Transfer of photosynthesis genes to and from *Prochlorococcus* viruses

Debbie Lindell*†, Matthew B. Sullivan†‡, Zackary I. Johnson*, Andrew C. Tolonen‡, Forest Rohwer§, and Sallie W. Chisholm*¶||

Departments of *Civil and Environmental Engineering and ¶Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; ‡Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution and Massachusetts Institute of Technology, Cambridge, MA 02139; and §Department of Biology, San Diego State University, San Diego, CA 92182

Comparative genomics gives us a new window into phage–host interactions and their evolutionary implications. Here we report the presence of genes central to oxygenic photosynthesis in the genomes of three phages from two viral families (*Myoviridae* and *Podoviridae*) that infect the marine cyanobacterium *Prochlorococcus*. The genes that encode the photosystem II core reaction center protein D1 (*psbA*), and a high-light-inducible protein (HLIP) (*hli*) are present in all three genomes. Both myoviruses contain additional *hli* gene types, and one of them encodes the second photosystem II core reaction center protein D2 (*psbD*), whereas the other encodes the photosynthetic electron transport proteins plastocyanin (*petE*) and ferredoxin (*petF*). These uninterrupted, full-length genes are conserved in their amino acid sequence, suggesting that they encode functional proteins that may help maintain photosynthetic activity during infection. Phylogenetic analyses show that phage D1, D2, and HLIP proteins cluster with those from *Prochlorococcus*, indicating that they are of cyanobacterial origin. Their distribution among several *Prochlorococcus* clades further suggests that the genes encoding these proteins were transferred from host to phage multiple times. Phage HLIPs cluster with multicopy types found exclusively in *Prochlorococcus*, suggesting that phage may be mediating the expansion of the *hli* gene family by transferring these genes back to their hosts after a period of evolution in the phage. These gene transfers are likely to play a role in the fitness landscape of hosts and phages in the surface oceans.

The genomes of bacterial viruses (phages) contain a variety of genes homologous to those found in their hosts (1–5). Many encode functional proteins involved in processes of direct importance for the production of phage progeny. They include genes involved in DNA replication, nucleotide metabolism, and RNA transcription and are found in both lytic phage and prophage (3, 6). It is likely that many originated from their hosts (2, 4) and that some host genes that occur in multiple copies have been (re)acquired from phages (2, 7) either after a period of evolution in the phage or after acquisition of the gene from a different host.

Host genes that are not directly related to the production of new phages, such as genes involved in phosphate sensing and metabolism (8, 9), and the scavenging of oxygen radicals (10) are also found in phage genomes and may benefit phages by temporarily enhancing host functionality before lysis. In addition, prophages can provide their hosts with new functions by encoding genes, such as virulence factors, toxin production genes, and immune response genes (5, 6, 11).

Genes involved in photosynthesis have recently been found in a lytic phage isolated on *Synechococcus* WH7803 (12), a member of the marine cluster A unicellular cyanobacteria that is widespread in the oceans. A member of the *Myoviridae* family of double-stranded DNA viruses, this phage contains two photosynthetic genes (*psbD* and an interrupted *psbA* gene) that code for the two photosystem II (PSII) core reaction center proteins found in all oxygenic photosynthetic organisms. These genes were not found in a different phage (a member of the *Podoviri-*

dae family) isolated on the same strain of *Synechococcus* (13). These observations lead one to wonder whether the presence of photosynthetic genes in phage is a rare phenomenon and to what extent it is specific for a particular phage or host type. If these genes are widespread in cyanophage, what is their origin? Were they acquired through a single ancestral transfer event?

The phage–host system for *Prochlorococcus* and *Synechococcus* (14, 15), which form a monophyletic clade within the cyanobacteria (16–19), is well suited to begin to answer these questions. Members of each genus form distinct subgenera clusters within this clade, which in *Prochlorococcus* also correspond to their efficiency of light utilization (17). Numerous phages have been isolated by using this diverse group, including members of the *Myoviridae*, *Podoviridae*, and *Siphoviridae* families, and the degree of cross-infection, a mechanism for horizontal gene transfer, has been analyzed (14, 15). The genomes of four host strains (20–22) and three phages (U.S. Department of Energy Joint Genome Institute; www.jgi.doe.gov) have been sequenced, providing a database to analyze the distribution and phylogenetic relationships of host genes among hosts and their phages.

Here we report that the genomes of three *Prochlorococcus* phages collectively contain a number of host-like photosynthetic genes. We further hypothesize from bioinformatic analyses that these genes likely play a functional role during infection and impact the evolutionary trajectory of both phages and hosts in the surface oceans.

Materials and Methods

Selection and Preparation of Cyanophage for Genome Sequencing. Three phages were chosen for sequencing with no prior knowledge of their gene content. P-SSP7, a T7-like podovirus characterized by a small capsid (≈50 nm), a noncontractile tail, and a 45-kb genome infects a single high-light-adapted (HL) *Prochlorococcus* strain. P-SSM2 and P-SSM4 are T4-like myoviruses characterized by larger capsids (≈85 nm and ≈80 nm respectively), long contractile tails, and larger genomes (252 kb and 178 kb, respectively). P-SSM2 infects three low-light-adapted (LL) *Prochlorococcus* strains, and P-SSM4 infects two HL and two LL *Prochlorococcus* strains (see Table 1) (15). None of the three phages infect *Synechococcus*. The vastly different protein com-

