

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

Jonathan Filée,* Eric Baptiste,† Edward Susko,‡ and H. M. Krisch*

*Laboratoire de Microbiologie et Génétique Moléculaire, CNRS UMR-5100, Toulouse, France; †Canadian Institute for Advanced Research and Genome Atlantic, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada; and ‡Genome Atlantic, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

Genomic analysis of bacteriophages frequently reveals a mosaic structure made up from modules that come from disparate sources. This fact has led to the general acceptance of the notion that rampant and promiscuous lateral gene transfer (LGT) plays a critical role in phage evolution. However, recent sequencing of a series of the T4-type phages has revealed that these large and complex genomes all share 2 substantial syntenous blocks of genes encoding the replication and virion structural genes. To analyze the pattern of inheritance of this core T4 genome, we compared the complete genome sequences of 16 T4-type phages. We identified a set of 24 genes present in all these T4-type genomes. Somewhat surprisingly, only one of these genes, that encodes for ribonucleotide reductase (*NrdA*), displayed evidence of LGT with the bacterial host. We test the congruence of the inheritance of the other 23 markers using heat map analyses and comparison of a reference topology with the 23 individual gene phylogenies. The vast majority of these core genes share a common evolutionary history. In contrast, analyses of all the noncore genes present in the same 16 genomes, located in the hyperplastic regions of the genome, show considerable evidence of frequent LGT. The similar evolution of the core replication and virion structural genes in the T4-type phage genomes suggests that, unlike the situation in many other phage groups, such portions of T4-type genome have been inherited as a block, without significant LGT, from a distant common ancestor. The preservation of the synteny of the core T4 genome could result from several factors acting in synergy, such as the constraints imposed by the sophisticated regulation of the transcription. Moreover, numerous and complex protein–protein interactions during virion morphogenesis could also impose a supplementary barrier against LGT. Finally, there may be some real evolutionary advantage to maintaining large regions of conserved sequence. Such segments could be a sort of genetic glue that maintains the genetic cohesion of the T4-type phages via recombination within the most conserved sequences. This could mediate the swapping of nonconserved sequences that they flank.

Introduction

The number of phages in the biosphere has been estimated to exceed the number of bacteria by a factor of 5–25, making phages the most abundant biological entities (Fuhrman 1999; Balter 2000). Many lines of evidence indicate that phage diversity is enormous (Rohwer 2003) and that their genomes represent the largest resource of genetic information on the planet. But attempts to assess viral genomic diversity are complicated by the fact that phage genomes evolved as a patchwork of DNA from disparate origins (Hendrix et al. 1999). Many studies have documented the impressive extent of mosaicism in the temperate *Escherichia coli* phages (e.g., λ and Mu-like phages, see for review Hendrix 2003), dairy phages (Proux et al. 2002), or mycobacteriophages (Pedulla et al. 2003). Lateral gene transfer (LGT) is a major mediator of phage evolution and has occurred frequently between diverse phages that share overlapping host ranges. For example, in the λ family, the effect of LGT is so large that it is difficult to define the characteristic features of lambdoids. Long ago, Hershey and Dove (1971) remarked that the only unifying characteristics among different λ family members was “the ability to form intervarietal hybrids” and a common genome organization that “is nearly, but not quite independent of specific nucleotide sequences.” More to the point, Hendrix et al. (1999) observed that the high frequency of LGT between

bacteriophage families can, in some cases, blur phylogenetic boundaries and limit the utility of classification by species. LGT also occurs between phages and their host at such a level that phages must now be considered as a potential major driving force of cellular evolution, serving as the primary vector for horizontal genetic exchange in the environment (Waldor and Mekalanos 1996). Phage genes have been frequently subverted to provide cellular functions (Forterre 1999; Filée et al. 2002, 2003, 2005), and the reverse is also true: host genes have been acquired by phages and subverted for their own purposes (Lindell et al. 2004; Millard et al. 2004).

In order to assess the impact of LGT in phage evolution, comparisons have been made on the genome sequences of a large collection of T4-type phages belonging to the 3 subgroups of T4-type phages (“T-evens and PseudoT-evens,” “SchizoT-evens,” and “ExoT-evens”). We have recently shown that there is substantial diversity among the T4-type bacteriophages that coinhabit various niches of marine ecosystem (Filée et al. 2005). This propinquity could result in substantial levels of recombination and LGT between divergent T4-type bacteriophages with overlapping host range. Genomic sequencing demonstrates that all the T4-type phages share a common ancestor for a significant fraction of their genomes [A Comeau, personal communication; Nolan et al. (forthcoming)]. Conserved genes form syntenic assemblages that represent 14–21% of the total length of the genome. Hypervariable segments of the genomes encode proteins that often seem to be concerned with adaptations of the phages to a particular host or environmental niche (Desplats et al. 2002; Miller et al. 2003; Mann et al. 2005; Sullivan et al. 2005). In contrast to the rampant genetic plasticity observed in such hypervariable

Key words: T4 phages, phylogeny, lateral gene transfer, synteny, congruence.

E-mail: jonathan.filee@ibcg.biotoul.fr.

