J Mol Evol (2001) 52:117–128
DOI: 10.1007/s002390010140

JOURNAL OF MOLECULAR EVOLUTION

Phylogeny, Genome Evolution, and Host Specificity of Single-Stranded RNA Bacteriophage (Family Leviviridae)

Jonathan P. Bollback, John P. Huelsenbeck

Department of Biology, University of Rochester, Rochester, NY 14627, USA

Received: 3 March 2000 / Accepted: 17 October 2000

Abstract. Bacteriophage of the family Leviviridae have played an important role in molecular biology where representative species, such as Qβ and MS2, have been studied as model systems for replication, translation, and the role of secondary structure in gene regulation. Using nucleotide sequences from the coat and replicase genes we present the first statistical estimate of phylogeny for the family Leviviridae using maximum-likelihood and Bayesian estimation. Our analyses reveal that the coliphage species are a monophyletic group consisting of two clades representing the genera *Levivirus* and *Allolevivirus.* The *Pseudomonas* species PP7 diverged from its common ancestor with the coliphage prior to the ancient split between these genera and their subsequent diversification. Differences in genome size, gene composition, and gene expression are shown with a high probability to have changed along the lineage leading to the *Allolevivirus* through gene expansion. The change in genome size of the *Allolevivirus* ancestor may have catalyzed subsequent changes that led to their current genome organization and gene expression.

Key words: Phylogeny — Genome evolution — Leviviridae — RNA bacteriophage — Replicase gene — Coat gene — Bayesian inference — Markov chain Monte Carlo

Correspondence to: Jonathan P. Bollback; *e-mail:* bollback@brahms.biology.rochester.edu

Introduction

The family Leviviridae currently contains two genera (*Levivirus* and *Allolevivirus*) and three unclassified groups (a, b, and c) which include 24, 19, 20+, 11, and 3 taxa, respectively (Murphy et al. 1995). *Levivirus* and *Allolevivirus* each contain two distinct subgroups based upon serological cross-reactivity (Furuse 1987), molecular weight and density of the virion (Shapiro and Bendis 1975), sedimentation velocity of the viral particle (Shapiro and Bendis 1975), and replicase template activity (Miyake et al. 1971). The genus *Levivirus* contains the MS2-like (serogroup I) and GA-like (serogroup II) phage, whereas the genus *Allolevivirus* contains the Qβ-like (serogroup III) and SP-like (serogroup IV) phage (Murphy et al. 1995). The observations that all single-stranded RNA bacteriophage species have a common genomic organization, a high degree of similarity among replicases, identical utilization of host factors (S1, HF, EF-Tu, and EF-Ts) during replication, and strong similarities in translational control mechanisms (e.g., replicase synthesis is repressed by the coat protein) strongly suggest a common ancestor for the family (reviewed by van Duin 1988).

Single-stranded RNA bacteriophage are found throughout the world in bacterial isolates associated with the sewage and feces of mammals (Furuse 1987; Osawa et al. 1981). The Leviviridae appear to be absent from the bacteria associated with avian species (Osawa et al. 1981). However, this family can represent 90% of all RNA bacteriophages found in certain geographical isolates (e.g., Japan), although there is considerable geographic variation in subgroup representation and relative

abundance [see Furuse (1987) for a concise account of the families' ecological distribution]. The natural hosts of all species have not been determined conclusively, but they appear to be restricted to the gram-negative bacterial genera *Escherichia, Pseudomonas, Caulobacter, Salmonella,* and *Vibrio* with either an F pilus (Crawford and Gesteland 1964), a polar (somatic) pilus (Bradley 1972), or pili encoded by a plasmid carrying the drug resistance RP factor [e.g., RP1 and RP4 (van Duin 1988)]. All of the bacteriophage gain entrance to the host cell's cytoplasm via attachment to surface pili structures (Bayer at al. 1995; van Duin 1988; Shapiro and Bendis 1975; Bradley 1966).