EVOLUTION

Table 1. Phages used in this study and their photosynthesis-related genes

Phage	Family	Host strains infected	Gene products
P-SSP7	Podovirus	*Pro* **MED4** (HL)	D1 and one HLIP
P-SSM2	Myovirus	*Pro* **NATL1A**, NATL2A, and MIT9211 (LL)	D1, six HLIPs, ferredoxin, and plastocyanin
P-SSM4	Myovirus	*Pro* NATL1A, **NATL2A** (LL), *Pro* MED4, and MIT9215 (HL)	D1, D2, and four HLIPs
S-PM2*	Myovirus	*Syn* **WH7803** and WH8109	D1 and D2

Phage family and host-range information is per ref. 15. Boldface indicates the host on which the phage was isolated.
*From Mann *et al.* (12).

plements of the T7- and T4-like phages distinguishes them as distinctly different organisms in whole proteomic taxonomic reconstructions (23).

Phages were propagated on their *Prochlorococcus* hosts (P-SSP7 on MED4, P-SSM2 on NATL1A, and P-SSM4 on NATL2A) and were purified for DNA extraction and construction of clone libraries as described in ref. 8. Briefly, cell lysate was treated with nucleases to degrade host nucleic acids. Phages were precipitated by using polyethylene glycol 8000, purified on a cesium chloride step gradient (steps were $\rho = 1.30, 1.40, 1.50,$ and 1.65) spun at $104,000 \times g$ for 2 h at 4°C, and dialyzed against a buffer containing 100 mM Tris·HCl (pH 7.5), 100 mM MgSO₄, and 30 mM NaCl. Purified phages were burst by using SDS (0.5%) and proteinase K (50 μg/ml). DNA was extracted with phenol:chloroform and concentrated by ethanol precipitation. A custom Los Alamos Scientific Lab clone library was constructed by Lucigen (Middleton, WI) as described in ref. 24. Inserts were sequenced and genomes were assembled by the Department of Energy Joint Genome Institute. Analyses were conducted on the phage genomes as provided on October 17, 2003 (P-SSM2 and P-SSM4), and November 19, 2003 (P-SSP7). At that time, these genomes were in large high-quality contigs compiled from 26-fold (P-SSP7), 30-fold (P-SSM2), and 39-fold (P-SSM4) coverage, respectively.

PCR Amplification of *psbA*. Genomic DNA was isolated from *Prochlorococcus* cultures by using the DNeasy kit (Qiagen, Valencia, CA). Partial *psbA* sequences were amplified by using primers from (19) or for *Prochlorococcus* MIT9211 by using the following primers: 5′-AACATCATYTCWGGTGCWGT-3′ and 5′-TCGTGCATTACTTCCATACC-3′. Reactions (50 μl) consisted of 4 mM MgCl₂, 200 μM dNTP, 0.25 μM (each) primer, 2.5 units of *Taq*DNA polymerase (Invitrogen), and 4 ng of genomic DNA. Amplification conditions, which were run on a RoboCycler Gradient 96 thermocycler (Stratagene), comprised steps at 92°C for 4 min; 35 cycles at 92°C for 1 min, 50°C for 1 min, and 68°C for 1 min; followed by a final extension step at 68°C for 10 min. Fragments were gel-purified and sequenced in both forward and reverse directions (Davis Sequencing, Davis, CA).

Identification of Genes and Transcriptional Regulatory Elements. ORFs in the phage genomes were identified by using GENEMARK (25), and gene identifications were based on homology to known proteins by using the BLASTP program (ftp://ftp.ncbi.nih.gov/blast) with an *E*-value cutoff of 10^{-5}. Ferredoxin-encoding genes (*petF*) were included in our analyses if they encoded the 2Fe–2S iron–sulfur cluster-binding domain (fer2) (with an *E* value $<10^{-10}$ as determined by the BLAST tool RPSBLAST from the conserved domain database of the National Center for Biotechnology Information. High-light-inducible protein (HLIP)-encoding genes (*hli*) were identified as present if they encoded at least six of 10 amino acids in the motif AExxNGRxAMIGF (26). Bhaya *et al.* (27) report that many *Prochlorococcus hli* genes code for a conserved 9-aa C-terminal sequence with the consensus sequence TGQIIPGI/FF. Here this sequence was defined

as present when at least six of the nine conserved amino acids were found.

ρ-Independent transcriptional terminators were identified by using the TRANSTERM program (28), and all had an energy score of <-10 and a tail score of <-5. Potential bacterial σ^{70} promoters were identified in intergenic regions by using the program BPROM (SoftBerry, Mount Kisco, NY). Promoter sequences had a linear discriminant function >2.5. Although identification of terminators is robust, promoter identification in cyanophage is presently more precarious.

Sequence Manipulation and Analyses. Sequences were aligned by using CLUSTALX and edited manually as necessary. Amino acid alignments served as the basis for the manual alignment of nucleotide sequences. Regions that could not be confidently aligned were excluded from analyses, as were gaps. The divergence estimator program K-ESTIMATOR 6.0 (29) was used to estimate the frequency of synonymous and nonsynonymous nucleotide substitutions and employs the Kimura 2p correction method for multiple hits.