Mol. Biol. Evol. 23(9):1688–1696. 2006

doi:10.1093/molbev/msl036

Advance Access publication June 16, 2006

Table 1
Summary of the T4 Genomes Analyzed in This Study: Genome Size, Bacterial Host, and References

Phage	Genome Size (kb)	Host	References
T and PseudoT			
T4	169	<i>Escherichia coli</i>	Miller et al. (2003)
RB69	167.6	<i>E. coli</i>	http://phage.bioc.tulane.edu/
RB49	164	<i>E. coli</i>	http://phage.bioc.tulane.edu/
RB43	180.5	<i>E. coli</i>	http://phage.bioc.tulane.edu/
RB16	177	<i>E. coli</i>	http://phage.bioc.tulane.edu/
44RR	173.6	<i>Aeromonas salmonicida</i>	http://phage.bioc.tulane.edu/
25	162	<i>A. salmonicida</i>	http://phage.bioc.tulane.edu/
31	173	<i>A. salmonicida</i>	http://phage.bioc.tulane.edu/
133	170	<i>Acinetobacter johnsonii</i>	http://phage.bioc.tulane.edu/
ShizoT			
Aeh1	233.2	<i>Aeromonas hydrophila</i>	http://phage.bioc.tulane.edu/
65	235	<i>A. salmonicida</i>	http://phage.bioc.tulane.edu/
KVP40	244.8	<i>Vibrio parahaemolyticus</i>	Miller et al. (2003)
nt1	247.1	<i>Vibrio natriegenes</i>	http://phage.bioc.tulane.edu/
ExoT			
SPM2	196.2	<i>Synechococcus sp.</i>	Mann et al. (2005)
PSSM4	178.2	<i>Prochlorococcus sp.</i>	Sullivan et al. (2005)
PSSM2	52.4	<i>Prochlorococcus sp.</i>	Sullivan et al. (2005)

regions, genetic swapping within conserved syntenous regions appears to be much rarer. Different regions of the T4-type genome thus seem to have dissimilar susceptibilities to LGT. Here we attempt to understand the evolutionary significance of the different behavior of these segments of the T4 genome.

Materials and Methods

Sequence Retrieval, Data Set Construction, and Phylogenetic Analyses

All the open reading frames (ORFs) of T4 (GenBank accession number AF158101) were blasted using BlastP (Altschul et al. 1990) against a database containing the sequences of 15 complete genomes of T4-type bacteriophages (table 1) at <http://phage.bioc.tulane.edu/blast/blast.html>. Depending on the Blast cutoff, we identified a set of approximately 30–35 genes shared by all these genomes (see table 1). For subsequent analyses, we retained only genes with *E* values lower than 10^{-5} and that were more than 100 amino acids long. This procedure led to a data set of 24 conserved genes: *gp4* (head completion), *gp6* (baseplate wedge), *gp13* (head), *gp14* (head), *gp15* (head), *gp17* (head), *gp18* (sheath), *gp20* (head portal), *gp21* (head protease), *gp22* (head scaffold), *gp23* (major head capsid), *gp25* (baseplate plug), *gp32* (helix-destabilizing protein), *gp41* (DNA primase/helicase), *gp43* (DNA polymerase), *gp44* (DNA polymerase accessory protein), *gp46* (DNase), *gp47* (DNase), *gp53* (baseplate wedge), *gp55* (RNA polymerase sigma factor), *gp61* (helicase primase subunit), *regA* (translational regulator), *UvsW* (unknown function), and *NrdA* (ribonucleotide reductase). To identify possible LGT with organisms other than T4-type phages, each of these conserved genes was blasted against the Non Redundant (NR) database at <http://www.ncbi.nih.gov/BLAST/>. This allowed us to identify a possible example of LGT for *NrdA*. This gene was not considered in our further analyses of congruence. All these alignments were inspected and manually refined. Gaps and ambiguously aligned posi-

tions were eliminated for the phylogenetic analysis. For all individual markers, we performed preliminary analyses to assess the gene orthology by Neighbor-Joining (NJ) (without any correction) using MUST 3.0 (Philippe 1993) and by maximum likelihood (ML) using PROML (PHYLIP v3.6) with the Jones Taylor and Thornton (JTT) amino acid substitution matrix, a rate heterogeneity model with gamma-distributed rates over 4 categories with the α parameter estimated using Tree-Puzzle, global rearrangements, and randomized input order of sequences (10 jumbles). A concatenation of all these markers was then realized (length = 4,653 amino acids, 16 species), and the best ML tree was calculated for it with the same settings. All alignments and corresponding phylogenetic trees are available on request from J. Filée.

Tree Files

We employed a set of 946 unrooted topologies consisting of the concatenation tree plus all the 945 possible rearrangements of 7 paired taxa identified by Filee et al. (2005): the marine cyanophages (PSSM2, SPM2, PSSM4), the marine vibriophages (nt1, KVP40), a group of coliphages (RB43, RB16), freshwater phages (44RR, 31, 25), (Aeh1, 65), RB49, plus an additional group of enterobacterial phages (T4, RB69, 133), suggested by our preliminary phylogenetic analyses of concatenated data. These entrees were used as user tree in Tree-Puzzle 5.1, option-wsl, with a JTT + Γ 8 + I model of evolution to estimate the likelihood of each site of a given gene and global tree likelihoods for each tree. These likelihood values were used as input for CONSEL (Shimodaira and Hasegawa 2001) to perform the approximately unbiased (AU) test (Shimodaira 2002). This is a statistical test of the hypothesis that the given topology is the correct topology for the taxa under consideration. When the *P* value associated by the AU test to one of the topologies under study is <0.05 , this tree can be said to be significantly different and worse than other topologies, at a threshold of 5%.

