Gene Network Visualization and Quantitative Synteny Analysis of more than 300 Marine T4-Like Phage Scaffolds from the GOS Metagenome

André M. Comeau,†,[1,2] Christine Arbiol,[1,2,3] and H.M. Krisch*,[1,2]

[1]Centre National de la Recherche Scientifique, UMR5100, Toulouse, France

[2]Laboratoire de Microbiologie et Génétique Moléculaires, Université de Toulouse, Université Paul Sabatier, Toulouse, France

[3]Institut d'Exploration Fonctionnelle des Génomes, Centre National de la Recherche Scientifique, IFR109, Toulouse, France

†Present address: Institut de Biologie Intégrative et des Systémes/Québec-Océan, Department of Biology, Université Laval, Québec, Canada.

***Corresponding author:** E-mail: krisch@ibcg.biotoul.fr.

Associate editor: Hervé Philippe

Research article

Abstract

Bacteriophages (phages) are the most abundant biological entities in the biosphere and are the dominant "organisms" in marine environments, exerting an enormous influence on marine microbial populations. Metagenomic projects, such as the Global Ocean Sampling expedition (GOS), have demonstrated the predominance of tailed phages (Caudovirales), particularly T4 superfamily cyanophages (Cyano-T4s), in the marine milieu. Whereas previous metagenomic analyses were limited to gene content information, here we present a comparative analysis of over 300 phage scaffolds assembled from the viral fraction of the GOS data. This assembly permits the examination of synteny (organization) of the genes on the scaffolds and their comparison with the genome sequences from cultured Cyano-T4s. We employ comparative genomics and a novel usage of network visualization software to show that the scaffold phylogenies are similar to those of the traditional marker genes they contain. Importantly, these uncultured metagenomic scaffolds quite closely match the organization of the "core genome" of the known Cyano-T4s. This indicates that the current view of genome architecture in the Cyano-T4s is not seriously biased by being based on a small number of cultured phages, and we can be confident that they accurately reflect the diverse population of such viruses in marine surface waters.

Key words: T4-like phage, cyanophage, metagenome, synteny.

Introduction

Bacteriophages, or phages, are the most abundant biological entities in the biosphere ($\sim 10^{30}$ virions) and are the dominant "organisms" in marine environments where they exert a substantial influence on the ecosystem (see Suttle 2007 for a review). Ninety-six percent of the more than 5 500 phages that have been described in the literature are dsDNA tailed phages (Caudovirales; Ackermann 2007). Analyses of marine environments by electron microscopy (Frank and Moebus 1987), by sequencing of PCR-amplified conserved marker genes (Short and Suttle 2005), and by metagenomics (Breitbart et al. 2002; Angly et al. 2006) confirm the preponderance of the Caudovirales. Simultaneously, such studies point out the extraordinary diversity of the "virosphere," with literally millions of different viruses that have never been studied.

The first marine phage whose genome was fully sequenced was a small short-tailed T7-like phage infecting *Roseobacter* (Rohwer et al. 2000). The genome sequences of a cyanophage (Chen and Lu 2002) and a vibriophage (Hardies et al. 2003) that belong to the same Podoviridae family rapidly followed. Since then, many additional genomes have been sequenced, most notably those of the much larger contractile-tailed (Myoviridae family) T4-like cyanophages (Cyano-T4s) infecting *Synechococcus* (S-PM2 and Syn9) (Mann et al. 2005; Weigele et al. 2007) and *Prochlorococcus* (P-SSM2 and P-SSM4; Sullivan et al. 2005). Comparative analyses of these and other genomes have revealed the existence and importance of a set of phage "superfamilies" (Hardies et al. 2003; Comeau, Bertrand, et al. 2007), and we are beginning to identify a pool of "marine-like" genes (mostly of unknown function) that are specifically associated with these marine phages and the marine metagenomes (Angly et al. 2006).

Large-scale marine metagenomic projects, such as the Global Ocean Sampling expedition (GOS; Rusch et al. 2007), have given us access to unparalleled quantities of sequence data on environmental phage populations (Comeau et al. 2008). The GOS surface water data has revealed that the Cyano-T4s were the predominant phage type based on gene content analysis (Comeau and Krisch 2008; Williamson et al. 2008). The size distribution of the phage genomes in these samples was not characterized, and the T4-like phages are among the largest. This fact alone could contribute (although modestly) to the

disproportionately high fraction of T4-like genes present in the samples. However, much more useful information could be gained by knowing the genomic context (synteny) in which these phage genes are found. Such an analysis is now feasible because the vast quantities of viral-fraction GOS sequence allowed these random small segments of phage genomes to be assembled into larger-sized scaffolds that have been made available through the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis database (http://camera.calit2.net/). These scaffolds, which were generated in a non–culture-dependent manner, permit a comparative analysis with the few cultivated Cyano-T4s whose genomes have been sequenced. The four Cyano-T4 genomes that have been analyzed to date (Mann et al. 2005; Sullivan et al. 2005; Weigele et al. 2007) have a common genome architecture—conserved replication and virion modules (i.e., the "core genome") that are separated by hyperplastic regions (HPRs) that contain much of the sequence divergence in the phage genomes, including most of the genes believed to be responsible for phage adaption to new hosts and environments (Comeau, Bertrand, et al. 2007). As an illustration of such adaptation in the T4 superfamily phages, it is within such HPRs that the host photosynthesis genes were discovered in the Cyano-T4 phage S-PM2 (Mann et al. 2003).

Our objective was to determine whether the GOS metagenome scaffolds reflected the general organization of known cultured Cyano-T4 genomes or whether our current view of genome architecture of these phages is seriously distorted because of a sampling bias due to the small number of genomes that have been sequenced. A variety of methods, including a novel use of network visualization software, demonstrated very strong synteny among the core genome components assembled from metagenomic data and the genome sequences of the cultured Cyano-T4 phages. These results indicated that the small number of T4 superfamily genome sequences already in the databases share a common architecture with a vast number of T4-like phages present in marine surface waters.