PAUP Version 4.0b10 was used for the construction of distance and maximum parsimony trees. Amino acid distance trees were inferred by using minimum evolution as the objective function and mean distances. Heuristic searches were performed with 100 random addition-sequence replicates and the tree-bisection and reconnection branch-swapping algorithms. Starting trees were obtained by stepwise addition of sequences. Bootstrap analyses of 100 resamplings were carried out. Maximum likelihood trees were constructed by using TREE-PUZZLE 5.0. Evolutionary distances were calculated by using the JTT model of substitution (except for the highly divergent HLIPs, for which the VT model of substitution was used) assuming a γ-distributed model of rate heterogeneities with 16 γ-rate categories empirically estimated from the data. Quartet puzzling support was estimated from 10,000 replicates.

For cases in which phylogenetic analyses of small genes received low bootstrap support we used GENERAGE (30) to cluster proteins with significant relationships at user-defined *E*-value thresholds. The input to GENERAGE was an all-against-all table of BLAST comparisons of amino acid sequences. GENERAGE uses a Smith–Waterman dynamic programming alignment algorithm to correct for false positive linkages whenever pairwise relationships are not symmetrical. For HLIPs, an *E*-value cutoff of 10^{-14} was used. The clusters containing the phage HLIPs were preserved down to an *E*-value cutoff of 10^{-17}. For plastocyanin and ferredoxin respectively, *E*-value cutoffs of 10^{-26} and 10^{-34} linked the phage proteins with other proteins, whereas, at *E*-value cutoffs of 10^{-28} and 10^{-36}, the respective phage proteins did not cluster with other sequences.

Results

A suite of host photosynthesis genes was found in the three *Prochlorococcus* phage genomes (Fig. 1). The *psbA* gene, encoding the PSII core reaction center protein D1 (hereafter referred to as the D1-encoding gene) and one *hli* gene type encoding the HLIP cluster 14-type protein (*sensu*, see ref. 27)
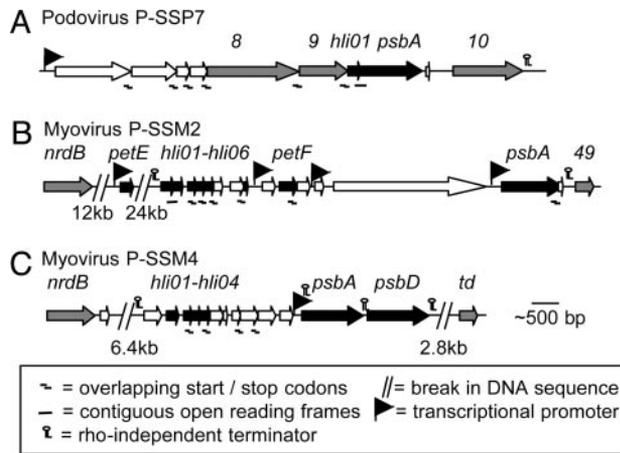
Fig. 1. Arrangement of photosynthesis genes in three *Prochlorococcus* phages. (*A*) Podovirus P-SSP7. (*B*) Myovirus P-SSM2. (*C*) Myovirus P-SSM4. Black bars indicate genes related to photosynthesis, gray bars indicate genes commonly found in phage, and white bars indicate predicted ORFs of unknown function. Genes and their protein designations are as follows: *psbA*, D1; *psbD*, D2; *hli*, HLIP; *petE*, plastocyanin; *petF*, ferredoxin; *8*, T7-like head-to-tail connector; *9*, T7-like capsid assembly protein; *10*, T7-like capsid protein; *nrdB*, T4-like ribonucleotide reductase β-subunit; *49*, T4-like restriction endonuclease VII; and *td*, T4-like thymidylate synthetase.

were present in all three phages. HLIPs are thought to protect the photosynthetic apparatus from excess excitation energy during stressful conditions in cyanobacteria (31). In addition, one of the myoviruses, P-SSM4, contains the *psbD* gene encoding the second PSII core reaction center protein, D2, (hereafter referred to as the D2-encoding gene), whereas the other myovirus, P-SSM2, contains two photosynthetic electron transport genes coding for plastocyanin (*petE*) and ferredoxin (*petF*) (Fig. 1 *B* and *C*). Both myoviruses contain additional gene types from the *hli* multigene family.

The deduced amino acid sequences of the phage photosynthesis genes are highly conserved and therefore have the potential to be functional proteins. The coding sequences of all of these genes are uninterrupted and show a high degree of identity to their host homologs (up to 85% and 95% nucleotide and amino acid identities, respectively; see Table 2 and Figs. 4–8, which are published as supporting information on the PNAS web site). The greatest amino acid divergence in D1 and D2 from all three phages is in the N-terminal leader sequences that do not form part of the functional protein. Furthermore, divergence analyses based on estimates of the frequency of nonsynonymous (K_a) and synonymous (K_s) nucleotide substitutions between phage- and host-encoded genes revealed that the phage genes have diverged relative to those from their hosts (K_s values range from 0.65 to 3.11 and are higher than for *Prochlorococcus* gene pairs; see Table 3, which is published as supporting information on the PNAS web site), but that the majority of nucleotide substitutions did not cause a change in amino acid sequence (K_a/K_s ratios <0.45 for all genes, with values of <0.1 for the D1 and D2 encoding genes; Table 3). Although we cannot rule out the possibility of a recent transfer of these genes from as yet unknown *Prochlorococcus* types with sequences nearly identical to those found in the phages, these findings suggest that the phage-encoded genes, particularly those encoding D1 and D2, have been subjected to strong selective pressure to conserve their amino acid sequences, which is consistent with the hypothesis that they are functional.