Heat Map Analyses and Gene Pairwise Comparisons

Heat maps of P values of the AU test (Baptiste et al. 2005; Susko et al. 2006) were used to more thoroughly test which genes support similar test topologies (P values >0.05) and which genes had a different phylogenetic history. Such heat maps display the information about the support/rejection for each topology by each gene by a colored matrix, where each unit of the x axis corresponds to a given gene and each unit of the y axis corresponds to a given test topology. More precisely, light-colored cells indicate supported trees, whereas dark-colored cells identify rejected trees. We also implemented a new alternative conservative method to detect incongruencies. These heat maps were constructed with different P values, the pairwise P values, for tests that determine if pairs of genes come from the same topology. These pairwise P values were calculated based on the fact that existing methodology tests the appropriateness of a given topology for a gene. A $(1 - \alpha) \times 100\%$ confidence region of topologies for a given gene can be construed as the set of supported topologies (i.e., the topologies with P values greater than α) for that gene. Given confidence regions for 2 genes, the P value for the pairwise test for which they share a common supported topology was taken as twice the smallest α such that at least one topology is contained in both confidence regions for the genes. Because the smallest α could be greater than 0.5, doubling it can give a value greater than 1; in such a case we set it to 1. As with all P values, the pairwise P value can be defined as the smallest α level at which a corresponding test of the equality of topologies for the 2 genes can be rejected. This test rejects the hypothesis of equality at the α level if $(1 - \alpha/2) \times 100\%$ confidence regions for the genes have no topologies in common. Supplementary Material 1 (Supplementary Material online) establishes that these P values are conservative. A heat map with 2-way clustering of these P values for pairwise tests aids the visualization of this possible compatibility between genes histories. More precisely, clusters for columns (genes) are created in a hierarchical manner by successively joining together the pair of genes (or clusters of genes) whose pairwise P values tend to be most similar in the sense of having the smallest sum of squared differences. The process is repeated for rows (genes as well), giving rearrangements of rows and columns of the matrix of P values that will more clearly illustrate patterns of compatibility. The cells of these heat maps give the pairwise P value for all the possible pairs of genes so that the rows and columns are by definition symmetric. A white-colored or light-colored cell indicates that there is evidence that the 2 genes could share a common phylogenetic history. A dark-green-colored cell indicates pairs of genes that do not support the same set of topologies. Consequently, a dark-green-colored row/column indicates the disagreement of one gene with most other genes.

All heat maps were generated using the freely available statistical package R (<http://www.r-project.org/>). R language functions and example scripts for generating them will be made available at <http://www.mathstat.dal.ca/~tsusko>.

Synthesis Reconstruction

The synthesis of T4 marine phages was inferred from the analyses of 23 ML gene trees and of their concatenated tree. Individual ML trees as well as the concatenation were calculated using PROML. Options were global rearrangements, randomized input order of sequences (10 jumbles), JTT amino acid substitution matrix, and a rate heterogeneity model with gamma-distributed rates over 4 categories, with the α parameter estimated using Tree-Puzzle. Bootstrap supports represent a consensus (obtained using CONSENSE) of 100 Fitch–Margoliash distance trees (obtained using PUZZLEBOOT and FITCH) from pseudoreplicates (obtained using SEQBOOT) of the original alignment. The settings of PUZZLEBOOT were the same as those used for PROML, except that global rearrangements and randomized input order of sequences are not available in this program. The clades supported with more than 50% bootstrap support in these 23 gene trees were compared with the concatenation tree using 2 programs: Horizstory and Lumbermill (MacLeod et al. 2005). Briefly, Horizstory allows the inference of the most parsimonious scenarios involving LGT and vertical descent to explain the common features and the discrepancies between the “organismal” tree and each of the 23 gene trees. Lumbermill draws the synthesis by mapping the outcomes of these scenarios onto the reference tree and calculates all the estimates presented here. A strict consensus option was applied, meaning that only the relationships supported or inferred in 100% of the evolutionary scenarios resulting from the comparison between the reference and a given tree were considered in this drawing.

Genome Dot Plot

Intergenome comparisons were performed by pairwise BlastP alignments without filtering and with an exclusion threshold of $E < 10^{-5}$. Each ORF of a genome were blasted against all the ORFs of the other genome, and the sequence having the best high-scoring segment pair were used to plot the alignment diagram.

Close Blast Hit Analyses

For each ORF of a genome, a BlastP search was performed at <http://www.ncbi.nih.gov/BLAST/> on an NR database. Distribution of close Blast hits were manually established with E value threshold cutoffs better than 10^{-5} . For genes susceptible to have been recently acquired by a phage from a cellular source, we confirmed the transfer by reconstructing individual NJ phylogenies.

Results and Discussion

DNA sequencing of the T4-type phages revealed substantial variations in genome size and genome content (table 1), and even among the most conserved core genes, there is considerable sequence divergence (Desplats et al. 2002; Miller et al. 2003; Mann et al. 2005; Sullivan et al. 2005). In spite of all their divergence, dot plot comparisons of all these genomes revealed the existence of 2 large syntenous genomic segments (fig. 1A). The smaller one encodes a group of early genes involved in DNA