Materials and Methods

The GOS-Assembled Scaffolds and Processing

The >5-kb GOS viral scaffolds assembled by Williamson et al. (2008) ($n = 440$) were obtained from the CAMERA database (http://camera.calit2.net/). Only a few of the total 314 myovirus scaffolds were further processed to remove stretches of polyN >1 kb . These stretches, present in only 31 of the scaffolds (10%), are due to the "mating" of sequences originating from both sides of a same clone (mate pairs), so that the size of the intervening nonsequenced space is known (see Rusch et al. 2007 and Kunin et al. 2008). However, these polyN stretches give us no information on gene content and were therefore ignored. This ended up not affecting the synteny of 7 of the 31 scaffolds thus treated because there were no known genes to the left or right of their polyN gaps. For the remaining 24 (7% of the 314 total), the polyN stretches may or may not imply a break in synteny; therefore, these 24 links between the genes to the left and right of the polyN gaps, representing a negligible portion (1.2%) of the total gene interactions ($n = 1\,989$), were conservatively considered to be synteny breaks in the analyses described below. The 314 polyN-removed scaffolds are available as supplementary file S1 (Supplementary Material online; FASTA format) or upon request from the authors.

Scaffold Content Analysis

Content was manually determined for each scaffold using the basic local alignment search tool (BLAST) at the National Center for Biotechnology Information (NCBI) Web site (http://blast.ncbi.nlm.nih.gov) with an E value cutoff of $<10^{-4}$ against the nr database, using the top hit as protein identity, with occasional restriction to viral genes for a more informative origin/function determination. We used BlastX (bacterial genetic code) directly on scaffolds of small size (generally <8 kb). Otherwise, GeneMark Heuristic approach (http://exon.gatech.edu/GeneMark/; Besemer and Borodovsky 2005) was used for open reading-frame (ORF) determination, followed by BlastP of the resulting ORF translations. The few sequences with obvious frameshifts (primarily single base pair–sequencing errors) were corrected, whereas more ambiguous gene splits were left unchanged. Gp20 portal proteins from the Cyano-T4 genomes (NCBI) were queried against all GOS proteins through the CAMERA database (http://camera.calit2.net/) with an E value cutoff of $<10^{-4}$. The Cyano-T4 genome coverage circular visualizations were generated using CGView (Stothard and Wishart 2005).

Scaffold Synteny Visualization and Analysis

Traditional comparative genomic representations become increasingly unwieldy when the number of objects under consideration becomes large. For example, whereas traditional dot plots are efficient for comparing up to a few dozen genomes (e.g., Hatfull et al. 2010), they are less useful in this case of hundreds of scaffolds as they focus more on the "DNA sequence" (good for showing indels, inversions, etc.) versus the "gene," which is our focus. In this analysis, we needed to represent the synteny of >300 scaffolds containing nearly 1 800 genes/ORFs. To do this, we applied the open-source Cytoscape program (version 2.6.1, http://www.cytoscape.org; Shannon et al. 2003) that has been developed for the visualization of complex metabolic pathways and "interactomes" to deal with the similar presentational problems posed by scaffold synteny. These visualizations convert multiple occurrences of the same gene/ORF to one "node" with multiple "links" (synteny) to its neighboring genes/ORFs. There are multiple advantages to using such well-established "network" programs, among which include 1) the capacity to handle very large data sets; 2) great flexibility in visualization control; and 3) the "compaction" of data into a smaller visual space for ease of analysis and presentation. Figure 1 illustrates how a traditional "arrow" representation of a genome (panel *A*) can be translated into a network of nodes and edges (panel *B*). The space required to represent the data is considerably reduced. For example, insertions

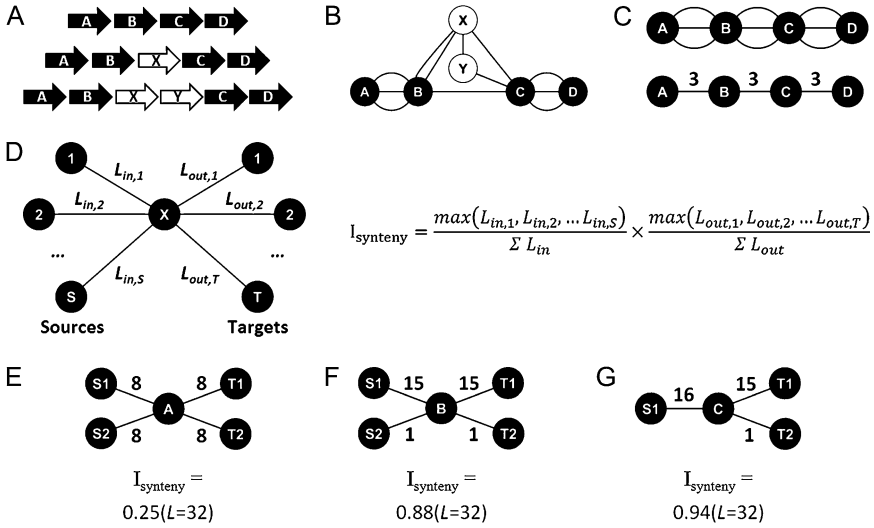

FIG. 1. Synteny representations and $I_{synteny}$ calculations. (*A*) Traditional arrow gene representations, with "core genes" in black and inserted novel ORFs in white. (*B*) Conversion of (*A*) into network representation (used by Cytoscape), with each line representing an occurrence/link between the respective genes/ORFs. (*C*) Reduction of (*B*), with the removal of non-core genes, showing either all (top) or a condensation (bottom) of the number of links. (*D*) Formulation of the Index of Synteny ($I_{synteny}$) (equation on right), which reports the proportion of links (L) to the left (in) and right (out) of a gene X that are to single sources (S) and targets (T). (*E*–*G*) Various examples of gene synteny, along with the corresponding $I_{synteny}$ values and the total number of links (under the format $L = n$).

of novel ORF database orphans (ORFans; white nodes) within groups of "core genes" (black nodes) can also be further removed in order to simplify and represent the "core synteny" (panel *C*). This type of pruning, which removes intervening genes/ORFs that are not part of the core genome, answers the fundamental question of whether (following fig. 1*A*–*C*) core gene *B* is invariably downstream of *A*, and upstream of *C*, regardless of the expansion/contraction of this genome by the addition/removal of the facultative intervening ORFs.