All of the photosynthesis genes (with the exception of plastocyanin) are arranged together in the phage genomes. Such gene clustering in phage often suggests that they are expressed

at a similar stage of infection (3, 32). In addition, identification of potential promoter and terminator elements suggests that distinct transcriptional units are present. In the genome of P-SSP7, for example, the *hli* and D1-encoding gene may be cotranscribed with the adjacent phage structural genes in a single operon. Most of the genes in this region have overlapping start and stop codons and are flanked by a putative σ^{70} transcriptional promoter and ρ-independent transcriptional terminator (Fig. 1*A*). This arrangement further suggests that the photosynthesis genes are expressed in the latter portion of the lytic cycle, if indeed they are expressed, as is known for structural proteins in other T7-like podoviruses (32). In contrast, the presence of transcriptional terminators flanking the regions containing photosynthetic genes in the myoviruses suggests that they may be transcribed as discrete transcriptional units largely independent of the surrounding phage genes. These hypotheses require further testing by measuring phage gene expression over the infective cycle.

The cyanobacterial origin of the phage D1- and D2-encoding genes is suggested by the presence of certain features in both phage and host genes. Phage D1 proteins contain a 7-aa indel close to the C terminus of the protein (Fig. 4) which is found in all cyanobacterial D1 proteins as well as in nongreen algal plastids (33). Similarly, phage D2 contains a 7-aa indel in the center of the protein that is also found in *Prochlorococcus* MED4 and SS120 (but not in other cyanobacterial or eukaryotic D2 proteins) (Fig. 5). These additional amino acids are not found in the D2 proteins encoded by either *Synechococcus* WH8102 or the *Synechococcus* phage S-PM2 (Fig. 5), suggesting that *Prochlorococcus* phages acquired the D2-encoding gene from *Prochlorococcus* and that *Synechococcus* phages acquired it from *Synechococcus.*

Phylogenetic analyses of the PSII core reaction center proteins further supports the cyanobacterial origin of the phage genes and, along with knowledge of phage host ranges (15), suggests that they were acquired multiple times from their hosts. Phage D1 and D2 proteins clustered with marine cyanobacteria (Fig. 2). Proteins encoded by *Prochlorococcus* phages clustered with *Prochlorococcus*, whereas those from a phage that infects only *Synechococcus* (12) clustered with *Synechococcus*, as did an environmental sequence (BAC9D04) encoding both D1 and phage structural genes (34). Despite low bootstrap support for *Synechococcus* D1 clades in the distance tree, a similar tree topology also emerged from maximum likelihood and maximum parsimony reconstructions (data not shown). Moreover, D1 from two *Prochlorococcus* phages clustered within *Prochlorococcus* clades that match their host range (Fig. 2*A*). However, D1 from the third *Prochlorococcus* phage did not cluster within a specific *Prochlorococcus* clade, suggesting that its gene was acquired from an as yet uncultured *Prochlorococcus* type or has diverged to an extent that prevents identification of the common ancestor. The fact that the phage D1 and D2 proteins are distributed in both the *Prochlorococcus* and *Synechococcus* clades and are largely consistent with their host range suggests that the genes were acquired in independent transfer events from their cyanobacterial hosts (*sensu*; see refs. 2 and 4). These transfer events could have occurred *de novo* between distinct hosts and phages several times, or these genes may have been transferred from host to phage in a process akin to gene conversion subsequent to an ancestral transfer event (see *Discussion*). If host genes in phages resulted from a single ancestral event followed by subsequent vertical or lateral transfers from phage to phage, the phage- and host-encoded genes would have formed monophyletic clades distinct from each other.

Phylogenetic analyses of plastocyanin proteins also suggests that the phage *petE* gene is of cyanobacterial origin (Fig. 9, which is published as supporting information on the PNAS web site). However, the data are not conclusive as to the origin of the phage
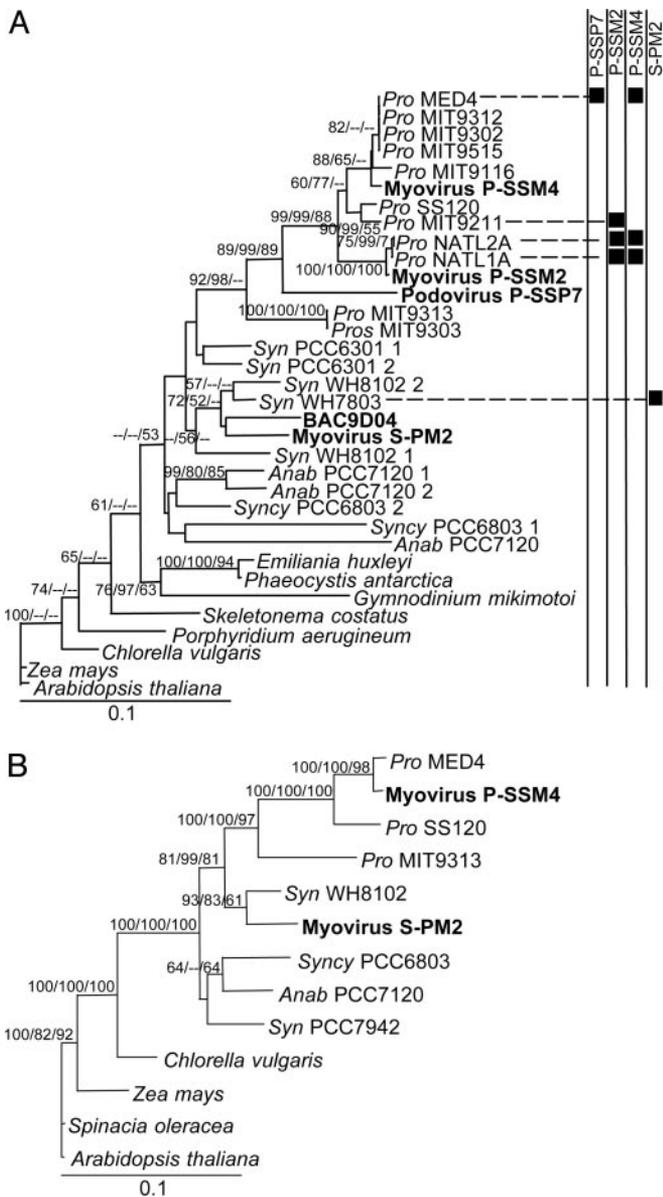
EVOLUTION