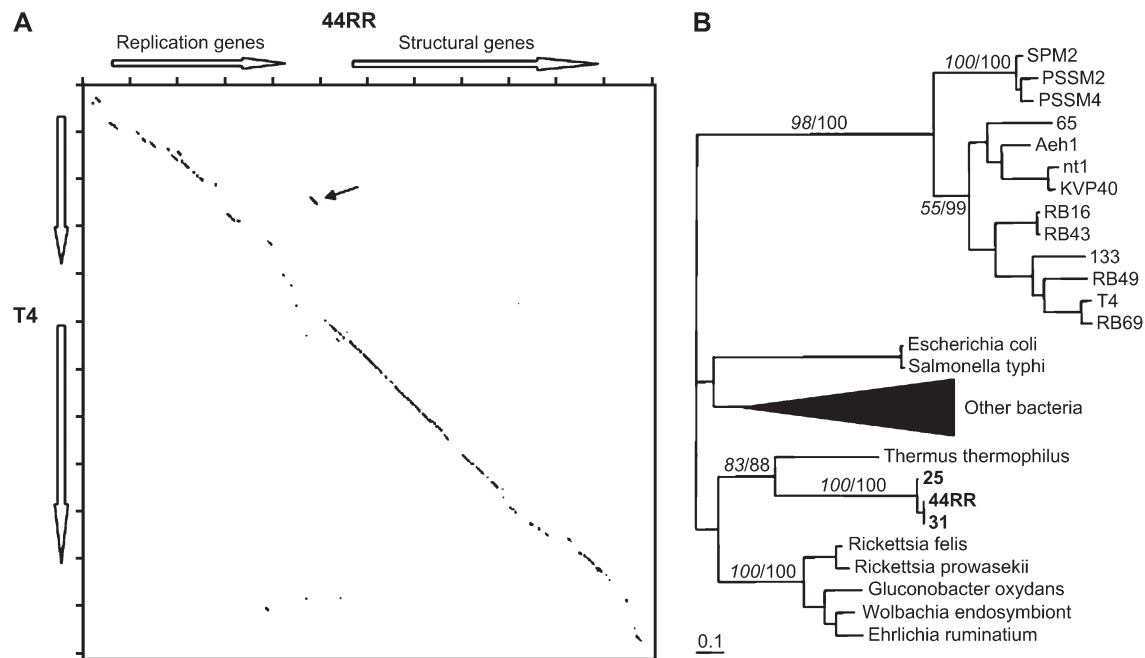


FIG. 1.—(A) Genome dot plot of phage T4 (vertical axis) versus phage 44RR (horizontal axis). Large arrows indicate the respective position of the 2 conserved blocks of genes in each genome. The small black arrow indicates the position the ribonucleotide reductase *NrdA*. (B) Maximum likelihood phylogeny of the ribonucleotide reductase *NrdA*. Bootstrap values for ML (italic) and Neighbor-Joining tree are shown at the nodes. Scale bar represents the number of amino acid substitutions per residue. Phages susceptible to be concerned by a LGT are indicated in bold.

replication and transcription. The second, larger, segment contains the late genes that encode the structural components of the virion. The presence of such large genomic regions that, although highly diverged, still retain considerable synteny is intriguing. It immediately poses the question of why this synteny has been preserved during evolution and what important function it might provide to the phage. In an attempt to answer this question, we first identified 24 T4-type genes of sufficient size and sequence conservation to be employed in a classical phylogenetic analysis. For the DNA replication genes, such as the polymerase (gp43), DNA primase (gp61), and clamp loader (gp62), no evidence for any LGT with cellular organism was obvious. Among the early genes, only the *NrdA* gene involved in nucleotide biosynthesis was clearly an alien in a subset of these genomes. Blast searches revealed that the *NrdA* gene in the *Aeromonas* phages 44RR, 25, and 31 were more closely related to bacterial sequences than to the other T4 phage sequences. The phylogeny of this gene shows with high bootstrap values that the common ancestor of phages 44RR, 25, and 31 must have acquired a novel version of the *NrdA* gene (fig. 1B). Significantly, dot plot comparisons of the genomes of phages 44RR, 25, 31, and T4 indicate that the *NrdA* gene is not in its habitual location among the replication genes but is transposed to a distant site (fig. 1A).

To increase our ability to detect LGT in the other conserved genes of the T4-type phages, we employed more sophisticated statistical methods. We calculated the *P* values obtained by the AU test for each of the 23 orthologous viral genes (*NrdA* was excluded) and a series of test topologies (see Materials and Methods) giving 946 *P* values for each gene. With this data, we performed 2 kinds of heat map analyses.

First, we tested for potential conflict between markers by evaluating the congruence of all possible pairs of genes of this data set (see Materials and Methods). The results of these pairwise tests are depicted in figure 2A. Light-colored and white-colored cells indicate pairs of genes for which there is no evidence to reject compatibility. Dark-green-colored cells pinpoint potentially incongruent pairs of genes. As this heat map is almost exclusively white, it shows no apparent conflict between most pairs of genes. Only *gp13* (a protein in the neck of the phage virion) behaves anomalously compared with most of the other phage genes, causing a dark-green-colored row/column for the comparisons in which this marker is involved (row/column no. 1). Nevertheless, *gp13* history remains still compatible with some of the other gene histories. This pattern occurs because, unlike the other genes, its phylogeny has no bootstrap basal support (see Supplementary Material 2 online). Such an atypical absence of bootstrap support for many basal nodes affects the *P* values inferred using this marker and makes it appear at odds with the rest of the data set. Based on this first conservative test, however, we cannot affirm that all genes share the same history. Even if they do not disagree two by two, it does not imply that all pairs agree on a similar unique tree.

We thus used a second approach based on the simultaneous global comparison of all the genes' *P* values for all the topologies. On the heat map of figure 2B, the dark-colored cells indicate low *P* values—that is rejection—for a test topology of a given gene. By contrast, a light-colored cell indicates high *P* values, that is, a good support for this topology by a given gene. Because the heat map is predominantly dark colored, such a display makes it clear that most genes (along the *x* axis) reject almost all the test topologies