Finally, we wanted to quantitatively express the amount of synteny of each gene/ORF on the scaffolds. In order to do so, we developed an "Index of Synteny" that expresses the proportion of links to the left (in) and right (out) of a gene/ORF that are to single sources and targets (perfect synteny being only one source and one target, i.e., $I_{synteny} = 1$):

$$I_{synteny} = \frac{\max(L_{in,1}, L_{in,2}, \ldots, L_{in,S})}{\sum L_{in}} \times \frac{\max(L_{out,1}, L_{out,2}, \ldots, L_{out,T})}{\sum L_{out}},$$

where L is the number of links to adjacent source (S) and target (T) genes/ORFs (fig. 1*D*). We can also display the total number of links involved, under the form $I_{synteny}(L = n)$, where n is the sum of L_{in} and L_{out}, as an indication of the confidence we have in the measure. For example, gene *B* in figure 1*C* would have a value of 1.0($L = 6$), indicating that it is in perfect synteny with its surrounding genes, but the sample size is small (only three occurrences in this example) and so the perfect synteny score must be interpreted with caution. Genes *A*–*C* of panels *E*–*G*, with values of 0.25, 0.88, and 0.94, have varying degrees of synteny that are <1, but these measures are derived from sufficient scaffold occurrences ($L = 32$) that they merit much greater confidence.

Protein Phylogenies

The conserved marker proteins (gp20, gp23, gp43) of the Cyano-T4s (and T4 as the outgroup) were aligned using ClustalW (Thompson et al. 1994) within BioEdit v7.0.9.0 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). These alignments were used to construct Neighbor-Joining phylogenies using QuickTree (Howe et al. 2002), which uses the ClustalW distance calculations using default parameters (neither column rejection nor multiple substitution correction), with 1 000 bootstrap replicates as implemented on the Institut Pasteur Mobyle web portal (http://mobyle.pasteur.fr).

Results and Discussion

The Marine Phage Scaffolds Extracted from the GOS Metagenome

We chose to examine only those scaffolds >5 kb in length from the pooled GOS data assembly conducted by Williamson et al. (2008). The majority of the scaffolds from their assembly were smaller (containing only a few genes), but these were judged to be of insufficient length for unambiguously assigning a viral origin or determining gene synteny. Local similarities (BlastX) to the 440 large viral scaffolds indicated that the vast majority (n = 314) were of probable myovirus origin, whereas there were 18 podovirus scaffolds and only 8 of likely siphovirus origin. Thus, there were a total of 340 scaffolds considered to be of probable phage origin, with the remaining 100 viral scaffolds mostly distributed among the phytoplankton- and protist-infecting phycodnaviruses (70 scaffolds) and mimivirus (25 scaffolds). We removed four unconvincing myovirus scaffolds (each containing a single "phage" gene with greater similarity to a bacterial gene) from our analysis, leaving 310 scaffolds ranging in size from 2.5 to nearly

Table 1. GOS Myovirus Scaffold Statistics (ignoring polyN mate-pair gaps).

Feature	Value
Size/base composition	
Total sequence length (Mb)	2.171
Number of scaffolds	310
Scaffold size range (kb)	2.5–29.5
Scaffold (kb)	
Mean size	7.0
Median size	5.9
Scaffolds (kb)	
<5	20
5–10	260
>10	30
%G+C mean (range)	37.2 (29.1–57.3)
Content	
Number of scaffolds with cellular hits[a] (% total)	107 (35)
Number of hits[a]	
T4-like phages	1 581 (1 064/517)[b]
Other phages	12 (11/1)
Eukaryotic viruses	3 (2/1)
Eubacteria	167 (84/83)
Archaea	4 (3/1)
Eukaryotes	12 (9/3)
Total	1 779 (1 173/606)

[a] Hits were determined using BlastX and BlastP against the nr database with an *E* value $<10^{-4}$.
[b] Numbers in parentheses indicated the known/unknown function hits.

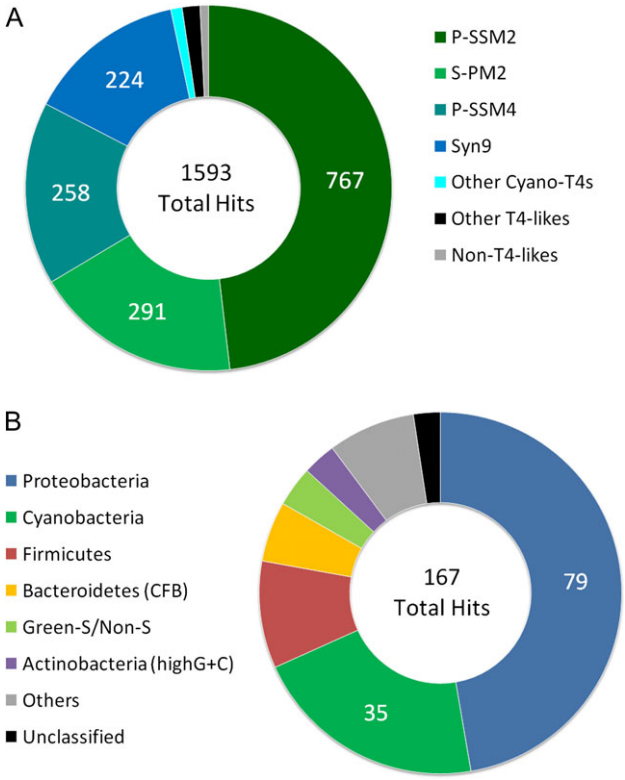

Fig. 2. GOS myovirus scaffold content. Taxonomic distribution of the phage (*A*) and eubacterial hits (*B*).

30 kb (ignoring the polyN gaps, see Materials and Methods), with the majority being in the 5- to 10-kb size range (table 1). The G+C content of these scaffolds ranged between 30% and 45%, which agrees with what has been reported (Miller, Heidelberg, et al. 2003; Mann et al. 2005; Sullivan et al. 2005; Comeau, Bertrand, et al. 2007; Weigele et al. 2007) for the T4 superfamily: 33–42% for coliphages, 35–40% for cyanophages, and 41–43% for *Aeromonas*/*Vibrio* phages.

BLAST analysis of the gene content of the myovirus scaffolds confirmed their T4-like character as the overwhelming proportion of hits were to phage genes of the T4 superfamily (1 581 of 1 779 total; 89%), with other phage types accounting for merely 12 hits (table 1). The only other significant source of genes/ORFs was the Eubacteria domain, corresponding to only ~10% of the hits. For the phage hits, *Prochlorococcus* phage P-SSM2 genes were the most frequently represented in the scaffolds, accounting for nearly half of the hits (fig. 2*A*). These were followed by genes from the three other sequenced Cyano-T4s in nearly equal proportions—*Prochlorococcus* phage P-SSM4 and *Synechococcus* phages S-PM2 and Syn9. The four cyanophages represented ~97% of the best phage hits, which, in combination with our previous analysis of the major capsid protein (MCP) (Comeau and Krisch 2008), confirms the overwhelming abundance of cyanophage from the T4 superfamily in the marine environment. There were 167 eubacterial hits within the phage scaffolds (fig. 2*B*), distributed largely among the Proteobacteria (half in the alpha class) and the cyanobacteria (mostly *Prochlorococcus* and *Synechococcus*). The major classes of identifiable eubacterial gene/ORF functions were permeases/transport channels, proteins involved in various iron and phosphate functions, carbohydrate metabolism and DNA-modifying (nucleases, methylases) enzymes, and bacterial-encoded proteins of possible phage origin (supplementary table S1, Supplementary Material online).