Bacteriophage of the family Leviviridae have among the highest known mutation rates [10^{-3} bp/replication (Drake 1993)] and some of the smallest RNA genomes known (~3500–4200 nt). All species in the family for which the complete genome is currently known have four genes. These genes code for subunit II of replicase, a major coat protein, a maturation protein (a minor constituent of the virion involved in pilus recognition), and either a lysis protein in the case of *Levivirus* species and *Pseudomonas* species PP7 or a readthrough protein in the case of *Allolevivirus* species (van Duin 1988). Unlike the GA and MS2 subgroups (*Levivirus*) the four coding regions in the Qβ and SP subgroups (*Allolevivirus*) are oriented in a single reading frame. Moreover, in *Allolevivirus* the readthrough protein coding region shares the same initiation codon with the coat protein and is produced during translation by a low level of ribosome misincorporation of tryptophan (~5%) at the coat protein termination signal (Hofstetter et al. 1974; Weiner and Weber, 1971). Translation of these in frame overlapping genes is coupled; translational coupling of these proteins may ensure that each phage genome produces the correct ratio of the major coat protein and readthrough protein (van Himbergen et al. 1993). The readthrough protein represents 3–7% of the protein molecules composing the virion. The required number of readthrough protein copies in the virion is remarkably similar to the rate of ribosome misincorporation, supporting the translational coupling hypothesis (van Duin 1988).

A lysis gene is present in *Levivirus* (Beremand and Blumenthal 1979; Olsthoorn et al. 1995) and the *Pseudomonas* phage PP7, while the readthrough protein, present in *Allolevivirus,* is absent (Olsthoorn et al. 1995; van Duin 1988). Lysis in *Allolevivirus* is mediated by the maturation (A2) protein (Karnik and Billeter 1983; Winter and Gold 1983). Unlike the *Allolevivirus,* the coding regions in *Levivirus* and PP7 are initiated in different reading frames that vary depending on the group. The coat and lysis coding regions are overlapping [e.g., in MS2 the lysis gene overlaps both the 3′ end of the major coat protein and the 5′ end of the replicase gene (Fiers et al. 1976; Beremand and Blumenthal 1979; Atkins et al. 1979)] and exhibit coupled translation (Adhin and van Duin 1990). Mutation analysis of the coat protein in MS2 by Klovins et al. (1997a) demonstrated that translational coupling results from the effects of secondary structure on ribosome binding.

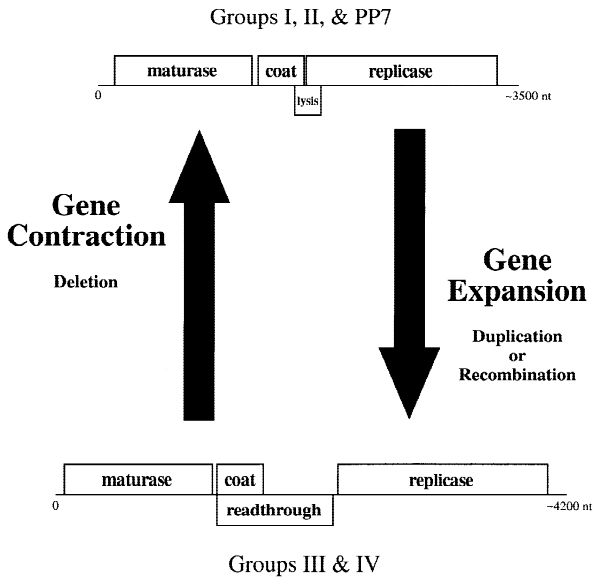

Fig. 1. Gene expansion and contraction hypotheses (modified from Mekler 1981).

Observations on the difference in genome size between the *Levivirus* (serogroups I and II) and the *Allolevivirus* (serogroups III and IV) have led to alternative hypotheses (see Fig. 1) about whether the ancestral phage genome size was large (i.e., Qβ-like) or small (i.e., MS2-like) (Hofstetter et al. 1974). It is unclear whether the difference in genome size between *Levivirus* and PP7 bacteriophage, and *Allolevivirus* bacteriophage represents an ancestral insertion or deletion event. Based on biological and physiochemical diversity, Furuse (1987) argued that the most probable pattern of genome evolution was through a major deletion in a Qβ-like ancestor giving rise to an MS2-like genome ("gene contraction" hypothesis; Fig. 1). An indirect prediction of this hypothesis is that the ancestral protophage contained a readthrough protein and was therefore more similar in genomic organization to *Allolevivirus* phage. A number of experiments (Mills et al. 1967, 1975; Schaffner et al. 1977; Klovins et al. 1997b) have demonstrated that deletions can occur in Qβ; when Qβ is passaged in vitro or in vivo, the phage undergo spontaneous deletions of parts of the genome. Moreover, at that time there were no known duplication or recombination and repair systems in RNA viruses and the "deletion" or "gene contraction" hypothesis had more biochemical support than an "insertion/duplication" or "gene expansion" hypothesis.