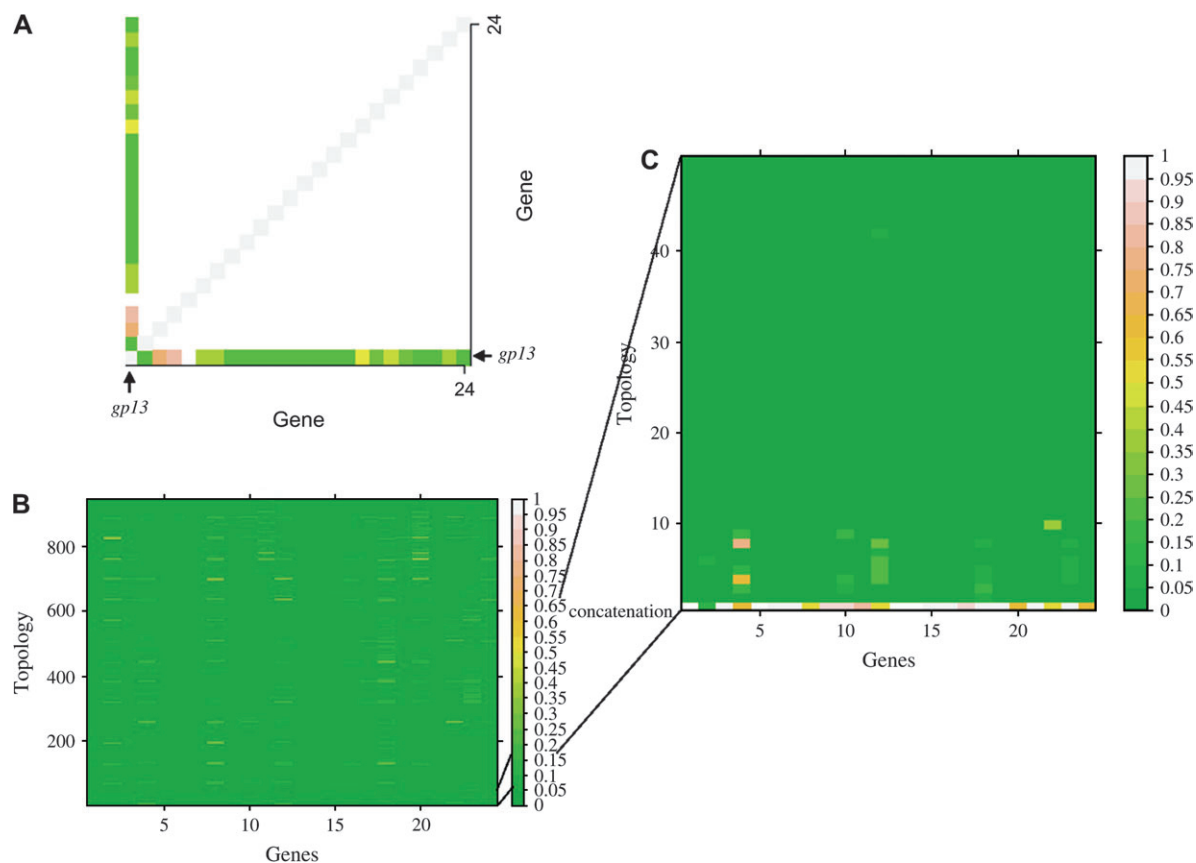


FIG. 2.—(A) Heat maps of the pairwise P values for the pairwise tests that genes share a common supported topology. Each square gives the pairwise P value for a test that gene in the corresponding row share a supported topology in common with the gene in the corresponding column. Lighter colors indicate no evidence against compatibility, whereas darker colors indicate that the 2 genes do not have compatible histories (see main text). The respective positions of the gene *gp13* are indicated with arrows. (B) Heat map for the P values for each of the 24 viral markers (23 genes plus the concatenation, on the x axis) assessed against 946 test topologies, without clustering of either genes or topologies. Lighter colors indicate a higher P value of the data given the tree (i.e., stronger support), whereas darker colors indicate lower P values (stronger rejection). The position of gene *gp13* is indicated with arrows. (C) Heat map when the P values of the 24 viral markers (23 genes plus the concatenation) for the 50 first topologies of figure 2B. It is obvious that one topology, the concatenation tree (indicated with an arrow), was supported by a majority of the data set. The color code associated with these P values (from green for rejection to white for support) is the same as reported in figure 2B.

(along the y axis). There is thus a strong phylogenetic signal in this data set. This analysis reveals that some genes support alternative topologies but that such topologies are rejected by the vast majority of the other markers. There is thus no evidence for deep conflicting signal within a given marker as recombination events between distantly related phages would have produced a different pattern in our heat map. In this case, a marker would then have difficulty to not reject multiple test topologies, and each column of the heat map would be a mosaic of light-colored spots. Interestingly, only one topology that is at the bottom of the figure is not rejected by almost all the genes, creating an almost continuous light-colored and white-colored line as is shown by an enlargement of figure 2C, which focuses on the first 50 rows (topologies) of the heat map in figure 2B.

The pattern displayed is most simply and best explained by a common dominant history, a unique tree being strongly supported by the large majority of the markers. Indeed, there is only one such plausible topology that is supported by a majority of genes. It is the concatenation tree. This observation is consistent with the pairwise analyses and fits well with our expectations. Indeed, if there is

a unique tree for these phage species and if we have sampled it in our large data set of topologies, this phage tree would be one of the most commonly supported topologies of the data set. Then, if the history of none of the individual genes disagrees with the most plausible topologies, we can reasonably conclude that they are likely to share a unique common history. As a matter of fact, only one marker (*gp13*) was apparently incongruent with the rest of the data set. All the other genes were likely congruent and would support a tree very close, if not identical, to the concatenation tree. However, with these analyses, we cannot rigorously exclude the possibility that some species acquired a closely related copy by occasional lateral transfer because such testing is inherently conservative in protecting against the Type I errors of rejecting vertical descent.