Coverage of the Cyano-T4 Genomes by the T4-Like Scaffolds

The T4-like scaffolds covered much of the sequenced Cyano-T4 genomes in the databases but not in a completely random fashion. There was good coverage of the replication and virion structural modules. As expected, because of the inherent difficulties in assembling and correctly identifying scaffolds from variable HPRs, there was much more limited coverage in these regions of the genomes (supplementary fig. S1, Supplementary Material online). In examining the most highly represented proteins on the T4-like scaffolds (≥10 hits), one sees twice as many structural protein hits as replication/recombination proteins and only a small number of stress/metabolism proteins typically found in the Cyano-T4 genomes (supplementary table S2, Supplementary Material online). The large discrepancy in favor of structural genes correlates with the fact that the T4 superfamily phage particle is a remarkably complex nanomachine (Leiman et al. 2003), and consequently, the T4-like phages dedicate a significant proportion (~40%) of their large genomes to encoding its constituents (Miller, Kutter, et al. 2003). Among the T4-type virion proteins, the gp7 baseplate component was the most prevalent sequence on the scaffolds; this is probably

due to its large size (>1 000 aa), which would enhance the importance of its contribution to metagenomes. Other frequent inhabitants of the scaffolds were the genes encoding the conserved phylogenetic markers often used for T4 superfamily classification, such as the gp20 portal vertex protein and the gp23 MCP. These genes have been targets for PCR amplification of T4-like sequences in the environment (Zhong et al. 2002; Dorigo et al. 2004; Filée et al. 2005; Short and Suttle 2005; Jia et al. 2007; Comeau and Krisch 2008). As for the replication/recombination proteins, the gp17 terminase large subunit and the UvsW helicase were the most commonly encountered, with the gp16 terminase small subunit being only half as frequent. The photosynthesis genes that were prevalent in the analysis of Williamson et al. (2008) were primarily located on the small scaffolds (<5 kb) assembled from the GOS data and were rarer on the selected large scaffolds that we analyzed here. This correlates with the fact that these host-acquired genes are often located in the HPRs of cyanophage genomes (Mann et al. 2005; Sullivan et al. 2005; Sharon et al. 2009), which, as mentioned above, can seldom be assembled into large scaffolds due to the extremely variable context in which they are found.

The Synteny of the Metagenomic Scaffolds and the Cultured Cyano-T4 Genomes

The raw metagenomic data provides information about both the abundance and the distribution of phage genes in the environment. The phage scaffolds assembled from such metagenomic data allow us to analyze gene synteny, the arrangement of the genes in these environmental genomes, and from this gain some valuable insights about the different phage types represented in the metagenome and their relative frequencies. The extraction of such data is greatly facilitated by having the full genomic sequence of at least a few of the major types of genomes contained in the environmental samples. Obviously, such an analysis depends on the validity of the assumption that the assembled genome scaffolds accurately reflect the actual genomes present in the samples, rather than in silico artifacts. Williamson et al. (2008) and Rusch et al. (2007) argue that the GOS scaffolds, created with the Celera Assembler using, among other parameters, a stringent overlap cutoff of 98% identity over 14–40 bp, accurately reflect the molecules present in the samples. However, Kunin et al. (2008) have described the various problems that are confronted by metagenome assembly procedures, including those of the Celera Assembler, and they strongly caution against outright acceptance of assembled metagenomic data. A potential problem with the GOS viral assembly is that the sequence data from all sample locations were pooled together before assembly. Thus, "near-identical" genomes present in different samples could have been assembled into "chimeric" scaffolds. This "composite" strategy increases the chances of assembling larger scaffolds from genomes that are widespread in nature but at the cost of losing the geographic resolution of the data, an aspect that we did not investigate here. Thus, our analysis of the scaffolds may have some limitations from a geographic point of view but not from a gene content perspective. Williamson et al. (2008) concede that a small fraction of the scaffolds probably reflect assembly errors. Based on our analysis, we believe that a good fraction of the phage scaffolds are correct. Regardless of the exact level of incorrect assembly or noise from geographic pooling, these limitations do not invalidate the major trends that can be extracted from the data as a composite representation of surface water genomes.

Conserved Marker Genes and Scaffold Origins

The primordial question that concerns us is the phylogenetic relation between the different genes that populate individual scaffolds. Because there is such a high degree of vertical transmission of the T4-type core genome, it would be expected that the various essential genes on each of the Cyano-T4 scaffolds would be phylogenetically closely related. This assumption is particularly important in regard to the genes used as phylogenetic markers for the various subgroups of T4 superfamily phages—the *g23* MCP gene, *g20* coding the portal vertex protein, and *g43* encoding the phage's DNA polymerase. As mentioned above, the first two genes have been used to create PCR primers that have been employed for the analysis of the diversity of the various subgroups of T4-like phages present in the environment (e.g., T-evens, Pseudo T-evens, Cyano-T4s, etc.) (Zhong et al. 2002; Dorigo et al. 2004; Filée et al. 2005; Short and Suttle 2005; Jia et al. 2007; Comeau and Krisch 2008). The critical question is therefore reduced to the following: If a Cyano-T4 marker gene is contained in the metagenomic data set, is it residing on a Cyano-T4 scaffold and, by extension, a valid indication of the presence of a Cyano-T4 phage in the environment?

The straightforward answer to this question is yes—on a "coarse" scale (separating T4-like subgroups), the scaffold phage genes in the three marker regions ($n = 535$) were almost exclusively (99%) of Cyano-T4 origin. There was only one gene from a non–T4-like phage (myovirus BcepB1A) and a few cellular hits and ORFans. There was "fine-scale mosaicism" in the sense that the scaffolds, and the cultured Cyano-T4 genomes themselves, are mixtures of different "specific" Cyano-T4–type genes (e.g. Syn9-like genes may be next to P-SSM2-like genes, etc.). However, one phylogenetic isolate is generally predominant; for example, most of the GOS scaffolds were P-SSM2 like in the marker regions. The most "faithful" marker gene was *g23*—only four, often small, scaffolds (of 20 total) had MCPs whose phylogenetic origins did not precisely match with the majority of the scaffold's genes (fig. 3).