A

B

Fig. 2. Distance trees of PSII core reaction center proteins. (*A*) D1 (*psbA*). (*B*) D2 (*psbD*). Phage sequences are shown in bold. The host strains that each phage infects are indicated by black squares. Trees were generated from 244 and 336 amino acids for D1 and D2, respectively (see Figs. 4 and 5). Bootstrap values for distance and maximum parsimony analyses and quartet puzzling values for maximum likelihood analysis >50% are shown at the nodes (distance/maximum likelihood/maximum parsimony). Trees were rooted with *Arabidopsis thaliana* proteins. Essentially, the same topology was obtained when nucleotide trees (third position excluded) were constructed, except for *psbA* from P-SSP7, which clustered with HL *Prochlorococcus*, albeit with low bootstrap support. *Pro, Prochlorococcus*; *Syn, Synechococcus*; *Anab, Anabaena*; *Syncy, Synechocystis.*

gene from within the cyanobacteria. The phage protein clusters with filamentous cyanobacteria, but contains a 10-aa indel found only in unicellular cyanobacteria (Fig. 6). GENERAGE analysis did not resolve the phage plastocyanin clustering. Both phylogenetic and GENERAGE analyses of ferredoxin proteins were inconclusive as to the origin of the phage *petF* gene. These results, together with the greater divergence estimates (K_a/K_s) for the phage and *Prochlorococcus petE* and *petF* gene pairs (0.19–0.43) than among *Prochlorococcus* gene pairs (0.03–0.07) (Table 3), suggest that these phage genes either originated from a host for which

a close relative does not currently exist in the database or have diverged to an extent that prevents inference as to their origin. The latter model may be due to either significant changes in gene sequence or through the formation of mosaic genes from more than one source. These may be new genes in the making.

Previous analyses of HLIPs in cyanobacterial genomes revealed the presence of genetically diverse types, with distinctly different clusters formed for single and multiple copy HLIPs (27). Genes found in a single copy in each of the four sequenced marine cyanobacterial genomes form four distinct clusters (GR C5, C6, C7, and C8 in Fig. 3) that are interspersed with HLIPs from freshwater cyanobacteria in a large cluster (Fig. 3), whereas multicopy *Prochlorococcus* HLIPs are in a separate cluster (Fig. 3). Although bootstrap support for these two broad clusters is low, all three phylogenetic reconstruction methods resulted in the same separation of the multicopy HLIPs from the other HLIPs (Fig. 3 and Figs. 10 and 11, which are published as supporting information on the PNAS web site), lending support to this tree architecture. When we add the phage HLIPs to this analysis, some interesting patterns appear. Ten of 11 phage HLIPs cluster with those that are encoded by multiple gene copies in *Prochlorococcus*, some with more bootstrap support than others. That these phage HLIPs do not group with those from freshwater cyanobacteria nor with the single-copy marine cyanobacterial HLIP types receives greater bootstrap support (Fig. 3). These results were obtained from four different analyses (distance, maximum parsimony and maximum likelihood phylogenetic analyses, and GENERAGE clustering). Indeed, GENERAGE clusters 7 of 11 phage HLIPs with the four HLIP types encoded by multicopy genes in *Prochlorococcus* genomes (GR 10, GR 12, GR 14, and GR 15), with the remaining four of indeterminate affiliation. As for nearly all of the multicopy HLIP sequences from *Prochlorococcus* (28 of 29), all but one of the phage HLIPs contain a 9-aa signature sequence at the C terminus of the protein that is absent from other cyanobacterial HLIPs (27), further supporting a connection between phage *hli* genes and multicopy *hli* genes in the host.

Although the lack of strong bootstrap support for most of the clustering patterns in Fig. 3 makes it impossible to draw definitive conclusions, the fact that both phage and *Prochlorococcus* HLIPs cooccur in four different clusters suggests that it is likely that *hli* genes have been transferred between hosts and their phages multiple times. Moreover, the clustering of phage HLIPs with a subset of the HLIPs that are found exclusively in *Prochlorococcus* suggests that these distinct *hli* gene types may have been reacquired from phage after a period of evolution, leading to the expansion of the *hli* multigene family in this genus.