The gene expansion hypothesis (Fig. 1) requires some mechanism by which the ancestral phage could give rise to a larger daughter species. An increase in genome size could be caused by either a duplication or recombination event. Nucleotide insertions have been observed in a

number of bacteriophage experiments (e.g., see Klovins et al. 1997b), however, these have been mostly small in size. The duplication mode of increase predicts the occurrence of historical footprints, observed as sequence repeats, in the coat and readthrough proteins of groups III and IV genomes (Mekler 1981). Mekler (1981) observed a number of repeats and "quasi-repeats" of homologous nucleotide sequence tracts arranged in the same order within the Qβ coat-readthrough cistron. This suggested tentative support for a duplication mechanism in the origin of group III and IV coliphage (Mekler 1981). In addition, high sequence similarity within two different repeat groups suggests two separate and distinct duplication events (Mekler 1981). The alternative mechanism—recombination—has also become a plausible explanation as recent experimental work has demonstrated in vivo homologous recombination in Qβ (Palasingam and Shaklee 1992; Chetverin et al. 1991); recombination is also a likely phenomenon in MS2 (Olsthoorn and van Duin 1996). In addition, isolation of the RQ120 satellite RNA and comparison of its nucleotide sequence with Qβ and known *E. coli* sequences has identified at least one event of intermolecular non-homologous RNA recombination in vivo between Qβ and the *E. coli* $tRNA^{Asp}$ molecule (Munishkin et al. 1988). Rates of recombination have been estimated to be of the order of 10^{-8} per 1500 nt RNA segment (Palasingam and Shaklee 1992). Both in vivo homologous and intermolecular nonhomologous recombination among single-stranded RNA bacteriophage are possible mechanisms that could have resulted in an increase in the genome size of the *Allolevivirus* group.

The phage PP7, which infects *Pseudomonas aeruginosa* via a polar (somatic) pilus specific receptor, has been completely sequenced and shows a similar genomic size (3588 nt), organization, and composition with the *Levivirus* bacteriophages (e.g., genome size of MS2 is 3569 nt) (Olsthoorn et al. 1995). The amino acid sequences of the coat and maturation protein of the *Pseudomonas* phage 7S, *Caulobacter* phage Cb5, and phage PRR1 are more similar to the *Levivirus* than *Allolevivirus* phage (Golmohammadi et al. 1993; Dhaese et al. 1980, 1979). These phages are thought to be very distantly related to the F pilus specific coliphage due to their different host specificity, non-F pilus pathway into the host cytoplasm, and low amino acid sequence similarity of the coat (19–23%) and replicase proteins (42–45%) to the coliphages (Olsthoorn et al. 1995; Golmohammadi et al. 1993; Dhaese et al. 1980, 1979). If their distant phylogenetic position holds, this would suggest a monophyletic origin of F receptor specific coliphages, a small ancestral genome state prior to coliphage diversification, an ancestral genetic organization most similar to that found in non-*Allolevivirus* phage, and additional support for the gene expansion hypothesis (Hofstetter et al. 1974; Mekler 1981).

In this study, we present the results of a statistical phylogenetic analysis for all single-stranded RNA bacteriophage species which have been completely sequenced to date and for which genome organization has been determined or inferred from analysis of sequence similarity. The phylogenetic analysis allows us to test alternative hypotheses for evolution of the genome and host specificity in Leviviridae.

Materials and Methods

Sequences

A total of nine single-strand RNA bacteriophage species have been completely sequenced and were used in this study. All RNA sequences were obtained from the GenBank database [accession numbers X15031 (FR), J02467 (MS2), X03869 (GA), X07489 (SP), AF059243 (NL95), AF052431 (M11), AF059242 (MX1), X14764 (Qβ; replicase), M99039 (Qβ; major coat protein), and X80191 (PP7)]. The core region of the replicase gene (701 nt) and the complete sequence for the major coat protein gene (442 nt) were analyzed. In this paper the core region of the replicase gene is defined as starting with amino acid residue number 205 and ending with residue number 443 (referenced to the Qβ genome).