Superficially, the situation for the *g20* portal vertex protein gene appeared to be quite different—approximately a third of the *g20* scaffolds were populated by genes coming from sources that were different than their *g20* sequences (supplementary fig. S2, Supplementary Material online). The same phenomenon is also manifested by all four of the available cultured Cyano-T4 genomes. All these odd Cyano-T4 hits were the consequence of two variant *g20* sequences that come from *Synechococcus* cyanophages

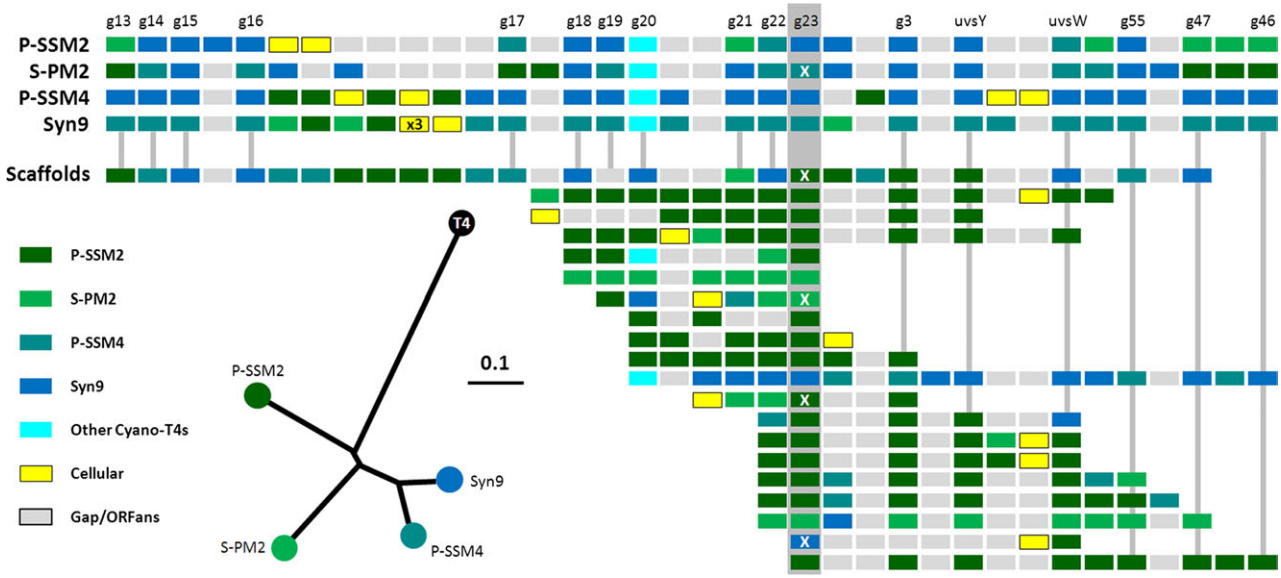

FIG. 3. GOS scaffolds containing the *g23* MCP gene. Schematic of the scaffolds containing *g23*, with each gene/ORF (rectangles) colored to match its origin type, compared with the cultured Cyano-T4 genomes in this region. Those *g23s* that do not represent the majority origin type on the scaffolds or genomes are marked with an "X." Due to space considerations, some of the genes/ORFs of known origin have been condensed (e.g., "x3"), and some ORFans have been omitted. The phylogenetic tree inset shows the relationships among the Cyano-T4 gp23 proteins, with T4 as the outgroup and the scale bar indicating 0.1 substitutions per site.

S-BnM1 (nine hits) and S-WHM1 (four hits). The complete genome sequence of neither of these cyanophages is known, but their *g20* genes were sequenced over a decade ago in order to design consensus cyanomyovirus-specific *g20* PCR primers (Fuller et al. 1998). So, does this suggest that *g20* has undergone substantial lateral gene transfer, contrary to the in-depth phylogenetic findings of Filée et al. (2006)? We do not believe this to be the case—all the anomalies in the *g20* scaffolds could have a trivial yet extremely interesting explanation. When the genome sequences of S-BnM1 and S-WHM1 become available, we believe that most of the aberrant scaffolds would no longer be assigned to their current types but (for the majority) to the S-BnM1 type. Indeed, the gp20 protein sequence of phage S-BnM1 is sufficiently divergent from the other Cyano-T4s ($\leq$60% identity) that it seems likely it represents a phylogenetically distinct phage type (supplementary fig. S2, inset tree, Supplementary Material online). In order to determine just how predominant S-BnM1 (and S-WM1) may be, we can interrogate the GOS metagenome by BLASTing the CAMERA database (*E* value $< 10^{-4}$) to see how many hits each of these *g20* sequences receives. Remarkably, S-BnM1 received the most hits (5 852), followed very closely by P-SSM2 (5 834), Syn9 (5 770), S-PM2 (5 764), P-SSM4 (5 646), and finally by the other potentially new phage-type S-WM1 (5 612). By comparison, T4 gp20 protein generates only ~3 500 hits (Comeau and Krisch 2008). These data imply that S-BnM1-like phages may be one of the most prevalent types of cyanophages in marine surface waters, perhaps even more so than the P-SSM2 type.

The *g43* DNA polymerase (*pol*) gene, which has been used for deep phylogenetic analysis (Karam and Konigsberg 2000; Filée et al. 2002), gave scaffolds whose profiles were intermediate between those of the MCP and the portal gene (supplementary fig. S3, Supplementary Material online). While they clearly remained dominated by Cyano-T4–type content on a coarse scale, nearly half of the *pol*-containing scaffolds (and two of four complete genome sequences) had *g43* genes that were not in perfect fine-scale concordance with their neighboring genes. Given the lack of evidence for *g43* horizontal gene transfer (Filée et al. 2006), these results are probably due to a specific feature of the genomic context where *g43* is located. Unlike *g20/23* above, *g43* is surrounded by less conserved genes and many more ORFs/ORFans, as well as some rearrangements and duplications, which may not represent the origin of the phage core sequences as faithfully as the adjacent genes in the *g23* region. Such results do not invalidate the continued use of the *g43 pol* gene as a phylogenetic marker, but they indicate that the MCP marker remains the preferred choice of the three markers for fine-scale "genotyping" of individual phages or scaffolds.