Discussion

Our findings, along with those by Millard *et al.* (35), indicate that the presence of photosynthesis genes is common, although not universal (13), among phages that infect both HL and LL *Prochlorococcus* and *Synechococcus*. Photosynthesis genes are found in representatives of both the *Myoviridae*, which predominantly infect *Synechococcus* and LL *Prochlorococcus* ecotypes, and *Podoviridae*, which generally infect a single HL *Prochlorococcus* strain (15). The presence of these genes in the members of the latter viral family, which have greater constraints on carrying extra genetic material than members of the former, supports our suggestion that they play a functional role in the phage.

The gene encoding the PSII core reaction center protein, D1, has been found in all phages with photosynthesis genes, suggesting that it plays a particularly significant role. Other photosynthesis genes were more sporadically distributed among the phages. Genes encoding HLIPs were found in all three *Prochlorococcus* phages but in only one of five *Synechococcus* phages (35). In contrast, the gene encoding the second PSII core
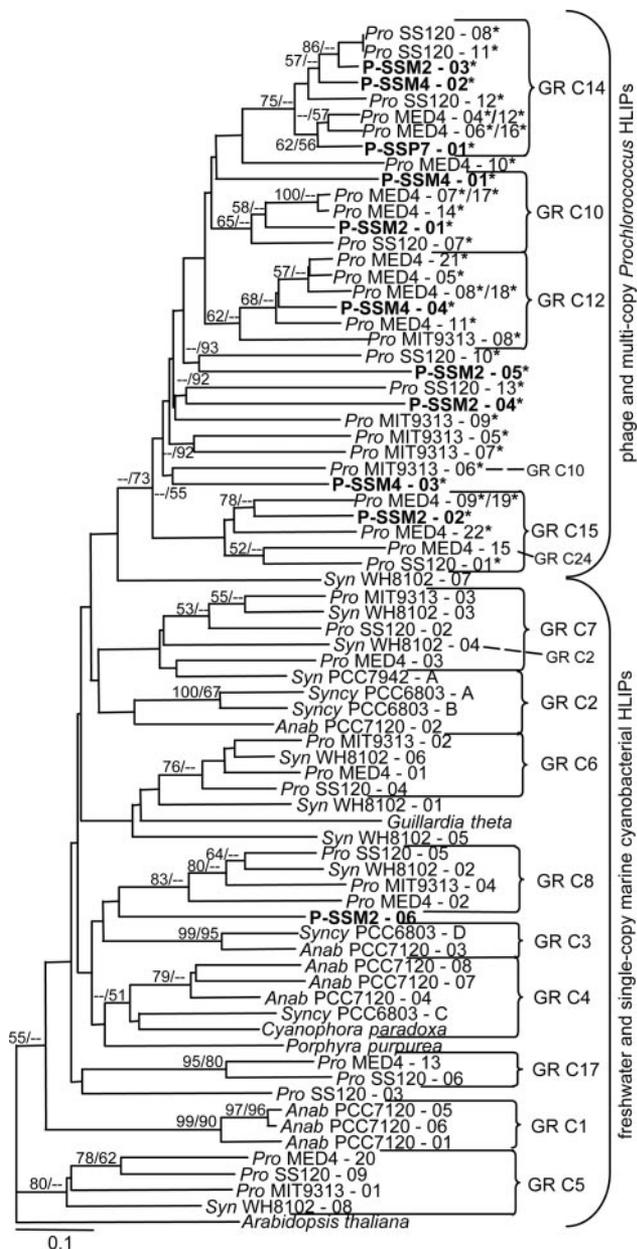
Lindell *et al.*

Fig. 3. Distance tree of HLIPs. Phage HLIPs appear in bold. The tree was generated from 36 amino acids (see Fig. 8), with gaps treated as missing data. GENERAGE clusters are indicated to the right of the tree, with cluster designations following ref. 27. Three discrepancies found between GENERAGE and distance tree clustering are indicated by the dashed line and their GR cluster designations. Asterisks denote proteins encoding at least six of the nine amino acids of the C-terminal 9-aa consensus sequence. Bootstrap and quartet puzzling values >50% are shown at the nodes for distance and maximum likelihood analyses, respectively. The tree was rooted with the single HLIP from *A. thaliana*. Abbreviations are as for Fig. 2.

reaction center protein, D2, was found in all *Synechococcus* phages but in only one *Prochlorococcus* phage. The small number of phage genomes presently available for analysis precludes making strong conclusions from this asymmetry, but if the trend holds up, it is likely that phages gain a differential benefit from these two genes that is influenced by genera-level attributes of their cyanobacterial hosts.

Photosynthetic electron transport genes were found in one *Prochlorococcus* phage and in none of the *Synechococcus* phages,

whereas the transaldolase gene was found both in *Prochlorococcus* myoviruses (M.B.S., F.R., and S.W.C., unpublished data) and in one *Synechococcus* phage (35). Assuming that these genes are functional, this scattered distribution may have arisen from differential gain and loss resulting from tradeoffs between the burden of carrying such genes and their utility during infection. Alternatively, we may be observing the transient passage of host genes through the phage genome pool.

