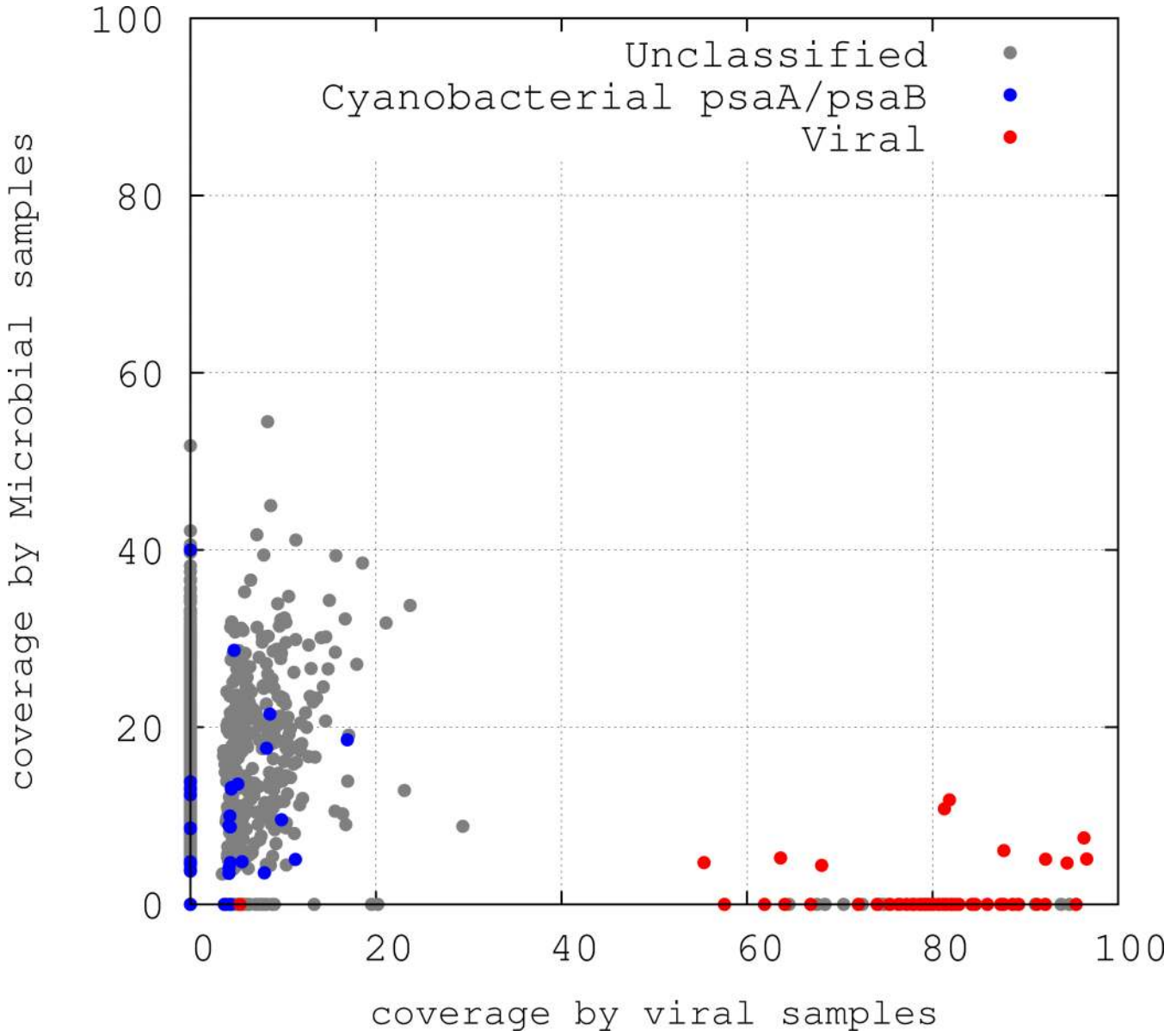