Analysis of the Largest (>20 kb) T4-Like Scaffolds

Four scaffolds contained more than 20 kb of sequence with homologues of identified Cyano-T4 genes. Two of the scaffolds were primarily P-SSM2/4 like but had similarities to S-PM2/Syn9 phages. Reciprocally, two S-PM2/Syn9-like scaffolds were partly similar to the P-SSM2/4 genomes (supplementary fig. S4, Supplementary Material online). Three of these large scaffolds were located in a segment extending from *g5/g25* to *g7/g8* in the T4 superfamily genome that encodes various components of the tail baseplate structure. Synteny was well maintained in this segment, but one scaffold showed differences due to presumably recent duplications. Regardless of the origin and means of

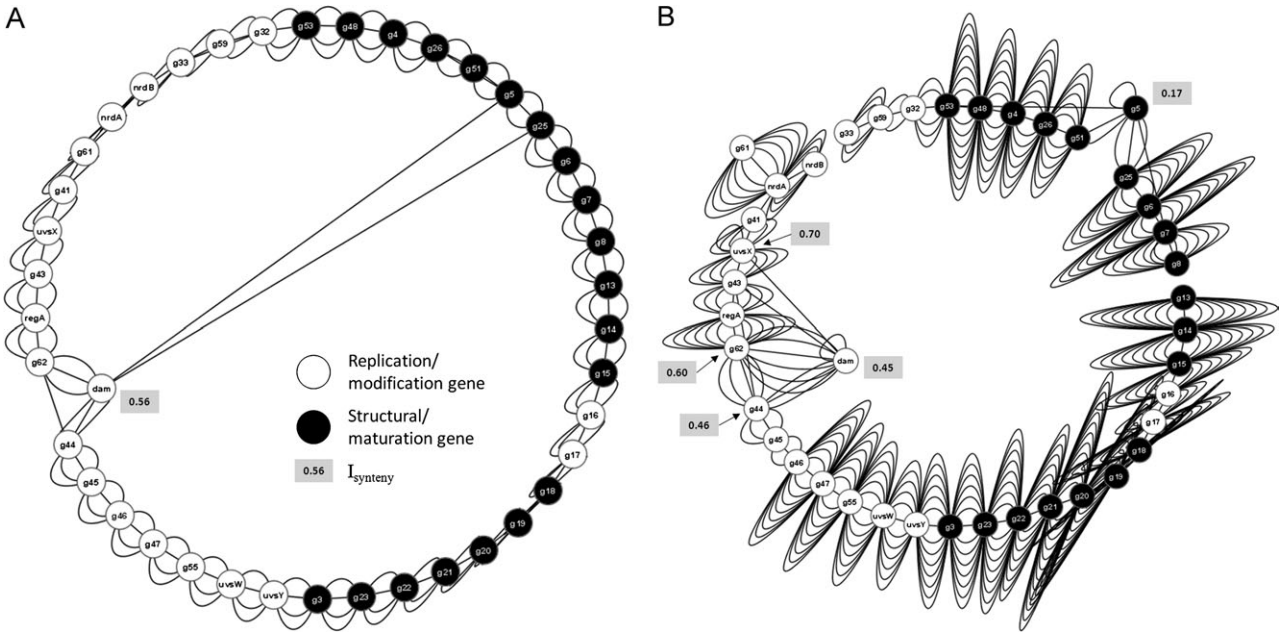

FIG. 4. Cultured Cyano-T4 and GOS scaffold "core genes." The 41 core genes of the cultured Cyano-T4 genomes (*A*) and the scaffolds (*B*) are presented, along with the $I_{synteny}$ values for those genes <0.75. There are 162 distinct occurrences for the Cyano-T4s and 549 occurrences for the scaffolds. Note that the actual number is 609 for the latter due to large mate-pair gaps that break some core gene links ($n = 18$) and because scaffolds containing only a single core gene ($n = 42$) are not included (no links to other core genes).

acquisition of the duplicated genes, they have not significantly altered the overall synteny of their scaffolds. The largest scaffold (~30 kb) had, on its left half, a plastic genomic region that contained many ORFans and, on its right, the segment of the T4 superfamily genome between *g13* and *g18* that encodes the neck and tail sheath components of the virion and the two subunits of the terminase (*g16*/*17*). On the right predominantly S-PM2-like half, there was a cluster of auxiliary genes (plastocyanin, an oxidase and two sugar dehydrogenases) acquired from the host, which are normally at this position between *g16* and *17* in the cyanophages. Therefore, all four of these scaffolds had very good synteny with respect to the characterized genomes of the Cyano-T4 phages in culture.

The Conserved Core in the Cyano-T4 Genomes and the Scaffolds

As the starting point for comparison of the scaffolds, we used the pruning of intervening ORFs (see Materials and Methods) from the 4 cultured Cyano-T4 genomes to depict the remaining 41 core genes, defined as T4-like functions present in at least 3 of 4 genomes (162 total occurrences in P-SSM2/4, S-PM2, and Syn9; supplementary table S3, Supplementary Material online). The resulting Cytoscape presentation (fig. 4*A*) is circular because the phage genomes are circularly permuted and terminally redundant. It is perfectly clear that the vast majority of the core genes in the four genomes are in complete synteny. There is only one genomic region where the synteny breaks down ($I_{synteny} < 0.75$), around the *dam* DNA methylase as a result of the transposition of the *dam* gene in S-PM2 compared with the other three phages.

We next analyzed the synteny of the same 41 cyanophage core genes on the environmental scaffolds (549 total occurrences)—This representation is organized to match as closely as possible the arrangement of that in the cultured genomes (fig. 4*B*). Substantial synteny, with a high number of occurrences, is observed for most of the genes. For example, *g20* had an $I_{synteny}$ of 0.89 and a large number of links ($L = 52$). The representation, however, cannot be closed to a circle because of two gaps in the scaffold data. The first of these occurred in the structural module between *g8* (baseplate) and *g13* (neck); and the second gap was between *g61* (primase)/*nrdB* (ribonucleotide reductase) and *g33* (transcriptional regulator). These gaps were both located in HPRs of 10–80 kb in size in the cultured genomes, and it was to be expected that the small-sized scaffolds (90% < 10 kb) could not connect the genome sequences on either sides of these large hypervariable segments. Nevertheless, two regions of reduced synteny ($I_{synteny} < 0.75$) were evident, one of which is present in the cultured genomes and the other is novel. The first of these was a consequence of the previously mentioned mobility of the *dam* gene. The second interruption in synteny occurred around *g5*, which encodes the lysozyme component of the baseplate. Aside from this modest change, the scaffold core genome was in very good synteny compared with the cultured Cyano-T4s.