The arrangements of photosynthesis genes in both *Prochlorococcus* and *Synechococcus* phages have some similar properties (compare Fig. 1 of this study with figure 1 of ref. 35), including adjacent D1- and D2-encoding genes, adjacent HLIP- and D1-encoding genes, and the D1-encoding gene adjacent to a T4-like phage gene encoding gp49. These gene organizations are distinctly different from those in cyanobacterial genomes in which photosynthetic genes are spread throughout the chromosome (20–22, 36). Most noticeably, the D1- and D2-encoding genes are hundreds of thousands of kilobases apart in the hosts. Yet phylogenetic analyses show that the D1 and D2 proteins from *Prochlorococcus* phages cluster with those from *Prochlorococcus*, and, in at least the one *Synechococcus* phage available for analysis, these proteins cluster with those from *Synechococcus* (Fig. 2). Assuming that the ancestral cyanobacterial donors of these genes had a similar gene arrangement to extant cyanobacteria, one likely explanation for these findings is that the genes were acquired from their respective hosts in separate transfer events, integrating at recombination hot-spots within the phage genome and forming advantageous gene arrangements. Alternatively, one early transfer event may have occurred, and the observed gene organization patterns formed before the divergence of these phages. In this latter case, for gene sequences to be similar to that from their respective hosts, they would have to have been swapped between phage and host in a process similar to gene conversion, whereby one gene is replaced by another in a nonreciprocal fashion. The direction of this gene conversion for both the D1- and D2-encoding genes is most likely with the host gene replacing the phage gene, as cyanobacterial phylogenies inferred from these gene products are congruent with those from other genes (Fig. 2) (16–19). This latter scenario would suggest that encoding PSII reaction center genes similar to those from the host is advantageous.

The presence of highly conserved PSII reaction center and *hli* genes in the three *Prochlorococcus* phages suggests that selection pressure has driven their acquisition and retention. The presence of these genes is liable to have important implications for phage–host interactions during infection. It has been known for some time that viral infection of many photosynthetic organisms leads to a decline in photosynthetic rates soon after infection (37, 38). This decline is attributed to damage to the PSII membrane–protein complexes (39, 40) and may be due to oxidative stress caused by an increase in destructive reactive oxygen species subsequent to infection (40). Alternatively, the shut-down of host protein synthesis soon after infection (41) could lead to a reduced supply of the highly turned-over D1 and D2 proteins. However, in many phage-infected unicellular freshwater cyanobacteria, the production of phage progeny depends on photosynthetic activity continuing until just before lysis (42, 43). Phage PSII reaction center proteins may, if expressed, prevent photoinhibitory damage to PSII in *Synechococcus* (12). We further suggest that expression of phage PSII reaction center proteins and the photoprotective HLIPs may help maintain photosynthetic activity during infection of *Prochlorococcus*, leading to increased phage fitness and resulting in selection for cyanophages that encode functional photosynthetic genes. Comparing the fitness of a phage with inactivated photosynthetic genes with that of a wild-type phage would enable one to test this hypothesis.

Lindell *et al.*

EVOLUTION

Our analyses of host genes in phages have implications not only for phage fitness but also for the evolution of the hosts, because there is suggestive evidence that phages may have mediated horizontal gene transfer and, hence, expansion of the *hli* multigene family in the hosts. It has recently been suggested that widely distributed, single-copy genes are resistant to horizontal transfer (44), whereas sporadically distributed multicopy genes are those most likely to have been dispersed by this method (44, 45). The clustering patterns displayed by the *hli* genes in our analyses, although not statistically robust, are consistent with this tenant. Each of the single-copy *hli* gene types common to the four sequenced unicellular marine cyanobacteria (20–22) are likely to have been vertically inherited, as is evident from the conserved gene arrangement surrounding these *hli* types and from their clustering to those from the other marine unicellular cyanobacteria (Fig. 3) (27). In contrast, *hli* gene types present in multiple copies per genome are found in only some *Prochlorococcus* genomes. These latter *hli* gene types are those that are found in the *Prochlorococcus* phage, with at least one phage *hli* gene in each of the four clusters of multicopy *Prochlorococcus hli* gene types (Fig. 3). We therefore suggest that phages have mediated the horizontal dispersal of these multicopy genes among *Prochlorococcus*.

The presence of numerous *hli* genes in *Prochlorococcus* MED4, a HL ecotype, is likely to have influenced its fitness in the surface waters of the open oceans (20, 27, 36). Indeed, upon shifts to high light, cyanobacterial mutants with inactivated *hli* genes are competitively inferior to wild-type cells (31). Our hypothesized phage-mediated expansion of the *hli* multigene family may have contributed to the numerical dominance of the HL ecotype in many ocean ecosystems (46). Other photosynthetic genes found in phages are also present in multiple copies in many cyanobacteria, including the D1-, D2-, and ferredoxin-encoding genes (Table 4, which is published as supporting information on the PNAS web site). The importance of gene duplication in the evolution of new gene functions is well recognized in other systems (47, 48); thus, it would not be surprising if it were playing a role in the evolution of physiological variants within the *Prochlorococcus* cluster.

The exchange of photosynthetic genes between *Prochlorococcus* and their phages could have significant implications for the evolutionary trajectory of both hosts and phages and may represent a more general phenomenon of metabolic facilitation of key host processes. That is, host genes retained in a particular phage could reflect key selective forces in the host environment. Indeed, phosphate sensing and acquisition genes have been found in phages that infect organisms in low phosphate environments (8, 9). Might we also find salt tolerance genes in phages that infect halotolerant organisms and thermal tolerance genes in phages that infect thermophilic organisms? Such coupled evolutionary processes in hosts and phages, if widespread, may play a role in defining host ranges for phages and niche space for hosts, leading to specialization and even speciation.

1. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. (2000) *Trends Microbiol.* **8,** 504–508.
2. Filee, J., Forterre, P. & Laurent, J. (2003) *Res. Microbiol.* **154,** 237–243.
3. Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Ruger, W. (2003) *Microb. Mol. Biol. Rev.* **67,** 86–156.
4. Moreira, D. (2000) *Mol. Microbiol.* **35,** 1–5.
5. Wagner, P. L. & Waldor, M. K. (2002) *Infect. Immun.* **70,** 3985–3993.
6. Casjens, S. (2003) *Mol. Microbiol.* **49,** 277–300.
7. Forterre, P. (1999) *Mol. Microbiol.* **33,** 457–465.
8. Rohwer, F., Segall, A., Steward, G., Seguritan, V., Breitbart, M., Wolven, F. & Azam, F. (2000) *Limnol. Oceanogr.* **45,** 408–418.
9. Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., Feldblyum, T. V., White, O., Paulsen, I. T., Nierman, W. C., *et al.* (2003) *J. Bacteriol.* **185,** 5220–5233.
10. Figueroa-Bossi, N. & Bossi, L. (1999) *Mol. Microbiol.* **33,** 167–176.
11. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., *et al.* (2003) *Cell* **113,** 171–182.
12. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. (2003) *Nature* **424,** 741.
13. Chen, F. & Lu, J. (2002) *Appl. Environ. Microbiol.* **68,** 2589–2594.
14. Waterbury, J. B. & Valois, F. W. (1993) *Appl. Environ. Microbiol.* **59,** 3393–3399.
15. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. (2003) *Nature* **424,** 1047–1051.
16. Ferris, M. J. & Palenik, B. (1998) *Nature* **396,** 226–228.
17. Moore, L. R., Rocap, G. & Chisholm, S. W. (1998) *Nature* **393,** 464–467.
18. Rocap, G., Distel, D., Waterbury, J. B. & Chisholm, S. W. (2002) *Appl. Environ. Microbiol.* **68,** 1180–1191.
19. Zeidner, G., Preston, C. M., Delong, E. F., Massana, R., Post, A. F., Scanlan, D. J. & Beja, O. (2003) *Environ. Microbiol.* **5,** 212–216.
20. Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., *et al.* (2003) *Nature* **424,** 1042–1047.
21. Palenik, B., Brahamsha, B., McCarren, J., Waterbury, J., Allen, E., Webb, E. A., Partensky, F. & Larimer, F. W. (2003) *Nature* **424,** 1037–1041.
22. Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10020–10025.
23. Rohwer, F. & Edwards, R. (2002) *J. Bacteriol.* **184,** 4529–4535.
24. Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. & Rohwer, F. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14250–14255.
25. Besemer, J., Lomsadze, A. & Borodovsky, M. (2001) *Nucleic Acids Res.* **29,** 2607–2618.
26. Jansson, S., Andersson, J., Kim, S. J. & Jackowski, G. (2000) *Plant Mol. Biol.* **42,** 345–351.
27. Bhaya, D., Dufresne, A., Vaulot, D. & Grossman, A. (2002) *FEMS Microbiol. Lett.* **215,** 209–219.
28. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. (2000) *J. Mol. Biol.* **301,** 27–33.
29. Comeron, J. M. (1999) *Bioinformatics* **15,** 763–764.
30. Enright, A. J. & Ouzounis, C. A. (2000) *Bioinformatics* **16,** 451–457.
31. He, Q., Dolganov, N., Bjorkman, O. & Grossman, A. R. (2001) *J. Biol. Chem.* **276,** 306–314.
32. Dunn, J. J. & Studier, F. W. (1983) *J. Mol. Biol.* **166,** 477–535.
33. Hess, W. R., Weihe, A., Loiseaux-de Goer, S., Partensky, F. & Vaulot, D. (1995) *Plant Mol. Biol.* **27,** 1189–1196.
34. Zeidner, G. & Beja, O. (2004) *Environ. Microbiol.* **6,** 528–534.
35. Millard, A., Clokie, M. R. J., Shub, D. A. & Mann, N. H. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 11007–11012.
36. Hess, W. R., Rocap, G., Ting, C. S., Larimer, F. W., Stilwagen, S., Lamerdin, J. & Chisholm, S. W. (2001) *Photosyn. Res.* **70,** 53–71.
37. Padan, E. & Shilo, M. (1973) *Bacteriol. Rev.* **37,** 343–370.
38. Zaitlin, M. & Hull, R. (1987) *Ann. Rev. Plant Physiol.* **38,** 291–315.
39. Rahoutei, J., Garcia-Luque, I. & Baron, M. (2000) *Physiol. Plant.* **110,** 286–292.
40. Arias, M. C., Lenardon, S. & Taleisnik, E. (2003) *J. Phytopathol.* **151,** 267–273.
41. Sherman, L. A. & Pauw, P. (1976) *Virology* **71,** 17–25.
42. Mackenzie, J. J. & Haselkorn, R. (1972) *Virology* **49,** 517–521.
43. Sherman, L. A. (1976) *Virol.* **71,** 199–206.
44. Lerat, E., Daubin, V. & Moran, N. A. (2003) *PLoS Biol.* **1,** e19.
45. Kurland, C. G., Canback, B. & Berg, O. G. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9658–9662.
46. Partensky, F., Hess, W. R. & Vaulot, D. (1999) *Microbiol. Mol. Biol. Rev.* **63,** 106–127.
47. Zhang, J. (2003) *Trends Ecol. Evol.* **18,** 292–298.
48. Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. & Li, W. (2003) *Nature* **421,** 63–66.