Consequently, we used an approach that goes further than the heat map when applied to test topologies. We realized a synthesis, which considers an alternative set of topologies, the best individual ML phylogeny of each gene, for which no relationship was a priori fixed and contrasts these trees with the best reference issued from the heat map analysis, that is, the concatenation tree (MacLeod

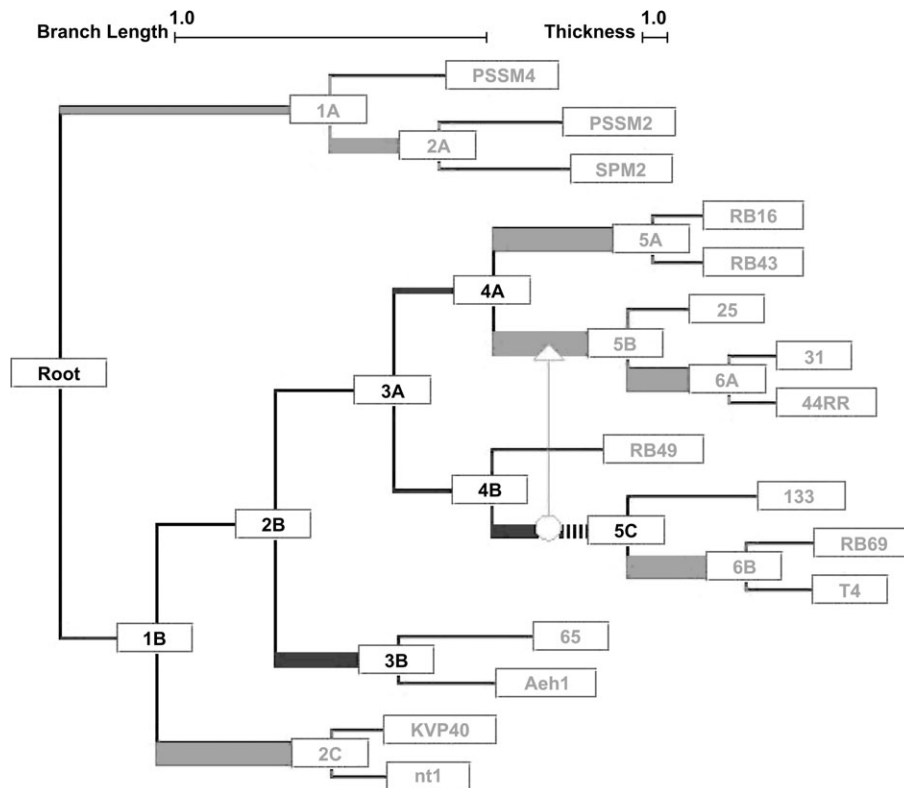


FIG. 3.—Synthesis of 23 T4-like phage genes. The proposed vertical-inheritance backbone—the concatenation tree—is shown in black, with the line thickness of an internal branch corresponding to the frequency of its support across the whole data set. Support was considered significant when clades received >50% bootstrap support. Putative LGT events concerned genes *gp25* and *gp46* and are reported in gray, connecting donors (circles, here 5C) with recipients (arrowheads, here 5B).

et al. 2005). The vertical dimension aims at pinpointing the portion of viral genome that is free from LGT, whereas horizontal exchanges (i.e., recent ones between members of the groups not tested before) would be placed on the horizontal axis. The resulting graph (fig. 3) was mostly treelike and only partly weblike, and 30 vertical branches are visible as well as one lateral connection. The comparison of the total support for the horizontal and vertical branches indicated that the vertical signal is about 23 times more important than the horizontal signal in these markers. A simple interpretation is that all these genes manifest support for vertical inheritance and could define robust clades. It would indeed be against standard practice in phylogeny to not associate a dominant congruent signal with vertical inheritance. However, if this congruence between all these markers were to be explained by recombination events, it would require recombination events of a region covering all of them. Such a region of recombination would have to extend over more than 100 kb because the 2 blocks (20 kb for the replication genes and 60 kb for the structural genes) are separated by a 20-kb spacer in the phage genomes. We feel that a vertical transmission is the more probable explanation for the congruent inheritance of the markers in these strictly lytic phages, which have fewer opportunities to recombine compared with prophage integrated in the host genome. Nevertheless, the synthesis also suggests that 2 genes, *gp25* (baseplate wedge subunit) and *gp46* (recombination exonuclease), were occasionally laterally trans-

ferred, from the last common ancestor of 133, RB69, and T4 to the last common ancestor of 25, 44RR, and 31. We suggest that these potential transfers that one can map onto the reference tree could be used as a synapomorphy for this last group. Phages 25, 44RR, and 31 infect *Aeromonas* species, which are abundant in aquatic habitats. As phages such as T4, RB69, and 133 are also widespread in marine environments (Filee et al. 2005), the existence of such transfers between these cohabiting phages is ecologically consistent. Interestingly, this transfer conserved the gene order.

In parallel, we investigated the level of LGT of the other genes of the T4-type genomes. Excluding the 24 most conserved genes, we did a survey of all the nonubiquitous ORFs in 4 randomly chosen representatives of the T4-like phages: 44RR (228 ORFs), Aeh1 (328 ORFs), 133 (204 ORFs), and SPM2 (215 ORFs). Each ORF was blasted against a nonredundant database using an *E* value cutoff of 10^{-5} , and potential LGTs were subsequently analyzed with individual phylogenies to define the source of the transfer. The frequency of the taxonomical categories of the first hit is reported in figure 4. Strikingly, a significant fraction of the first BlastP hit comes from cellular sources: 9.3% for SPM2, 2.6% for 44RR, 3% for 133, and 3.4% for Aeh1. A similar level of bacterial originated genes was also reported for KVP40 (5.8%) (Miller et al. 2003). These results indicated that the recruitment of genes of cellular origin is not unusual during the evolution of T4-type phages. Many of these events occur in genes encoding the