Overall Synteny on the Scaffolds

To get a fuller picture of the scaffold synteny, we can include all the genes/ORFs found on scaffolds, as opposed to the "minimal" core genome analyzed above. This analysis can be further extended to also include information on the relationships between the phage and the cellular genes on the scaffolds (fig. 5). If the phage genes were often found next to one another, one would expect multiple lines

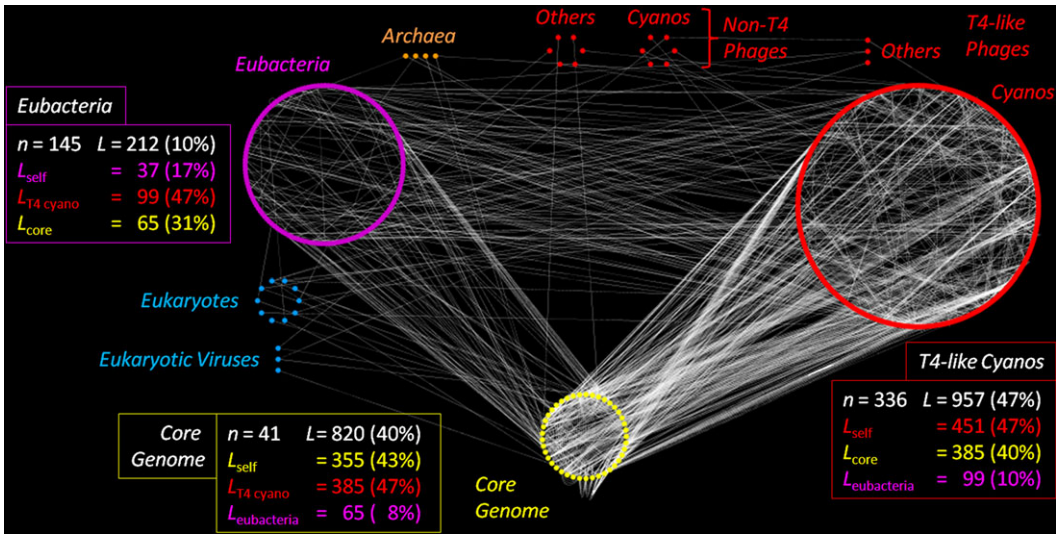

FIG. 5. All GOS scaffold genes/ORFs. All genes/ORFs on the scaffolds are presented, organized by their various origin types, where *n* is the number of genes and *L* the number of links to neighboring genes for each type.

(links) between them, including many lines crisscrossing within the phage circles (core + Cyano-T4) indicating that these genes of the same type (self) were present in contiguous blocks. The visualization confirms this predominance of links between T4-like genes (red and yellow) compared with a more modest portion of links to eubacterial genes (purple). There were few connections to non–T4-like phages, archaeal, eukaryotic, and eukaryotic viral genes (often weak hits) that accounted for only 3% of the total. Considering the 41 genes of the core genome, 43% of their links were to each other (L_{self}), confirming the strong internal cohesion (synteny) of the core but also an equally strong linkage (47%) to noncore Cyano-T4 genes. This indicated that the core genes generally found themselves in a phage-like context that was only rarely interrupted by eubacterial genes ($L_{eubacteria}$ = 8%). The Cyano-T4 genes were similar, with 47% L_{self}, 40% L_{core}, and only 10% linkage to eubacterial genes. The eubacterial genes, however, had a very low degree of synteny amongst themselves (L_{self} = 17%)—compare the few crisscrossing lines inside the purple circle with the web of crisscrossing lines inside the red Cyano-T4 circle. The lack of "self" connections in the eubacterial genes indicated that they were most often inserted singly and not in large contiguous blocks in a phage context. These eubacterial genes sometimes interrupted the synteny of the core genes, but less often (8%) than the other Cyano-T4 genes (10%), further confirming the strong internal cohesion of the core genome components and their resistance to interruption by non–phage-like functions.

Conclusions

This synteny analysis of the T4-like sequences assembled from the GOS metagenome relied on both classical comparative genomic techniques and novel network visualization methods. It appears that our current view of the architecture of Cyano-T4 genomes, although based on only a few cultured examples, is an accurate portrayal of such genomes that are found in marine surface waters. There was remarkably good synteny between the genomes of cultured isolates and the Cyano-T4 core genes on the scaffolds. We were not confident that this would be the case, although the isolated Cyano-T4s infect the dominant members (*Synechococcus* and *Prochlorococcus*) of the marine photosynthetic bacterioplankton (Partensky et al. 1999; Hess 2004). However, 16S ribosomal DNA surveys of the environment, for example, have shown that cultured isolates often do not accurately reflect ecosystem diversity/composition (Handelsman 2004). Similarly, we sought to directly demonstrate via metagenomic analysis whether the most prevalent phage genomes in the environment (including those whose hosts have not yet been cultured) were similar to the few cyanophages in culture.

We believe our metagenomic analysis gives "environmental legitimacy" to the Cyano-T4 model genomes available and thus justifies further pursuit of interesting questions concerning their environmental impact and genome diversity. For example, what are the causes of the limited number of synteny breakdowns that we observe or why are particular conserved ORF(an)s strongly associated with certain known genes, and most importantly what is the functional logic of the genome's organization. The answers to such questions may allow us to deduce evolutionary and functional relationships between the diverse genes in the Cyano-T4 genomes. This work also points out the necessity of focusing further work on particular members of the marine T4 superfamily, such as the unstudied *Synechococcus* cyanophage S-BnM1, which the metagenomic analysis indicated was more widespread and numerous than previously imagined. Similarly, much deeper metagenomic sequencing of certain GOS samples should now be undertaken on the basis of the distribution and synteny of scaffolds already available to get complete genomic coverage of these interesting Cyano-T4 variants. Equally important were those rare samples that contain

non-cyanophage T4-like scaffolds whose further analysis could expand our repertoire of these phage genomes. Regardless of the precise focus of future metagenomic research, we can also look forward to further analyses of nonmarine environments, such as freshwater and soil. Such studies could also simultaneously validate the environmental legitimacy of other phage superfamilies unrelated to T4.