Sequence Alignment

The amino acid sequences, translated from the primary RNA nucleotide sequence for both genes, were aligned in ClustalW [version 1.7 using ClustalX version 1.63b Macintosh interface (Thompson et al. 1994)] without a user-specified tree and using the program search defaults. Once aligned by the program, the amino acid sequences were confirmed by eye and adjustments were made where necessary. The alignment was then back translated into the original RNA coding sequence. A copy of the replicase gene and coat protein gene alignment, in the form of a NEXUS file ('Leviviridae.replicase.nex' and 'Leviviridae. coat.nex'), can be obtained from the authors or directly from the following URL address: http://brahms.biology.rochester.edu/data.html.

Phylogenetic Analysis

The phylogeny of the single-stranded RNA bacteriophage was estimated by maximum-likelihood and Bayesian methods. PAUP* [version 4.0b4 (Swofford 1998)] and BAMBE [version 2.02b (Simon and Larget 1998; Larget and Simon 1999)] were used for maximum-likelihood and Bayesian inferences, respectively. Both methods assume that substitutions occur according to a time-homogeneous Poisson process. The rate of change from one nucleotide to another is contained in the instantaneous rate matrix **Q:**

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} \cdot & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & \cdot & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & \cdot & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & \cdot \end{pmatrix}$$

where r_{ij} is the rate of change from nucleotide i to nucleotide j and π_i is the stationary frequency of the ith nucleotide. The diagonals are specified such that the rows sum to 0. The instantaneous rate matrix shown above represents the most general time reversible model com-

monly used in phylogenetic inference and is referred to as the GTR model (Lanavé et al. 1984; Tavaré 1986; Rodríguez et al. 1990). Other commonly used models of DNA substitution are simply special cases of this model. For example, the K80 (Kimura 1980) and HKY85 (Hasegawa et al. 1985) models assume that the transitions have one rate and the transversions have another rate (i.e., $r_{AG} = r_{CT}$, $r_{AC} = r_{AT} = r_{CG} = r_{GT}$, $\kappa = r_{AG}/r_{AC}$). The JC69 (Jukes and Cantor 1969) and K80 models assume that the stationary frequencies of the bases are equal ($\pi_A = \pi_C = \pi_G = \pi_T$). Among-site rate variation was accommodated in four ways: by assuming that the rate at a site is unknown but (1) drawn from a gamma distribution with shape parameter α [designated + Γ (Yang 1994)], (2) drawn from a distribution in which a proportion of the sites cannot vary [designated +I (Hasegawa et al. 1985)], (3) drawn from a distribution in which a proportion of the sites cannot vary and the rate for the remaining sites is drawn from a gamma distribution with shape parameter α (designated +I+Γ), and (4) by assuming that sites are assigned to classes (in this case based on codon position) and the rate for each class is estimated separately (designated +SS). The program PAUP* implements all of these models, whereas BAMBE implements F84, HKY85, and TN93 (Tamura and Nei 1993) models of DNA substitution and the +SS model for accommodating among-site rate variation. One difference in model implementation between PAUP* and BAMBE is that the latter program models base frequency variation in each category separately. Therefore, comparisons of bootstrap parameter estimates and Bayesian posterior probability estimates include only kappa (κ; ratio of transitions to transversions) and relative rates by codon position (see below).

We used likelihood-ratio testing to determine which model was most appropriate for each of our data sets (Goldman 1993). Maximum-likelihood scores were calculated for four models of sequence evolution [JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), HKY85 (Hasegawa et al. 1985), GTR (Lanavé et al. 1984; Tavaré 1986; Rodríguez et al. 1990)] and the four methods of accommodating rate heterogeneity. The molecular clock was tested using the best model of sequence evolution as outlined above. Model comparisons of our data were performed in PAUP* using the maximum-likelihood tree. The null model is designated H_0 and the alternative model H_1. The likelihood-ratio test statistic is

$$\Lambda = \frac{\max[L(H_0)]}{\max[L(H_1)]}$$

For nested models (i.e., models for which the null model is a special case of the alternative), $-2\log\Lambda$ is asymptotically χ^2 distributed with q degrees of freedom (Wilks 1938). The degrees of freedom is the difference in the number of parameters that are free to vary between the null and the alternative models. In determining the best model for each gene the number of simultaneous comparisons, k, was large ($k = 19$). Therefore, the appropriate probability values for significance at the 5% level were adjusted using the Bonferroni correction (α/k). A probability value of less than or equal to 0.003 was considered significant. Based on the above approach, and contingent upon being available in both software packages, the model that demonstrated the best fit to the data was utilized in our analyses (HKY85 with site-specific rates). Both the molecular clock and the unconstrained branch length model results are presented to demonstrate the robustness of the conclusions presented.