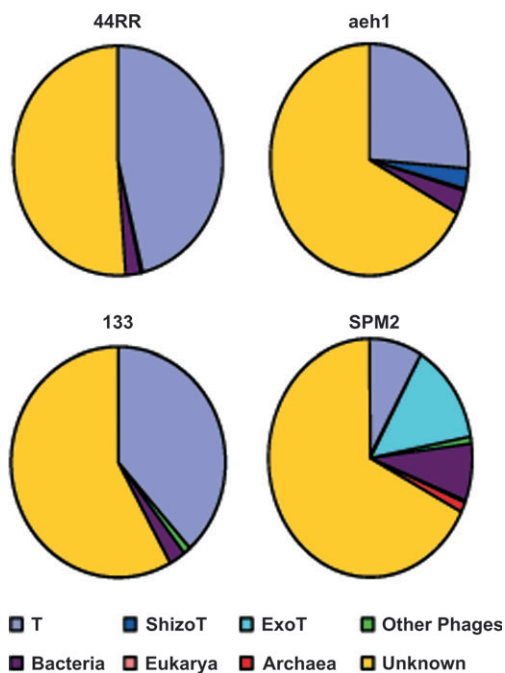


FIG. 4.—Taxonomy of best BlastP hits for 44RR, Aeh1, 133, and SPM2. Each ORF from the phage genomes was used as a query against an NR database to identify the taxon of the best hit (see Materials and Methods). Yellow slice indicates hypothetical proteins, blue slice indicates T4-hit, green slice other phage-hit, red slice archeal hits, pink slice eukaryotic hit, and purple slice bacterial hits. “T” indicated that the first hit is a T4-type phage belonging to the T-even or the PseudoT-even subgroup.

enzymes involved in DNA metabolism or DNA modification. In this regard, the evolutionary history of the gene encoding the thymidylate synthase is particularly illuminating. Two nonhomologous genes encoding a thymidylate synthase occur in nature: the ThyA family and the ThyX family (Myllykallio et al. 2002). Whereas phages T4, other T-evens, PseudoT-evens, and SchizoT-evens encode for the ThyA thymidylate synthase gene, the 3 cyanophages encode for the ThyX thymidylate synthase gene that is very closely related to the gene in cyanobacteria (E value $< 10^{-75}$). This could indicate a recent nonhomologous displacement in cyanophages of the T4-like ancestral thymidylate synthase (from ThyA family) by a cyanobacterial thymidylate synthase (from the ThyX family). Cyanophages appear particularly inclined to have acquired genes from their hosts by LGT. As already noted, a large fraction of their genes of cellular origin are involved in photosynthesis and consequently could play a central role in the adaptation of these phages or their hosts to novel environments (Mann et al. 2005; Sullivan et al. 2005). Several first hits from other non-T4-type phage families were also observed (mainly from temperate tailed phages). The frequent LGT of nonubiquitous T4-type genes is in striking contrast with the apparently predominant vertical transmission of the core T4-type genes.

The “anomalous” behavior of the core subset of T4-type genes is then surprising. The evolution of other well-studied phages such as λ or Mu proceeds predominantly by frequent LGT, and as a consequence, their genomes appear to be fully mosaic in their structure composed of genes from

disparate origins. It seems likely that, in contrast, the T4-type phages have vertically inherited a restricted subset of genes clustered in 2 syntenous blocks from a common ancestor. The argument for the antiquity of this common T4 ancestor is reasonably compelling. First, the high level of genome divergence between all these phages testifies for a long evolutionary history (e.g., cyanophages’ genomes share only 14–21% of their genome with the genome of T4; fig. 4; Sullivan et al. 2005). Moreover, there is considerable divergence of amino acid sequence of their homologous proteins, and sequencing of T4 environmental sequences indicated that our knowledge of this diversity is still largely underestimated (Filee et al. 2005). T4-like phage shares homologous protein with their host as DnaB (gp41), the clamp loader subunit (gp62), or RecA protein (UvsX). The level of divergence of these homologous proteins between cyanophage and T4 is higher than those from cyanobacteria and *E. coli*. Together with the enormous phylogenetic range of their hosts (i.e., cyanobacteria and γ -proteobacteria) (table 1), these observations suggest that the last common ancestor of the T4-type phages occurred a long time ago, perhaps predating the split between cyanobacteria and other bacteria (–2.5 to 3.2 Ga; Battistuzzi et al. 2004).

What can we offer as an explanation for such a stable association of the ensemble of core genes? We note that these coinherited genes are organized in 2 syntenous blocks of genes, one expressed during the early phase of the infection (DNA replication and in DNA metabolism genes) and the other during the late phase of the infection (virion structural proteins). The conservation of the gene content and the synteny of these 2 segments is probably due, at least in part, to an intrinsic biochemical coupling of many of the functions that they encode. DNA replication and phage morphogenesis involve the coordination of many phage-encoded functions. These separate functions are organized into complex multiprotein assemblies with a precise structural organization of the constituent subunits. Like a ribosome, the correct structural arrangement of the components of such organelles is probably critical to their ability to function correctly. The complex web of protein–protein interactions of the constituent subunits must be conserved intact. Domain or gene swapping with a divergent homologue within such a complex assembly would probably most frequently lead to complete loss of function which would be lethal. There is probably an overwhelming negative selection operating against genetic tinkering with such extremely successful and complex biological machines (Jain et al. 1999). Such constraints can be explained by high probability that any random changes in such components of a nanomachine will lead to collapse rather than to an improvement in function. In such a constrained context, even the displacement of gene within its module could disturb it and its immediate neighbor’s gene expression with negative impact on the function of the ensemble. The coevolution in interacting nanomachine protein subunits has probably been remarkably conservative to avoid orthologous replacement of single genes within the syntenous blocks, even when such events would preserve the gene order.