Finally, we would like to draw attention to the fact that the major sources of "alien" genes in the Cyano-T4 scaffolds were their host cyanobacteria (and other common marine groups), not other phages. This result is not so surprising because earlier analysis of cultured phage genomes (Mann et al. 2005; Sullivan et al. 2005; Zeidner et al. 2005; Sharon et al. 2009) had already indicated that important host genes, such those involved in photosynthesis, seem to have been exchanged during the course of the host–parasite relationship. Perhaps even more important are the commonality of such exchanges and the breadth of genes involved. We would like to reiterate our previous suggestion that there is an important, but underappreciated, mutualism involved in phage–host evolutionary interactions (Comeau and Krisch 2005; Comeau, Tétart, et al. 2007). Phage can evolve more rapidly and create gene diversity at much higher levels than their hosts. It is an advantage to both of the "partners" if the phage's host competes effectively for limited resources in its environment. Thus, if some of the phage's evolutionary innovations can be transferred to, and co-opted by, its host and thereby increase its competitive ability, its phage benefits as well. The traces of these phage contributions to the evolution of host genomes are presumably reflected in the numerous genes shown in figure 5 that are shared between them.

Supplementary Material

Supplementary figures S1–S4, tables S1–S3, and file S1 (corrected scaffolds FASTA) are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank our colleague Roland Barriot for discussions concerning the mathematical representation of the Index of Synteny. This work was supported intramurally by the Centre National de la Recherche Scientifique. A.M.C. was supported by the Les Treilles Foundation scientific prize and H.M.K. received supplementary support from the Kribu Foundation.

References

Ackermann HW. 2007. 5500 Phages examined in the electron microscope. *Arch Virol*. 152:227–243.

Angly FE, Felts B, Breitbart M, et al. (18 co-authors). 2006. The marine viromes of four oceanic regions. *PLoS Biol*. 4:2121–2131.

Besemer J, Borodovsky M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. 33:W451–W454.

Breitbart M, Salamon P, Andresen B, Mahaffey JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA*. 99:14250–14255.

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

Comeau AM, Bertrand C, Letarov A, Tétart F, Krisch HM. 2007. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* 362:384–396.

Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH, Prangishvili D. 2008. Exploring the prokaryotic virosphere. *Res Microbiol*. 159:306–313.

Comeau AM, Krisch HM. 2005. War is peace—dispatches from the bacterial and phage killing fields. *Curr Opin Microbiol*. 8:488–494.

Comeau AM, Krisch HM. 2008. The capsid of the T4 phage superfamily: the evolution, diversity and structure of some of the most prevalent proteins in the biosphere. *Mol Biol Evol*. 25:1321–1332.

Comeau AM, Tétart F, Trojet SN, Prère M-F, Krisch HM. 2007. Phage-Antibiotic Synergy (PAS): β-lactam and quinolone antibiotics stimulate virulent phage growth. *PLoS One*. 2:e799.

Dorigo U, Jacquet S, Humbert JF. 2004. Cyanophage diversity, inferred from *g20* gene analyses, in the largest natural lake in France, Lake Bourget. *Appl Environ Microbiol*. 70:1017–1022.

Filée J, Bapteste E, Susko E, Krisch HM. 2006. A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol*. 23:1688–1696.

Filée J, Forterre P, Sen-Lin T, Laurent J. 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol*. 54:763–773.

Filée J, Tétart F, Suttle CA, Krisch HM. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci USA*. 102:12471–12476.

Frank H, Moebus K. 1987. An electron microscopic study of bacteriophages from marine waters. *Helgolander Meeresunters*. 41:385–414.

Fuller NJ, Wilson WH, Joint IR, Mann NH. 1998. Occurrence of a sequence in marine cyanophages similar to that of T4 *g20* and its application to PCR-based detection and quantification techniques. *Appl Environ Microbiol*. 64:2051–2060.

Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 68:669–685.

Hardies SC, Comeau AM, Serwer P, Suttle CA. 2003. The complete sequence of marine bacteriophage VpV262 infecting *Vibrio parahaemolyticus* indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment. *Virology* 310:359–371.

Hatfull G, Jacobs-Sera D, Lawrence JG, et al. (27 co-authors). 2010. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J Mol Biol*. 397:119–143.

Hess WR. 2004. Genome analysis of marine photosynthetic microbes and their global role. *Curr Opin Biotechnol*. 15:191–198.

Howe K, Bateman A, Durbin R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18:1546–1547.

Jia ZJ, Ishihara R, Nakajima Y, Asakawa S, Kimura M. 2007. Molecular characterization of T4-type bacteriophages in a rice field. *Environ Microbiol*. 9:1091–1096.

Karam JD, Konigsberg WH. 2000. DNA polymerase of the T4-related bacteriophages. *Prog Nucleic Acid Res Mol Biol*. 64:65–96.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. 2008. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev*. 72:557–578.

Leiman PG, Kanamaru S, Mesyanzhinov VV, Arisaka F, Rossmann MG. 2003. Structure and morphogenesis of bacteriophage T4. *Cell Mol Life Sci*. 60:2356–2370.

Mann NH, Clokie MRJ, Millard A, Cook A, Wilson WH, Wheatley PJ, Letarov A, Krisch HM. 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. Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* 424:741.

Miller ES, Heidelberg JF, Eisen JA, et al. (13 co-authors). 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol*. 185:5220–5233.

Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W. 2003. Bacteriophage T4 genome. *Microbiol Mol Biol Rev*. 67:86–156.

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

Rohwer F, Segall A, Steward G, Seguritan V, Breitbart M, Wolven F, Azam F. 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol Oceanogr*. 45:408–418.

Rusch DB, Halpern AL, Sutton G, et al. (40 co-authors). 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*. 5:398–431.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13:2498–2504.

Sharon I, Alperovitch A, Rohwer F, et al. (12 co-authors). 2009. Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461:258–262.

Short CM, Suttle CA. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol*. 71:480–486.

Stothard P, Wishart DS. 2005. Circular genome visualization and exploration using CGView. *Bioinformatics* 21:537–539.

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*. 3:790–806.

Suttle CA. 2007. Marine viruses—major players in the global ecosystem. *Nature Rev Microbiol*. 5:801–812.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22:4673–4680.

Weigele PR, Pope WH, Pedulla ML, Houtz JM, Smith AL, Conway JF, King J, Hatfull GF, Lawrence JG, Hendrix RW. 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol*. 9: 1675–1695.

Williamson SJ, Rusch DB, Yooseph S, et al. (12 co-authors). 2008. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*. 3:e1456.

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

Zhong Y, Chen F, Wilhelm SW, Poorvin L, Hodson RE. 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene *g20*. *Appl Environ Microbiol*. 68: 1576–1584.