The maximum-likelihood tree was found using a heuristic search implemented in PAUP*; the heuristic search used the taxon-bisection-and-reconnection (TBR) perturbation. All other model parameters were estimated from the data using maximum likelihood. The reliability of different clades in our maximum-likelihood estimate of phylogeny and confidence intervals for the model parameters was assessed using the nonparametric bootstrap (Felsenstein 1985). The bootstrap assesses confidence in particular clades by resampling the sequence data with replacement to generate new data sets. These pseudoreplicate data are analyzed in the same fashion as the original data and the reliability of a particular clade is taken as the proportion of bootstrap trees containing that particular clade (Felsenstein 1985). Traditionally, individual nucleotide sites are resampled when generating bootstrap pseudo-data sets. However, this strategy does not maintain the coding structure within the data set and will result in assignment of nucleotides to the incorrect class. Assignment of nucleotides to the incorrect class may have the effect of averaging out the relative difference in rates between classes. This could potentially lead to incorrect bootstrap estimates that are depressed or inflated when there are large differences in rates between codon positions. More importantly, if nucleotide changes within the codon are not independent then resampling of individual nucleotides instead of codons could introduce biases in bootstrap estimates. Bootstrap data sets ($n = 100$) were generated using the program CodonBootstrap v2.1 written in C language by one of the authors (J.P.B.). This program resamples a data matrix by codon and generates a batch file in the NEXUS format. These data sets were used to determine uncertainty in topology and other parameters of the substitution model using PAUP*.

Bayesian inferences are based upon the posterior probability distribution of a parameter. We used the program BAMBE to approximate the posterior probability of alternative phylogenies and substitution model parameters. BAMBE uses Markov chain Monte Carlo (MCMC) to evaluate the high-dimensional summations and integrals required to calculate posterior probabilities (Simon and Larget 1998). Posterior probabilities are estimated from the proportion of times a chain at stationarity visits a particular state. Therefore, early samples taken when the chain is not at stationarity are discarded. This period is called the burn-in. Each Markov chain starting from a random tree was run for 1 million generations after the burn-in. The first 1000 generations were discarded as burn-in. The chain was sampled every 100 generations; inferences from each run were based upon a total of 10,000 sampled trees.

Ancestral Reconstruction and Genome Evolution

The parsimony criterion was used to map ancestral reconstructions of genome structure (presence/absence of the lysis and readthrough coding regions) and host specificity (F plasmid) onto the most probable topology. This method attempts to minimize the number of evolutionary changes required to explain the distribution of a character across the tips (species) of a particular topology. The parsimony method may perform poorly when the topology of interest includes long branches (Maddison 1994). The parsimony mapping also assumes that the phylogenetic tree on which the character was reconstructed is correct. Phylogenetic uncertainty was accommodated by performing the reconstruction on all trees, weighted by the probability that the tree was correct.

Results

Model Selection

The log-likelihood scores for four models of sequence evolution and four methods for accommodating rate variation (see Methods) were compared. Some of the possible comparisons represent nonnested model comparisons (e.g., HKY85+I+Γ vs HKY85+SS) and cannot be compared using the χ^2 distribution. Therefore the relevant comparisons being reported are for nested models. A comparison of the models (e.g., HKY85) indicates that, for a given method for accommodating rate varia-

tion, the general time-reversible model gave the best fit to our replicase data set (GTR, ln$L$ = −5830.42; GTR+I, ln$L$ = −5722.27; GTR+Γ, ln$L$ = −5688.80; GTR+I+Γ, ln$L$ = −5687.60; GTR+SS, ln$L$ = −5621.57) and was statistically significant in all comparisons (P < 0.001) except Γ vs I+Γ. For the coat protein data set the general time-reversible model was also significantly better (P < 0.003) in all model comparisons, except GTR+Γ vs GTR+I+Γ, GTR vs HKY85 (equal rates), and GTR+I vs HKY85+1 (GTR, ln$L$ = −3167.90; GTR+I, ln$L$ = −3160.73; GTR+Γ, ln$L$ = −3130.58; GTR+I+Γ, ln$L$ = −3130.58; GTR+SS, ln$L$ = −3059.22). The best model overall, based on log-likelihood scores, was the general-time-reversible model with site-specific rates by codon position (ln$L$ = −5621.57 and ln$L$ = −3059.22, replicase and coat, respectively). This model is not available in the BAMBE software package, however, so the HKY85 model (HKY85+SS; ln$L$ = −5639.28 and ln$L$ = −3071.06, replicase and coat, respectively) was used for both the maximum likelihood and the Bayesian analyses to allow for comparisons between the methods. Maximum-likelihood analyses (data not shown) using the GTR+SS model produced identical tree estimates as the HKY85+SS model.