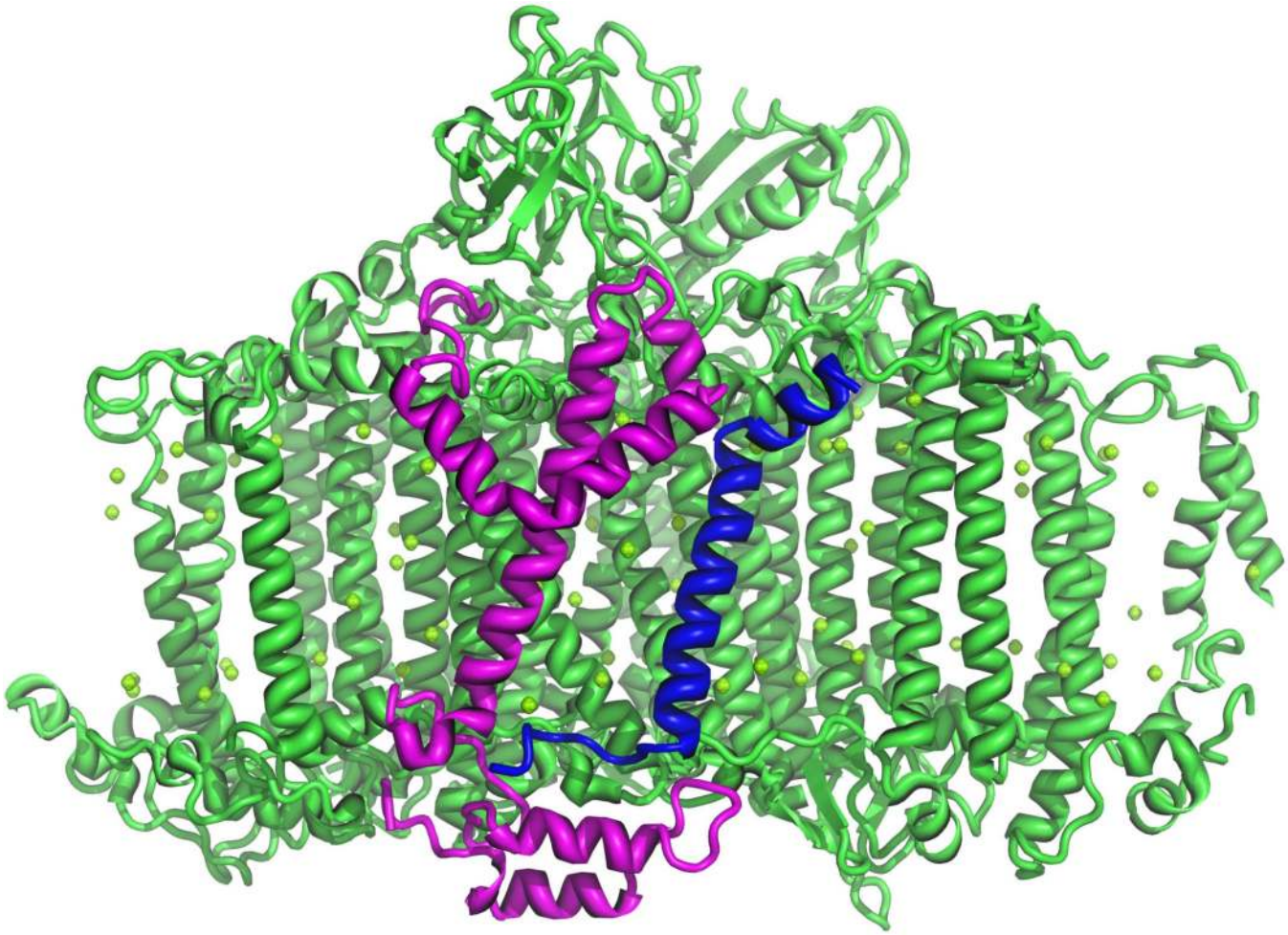