Strong, precise, and critical protein–protein interactions are probably a major component of the genetic glue

that holds conserved genome segments together. A similar hypothesis has previously been suggested for the conservation of the virion structural modules in lambdoid phages (Casjens and Hendrix 1974; Juhala et al. 2000). The other component of the glue is the necessity of coordinating the gene expression of the constituents to assure that they are present in the correct quantities and correct order to assure proper assembly of the nanomachine. Because in the extremely compact phage genomes the regulatory cassettes that control gene expression are frequently located in the C-terminal coding portion of the upstream gene, it is thus easy to understand why changing the genetic context of a gene could have disastrous consequences on its neighbor's expression. The synergy of such dual constraints probably suffices to explain the maintenance of the synteny of the modules encoding nanomachines, but a third factor might also contribute to this effect.

If synteny of a module is maintained in a population of closely related phages, then a homologous recombination event between somewhat diverged modules will invariably generate a completely intact chimeric module. Occasionally, such chimeras may have adventitious features. If the modules are nonsyntenous, then recombination between them will frequently generate deletions that are nonviable because they lack essential components of the full module. Thus, maintaining synteny preserves the possibility for genetic exchange with similar modules in the gene pool. If synteny is not conserved in the population, then an abundant and potentially useful source of genetic diversity is lost. Even the small selective advantage of maintaining access to genetic diversity within a population of closely related phages may suffice to preserve synteny.

Conclusion

The discovery of 2 specific instances where LGT is essentially excluded in specific segments of phage genomes that are otherwise exceptionally prone to such events provides us with a unique opportunity. We postulate the existence of an interesting and unexpected constraint on the process of horizontal gene transfer. It is effectively reduced in segments of the genome that encode subcellular organelles or nanomachines. Independent confirmation of this hypothesis is maybe to be found in cellular organisms, where the prokaryotic ribosomal proteins are very syntenic (Tamames 2001), fairly congruent, and thus presumably less prone to LGT (Brochier et al. 2002; Matte-Tailliez et al. 2002). It is paradoxical and amusing that this insight into the molecular constraints that limit LGT should come from the study of rampantly mosaic phage genomes.

Supplementary Material

Supplementary material 1 and 2 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank W. F. Doolittle, Bob Weisberg, and Allan Campbell for critical discussions and D. Walsh and C. Brochier for critical reading of the manuscript. J.F. and

H.M.K. were supported by the Centre National de la Recherche Scientifique and by a grant from the Ministère de la Recherche (ACI-Microbiologie). E.B. was funded by a Canadian Institutes of Health Research grant MOP-4467.

Literature Cited

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–10.
- Balter M. 2000. Virology. Evolution on life's fringes. *Science* 289:1866–7.
- Baptiste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol* 5:33.
- Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* 4:44.
- Brochier C, Baptiste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* 18:1–5.
- Casjens S, Hendrix R. 1974. Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. *J Mol Biol* 90:20–5.
- Desplats C, Dez C, Tetart F, Eleaume H, Krisch HM. 2002. Snapshot of the genome of the pseudo-T-even bacteriophage RB49. *J Bacteriol* 184:2789–804.
- Filee J, Forterre P, Laurent J. 2003. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res Microbiol* 154:237–43.
- Filee 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* 54:763–73.
- Filee J, Tetart 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* 102:12471–6.
- Forterre P. 1999. Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Mol Microbiol* 33:457–65.
- Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. *Nature* 399:541–8.
- Hendrix RW. 2003. Bacteriophage genomics. *Curr Opin Microbiol* 6:506–11.
- Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* 96:2192–7.
- Hershey AD, Dove WF. 1971. Introduction to lambda. In: Hershey AD, editor. *The bacteriophage lambda*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. p 3–11.
- Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–6.
- Juhala RJ, Ford ME, Duda RL, Youton A, Hatfull GF, Hendrix RW. 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* 299:27–51.
- Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* 101:11013–8.
- MacLeod D, Charlebois RL, Doolittle F, Baptiste E. 2005. Deduction of probable events of lateral gene transfer through

- comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol* 5:27.
- Mann NH, Clokie MR, 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* 187:3188–200.
- Matte-Tailliez O, Brochier C, Forterre P, Philippe H. 2002. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* 19:631–9.
- Millard A, Clokie MR, Shub DA, Mann NH. 2004. Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* 101:11007–12.
- 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* 185:5220–33.
- Myllykallio H, Lipowski G, Leduc D, Filée J, Forterre P, Liebl U. 2002. An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* 297:105–7.
- Nolan JM, Petrov VM, Bertrand C, Krisch HM, Karam, JD. 2006. Genetic diversity among five T4-like bacteriophages. *Virology* <http://www.virologyj.com/content/3/1/30/>.
- Pedulla ML, Ford ME, Houtz JM, et al. (20 co-authors). 2003. Origins of highly mosaic mycobacteriophage genomes. *Cell* 113:171–82.
- Philippe H. 1993. MUST, a computer package of management utilities for sequences and trees. *Nucleic Acids Res* 21:5264–72.
- Proux C, van Sinderen D, Suarez J, Garcia P, Ladero V, Fitzgerald GF, Desiere F, Brussow H. 2002. The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. *J Bacteriol* 184:6026–36.
- Rohwer F. 2003. Global phage diversity. *Cell* 113:141.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508.
- Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–7.
- Sullivan MB, Coleman ML, Weigle P, Rohwer F, Chisholm SW. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* 3:e144.
- Susko E, Leigh J, Doolittle WF, Baptiste E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the $\{\gamma\}$ -Proteobacteria. *Mol Biol Evol* 23:1019–30.
- Tamames J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol* 2:RESEARCH0020.
- Waldor MK, Mekalanos JJ. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272:1910–4.

Edward Holmes, Associate Editor

Accepted May 25, 2006