Comparison of the K80 and HKY85 log-likelihood scores indicates a nonsignificant deviation from equal nucleotide frequencies. The K80 model was not statistically worse than the HKY85 model in all comparisons (equal rates, P > 0.1; +I, P > 0.1; +I+Γ, P > 0.1; +Γ, P > 0.1; +SS, P > 0.05). The near-equality of nucleotide frequencies may have resulted from inherent properties of molecular folding or selection on secondary structure of the RNA molecule. The latter seems more likely since secondary structure and long-distance interactions have been shown to play a major role in replication (Klovins et al. 1998), translation (Klovins et al. 1997a; Groenveld et al. 1995; de Smit and van Duin 1994; Schmidt et al. 1987), virion assembly (Stockley et al. 1994; Witherell et al. 1991), and RNase protection (Klovins et al. 1997b). Regardless of the mechanism, if a majority of the nucleotides in the genome participate in stem structures and long-distance interactions, then equal nucleotide frequencies are expected due to the rules of complementarity.

We also tested the molecular clock assumption using a likelihood-ratio test (Felsenstein 1981). Comparison of the log-likelihood scores for the constrained (molecular clock) and unconstrained models of HKY85+SS and GTR+SS were unable to reject the molecular clock for both genes (replicase—HKY85+SS, P = 0.915; GTR+SS, P = 0.930; and coat—HKY85+SS, P = 0.723; GTR+SS, P = 0.686). Therefore, both maximum-likelihood heuristic searches under both the HKY85+SS and the GTR+SS models and a Bayesian analysis under the HKY85+SS model with the molecular clock enforced were conducted in addition to analysis under the unconstrained branch length model (see below).

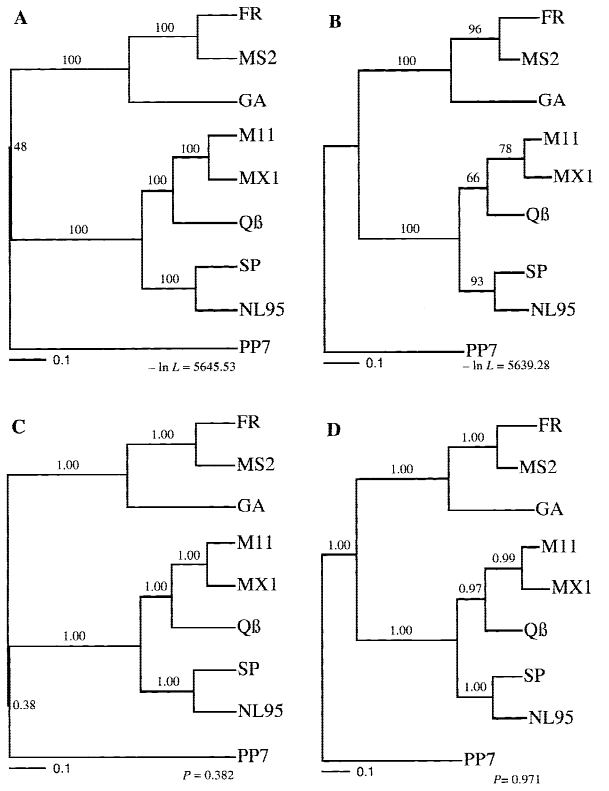

Fig. 2. Maximum-likelihood [**(A)** molecular clock and **(B)** unconstrained model] and Bayesian [**(C)** molecular clock and **(D)** unconstrained model] estimates of phylogeny for the replicase protein. The numbers on the interior branches indicate the bootstrap support or the posterior probability for each clade. Scale bars represent 0.1 expected substitution per site.