Figure 4. Recruitment of GOS clones carrying PSI genes with Northern Line Islands biomes

a, Recruitment of three GOS physical clones carrying suspected viral-PSI genes by Northern Line Islands virome (green) and microbiome (red) reads. Upper panel shows recruitment at 75-100% identity while lower panel shows fold coverage by these reads. Accession numbers of the GOS reads used are presented above each clone (JCVI_READ_#). **b**, Recruitment coverage of GOS single clones carrying PSI genes with Northern Line Islands biomes [viromes (X-axis) and microbiomes (Y-axis)]. Coverage is defined as % of GOS clone length covered by at least one recruited read. Note the single viral clone located at the "microbial area" is, in fact, GOS clone JCVI_TMPL_1061008099984, which carries a different organization of genes from the other viral-PSI clones. Notably, all control *psaA/psaB* genes from cultured cyanobacteria (blue dots) fall within the bacterial population,

which is a quality indication for accurate fractionation used to separate microbial and viral biomes[5].



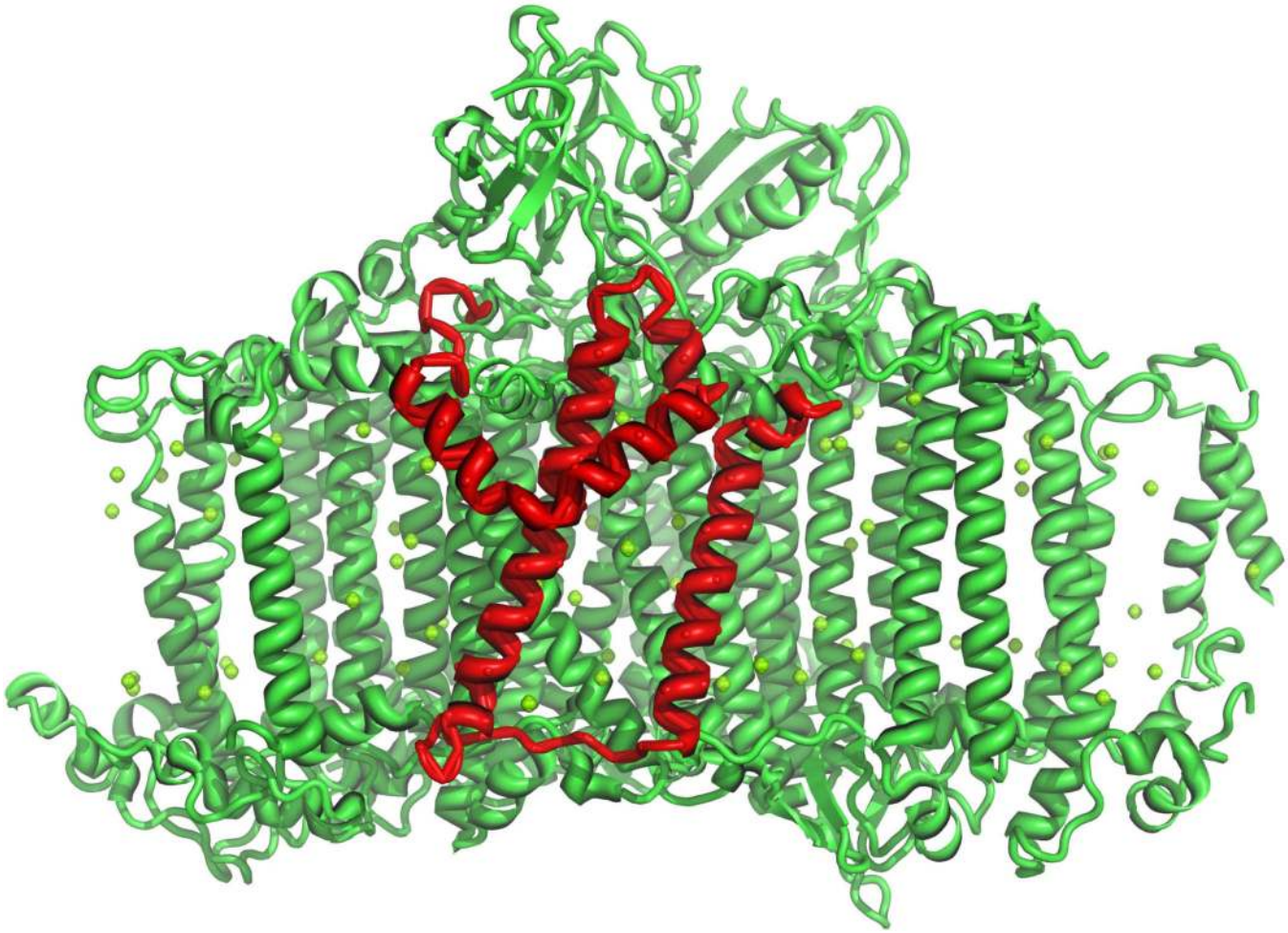


Figure 5. Structural consequences of assembling the viral fusion protein PsaJF into PSI

a, The structure of *T. elongatus* PSI (subunits) was illustrated by PyMOL (http://pymol.sourceforge.net/) using a PSI monomer (adopted from PDB 1jb0). PsaF (magenta), PsaJ (blue), and all the other subunits in green. **b**, A model for the structure of the viral PsaJF fusion protein (red) substituting the original PsaF and PsaJ subunits. The PsaJF was generated in COOT[39] by truncating the N-terminus of PsaF and substituting the amino acids into the corresponding viral amino acids according to the alignment of the fusion protein with PsaJ and PsaF.

Table 1

Distribution of viral-PSI clones (one or two reads sequenced from the same DNA fragment) in the GOS expedition.

GOS site[a]	# Clones	Site Description
GS025	1	Dirty Rock, Cocos Island
GS027	1	Devil's Crown, Floreana Island
GS029	1	North James Bay, Santiago Island
GS034	1	North Seamore Island
GS037	8	Equatorial Pacific TAO Buoy
GS045	1	400 miles from French Polynesia
GS047	11	201 miles from French Polynesia
GS111	1	Indian Ocean
GS117	1	Seychelles

[a]For exact site location and detailed description, see http://camera.calit2.net/ and Table 1 in ref[4].

Table 2

Number of different biomes reads recruited at different percent identity to GOS suspected viral-PSI clones.

Type	Microbial Metagenomes			Viral Metagenomes		
	85%	90%	95%	85%	90%	95%
Coral	0	0	0	0	0	0
Fish	0	0	0	0	0	0
Freshwater	2	0	0	1	1	1
Hypersaline	1	0	0	0	0	0
Marine	1	0	0	207	144	91
Microbialites	0	0	0	0	0	0
Terrestrial	0	0	0	0	0	0

Table 3

Percent cyanophages carrying PSI (*psaA*) or PSII (*psbA*) genes in the different Northern Line Islands viromes.

	Kingman	Palmyra	Tabuarean	Kiritimati
% phage[a] with PSI genes	6.5	16.1	54.4	61.6
% phage with PSII genes	96.0	58.0	111.2	99.8

[a]Cyanophage (myovirus & podovirus) abundance estimates were based on the myovirus viral-capsid-assembly-gene (g20), the myovirus major-capsid-protein-gene (g23), and the podovirus T7-like DNA polymerase markers. See online methods for details about normalization and recruitment conditions.