Maximum-Likelihood and Bayesian Analysis

The maximum-likelihood and Bayesian maximum posterior probability (MAP) estimates for both the unconstrained branch length model (i.e., nonclock) and the molecular clock model using HKY85+SS are shown in Fig. 2 (replicase) and Fig. 3 (coat). The log-likelihood score for the replicase trees are −5645.53 (clock) and −5639.28 (unconstrained). Posterior probabilities for the replicase gene are 0.382 (clock) and 0.971 (unconstrained). The log-likelihood score for the coat trees are −3075.40 (clock) and −3067.71 (unconstrained). Posterior probabilities for the coat gene are 0.990 (clock) and 0.450 (unconstrained). Table 1 summarizes all topologies recovered in the maximum-likelihood bootstrap searches and Bayesian analyses, with an occurrence of one or more replicate or greater than a 0.05 probability, respectively.

Uncertainty in the phylogenies was determined by the bootstrap and the posterior probabilities of clades for the

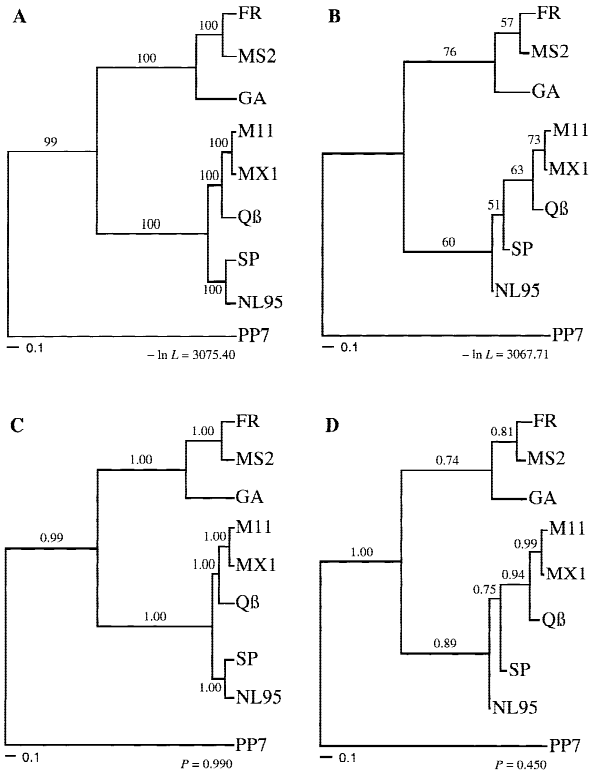

Fig. 3. Maximum-likelihood [**(A)** molecular clock and **(B)** unconstrained model] and Bayesian [**(C)** molecular clock and **(D)** unconstrained model] estimates of phylogeny for the coat protein. The numbers on the interior branches indicate the bootstrap support or the posterior probability for each clade. Scale bars represent 0.1 expected substitution per site.

maximum-likelihood and Bayesian analyses, respectively. Confidence intervals (95%), the mean, and the maximum-likelihood estimate for κ and relative rates for positions within the codon are shown in Table 2 for both genes. Credibility regions (95%), the mean, and the median of the posterior probabilities for κ and relative rates for positions within the codon are shown in Table 3 for both genes. Both maximum-likelihood estimates and Bayesian inferences of these different parameters were very similar between genes.

Ancestral Reconstruction of Host Specificity and Genome Evolution

Parsimony reconstructions for host specificity and genome evolution are represented in Fig. 4. The following traits were mapped onto the most probable tree under the molecular clock model; (A) evolution of F-plasmid specificity, (B) an increase in genome size through either recombination or duplication event(s), (C) loss of the lysis coding region and a shift in lysis function to the maturation gene, and (D) evolution of a novel readthrough protein involved in host infection. The probability of a particular reconstruction is the sum of the posterior probabilities for trees that are consistent with the proposed reconstructions. Probabilities for each reconstruction (see Fig. 4) are shown individually for each gene and combined in Table 4.

An inherent property of the molecular clock is inference of the root position on the tree. This allows the establishment of the polarity of character change and inference of a hypothetical ancestor (see Fig. 4). The common ancestor appears to have been very similar in genome architecture to *Levivirus* species and PP7 with a small genome size, absence of the readthrough gene, and presence of a lysis gene.

Discussion

Our phylogenetic estimates are concordant with the results from serological cross-reactivity data (Furuse 1987) and Inokuchi and co-workers' (1982) 3′-terminal nucleotide sequence data. These analyses placed FR and MS2 into a group [I], GA into a group [II], Qβ, M11, and MX1 into a single group [III], and SP and NL95 into a group [IV]. A UPGMA distance analysis using Furuse's (1987, Table 3.1) serological cross-reactivity data (not shown) for a single representative of each group produced a topology with a branching order consistent with the replicase and coat protein topologies. The lack of cross-reactivity data for PP7 prevents inferring the root position of the F-specific phages. However, our results suggest that serological typing contains a fair amount of phylogenetic information for the single-stranded RNA bacteriophages and the current assignment of other coliphage species into these serological groups is likely to be robust to future analysis of nucleotide sequence data. Yet this kind of data may not be sufficient to allow detailed and robust inference of the particular branching order among these groups. Therefore, the within genus branching orders of currently unsequenced phage remain uncertain.

Olsthoorn et al. (1995) suggested that the *Pseudomonas* phage PP7 branched off from the coliphages before divergence of this group into the current genera. The results from the maximum-likelihood and Bayesian molecular clock estimates provide support for the hypothesis that the coliphage are monophyletic (replicase, $\ln L = -5645.53$, $P = 0.299$; coat, $\ln L = -3075.40$, $P = 0.990$). The bootstrap proportions and posterior probabilities for the replicase gene indicate some uncertainty in the position of PP7 (BP = 48%, $P = 0.701$) (Figs. 2A and C), while results from the coat gene provide strong support for the basal placement of PP7 (BP = 99%, $P = 0.990$) (Figs. 3A and C). The suggestion that PP7 and 7S represent a recent horizontal transfer from *E. coli* to *Pseudomonas* (Olsthoorn et al. 1995) requires that PP7 is placed within one of the coliphage groupings. Our results indicate strong support for a basal position of PP7 and are not compatible with a hypothesis for the horizontal

Table 1. Support for alternative topologies found in maximum-likelihood bootstrap and Bayesian analyses[a]

	Maximum likelihood		Bayesian inference	
Tree	Coat (MC/U)	Replicase (MC/U)	Coat (MC/U)	Replicase (MC/U)
((((FR,MS2),GA),((SP,NL95),((M11,MX1),Qβ))),PP7)	**99**/0	**48/55**	**0.990**/0.068	0.299/**0.971**
((((FR,MS2),GA),PP7),((SP,NL95),((M11,MX1),Qβ)))	0/0	30/0	0.005/0.000	0.319/0.000
(((FR,MS2),GA),(((SP,NL95),((M11,MX1),Qβ)),PP7))	1/0	22/0	0.005/0.000	**0.382**/0.000
((((FR,MS2),GA),((SP,((M11,MX1),Qβ)),NL95)),PP7)	0/**21**	0/1	0.000/**0.450**	0.000/0.000
((((FR,MS2),GA),((((SP,NL95),Qβ),M11),MX1)),PP7)	0/0	0/16	0.000/0.000	0.000/0.000
(((((((FR,MS2),GA),NL95),SP),Qβ),M11),MX1),PP7)	0/13	0/0	0.000/0.000	0.000/0.000
(((FR,MS2),GA),((((SP,NL95),Qβ),(M11,MX1)),PP7))	0/0	0/12	0.000/0.000	0.000/0.027
((FR,(MS2,(GA,((SP,((M11,MX1),Qβ)),NL95)))),PP7)	0/10	0/0	0.000/0.041	0.000/0.000
((((FR,GA),MS2),((SP,((M11,MX1),Qβ)),NL95)),PP7)	0/8	0/0	0.000/0.051	0.000/0.000
(((((((FR,MS2),GA),NL95),SP),Qβ),M11,MX1),PP7)	0/8	0/0	0.000/0.012	0.000/0.000
(((FR,(GA,((SP,((M11,MX1),Qβ)),NL95))),MS2),PP7)	0/7	0/0	0.000/0.023	0.000/0.000
((((FR,MS2),GA),(SP,(NL95,((M11,MX1),Qβ)))),PP7)	0/0	0/6	0.000/0.014	0.000/0.000
((((FR,MS2),GA),((((SP,NL95),Qβ),MXI),M11)),PP7)	0/0	0/6	0.000/0.000	0.000/0.001
(((((((FR,MS2),GA),NL95),SP),(M11,MX1)),Qβ),PP7)	0/6	0/0	0.000/0.012	0.000/0.000
(((FR,(MS2,GA)),((SP,NL95),((M11,MX1),Qβ))),PP7)	0/0	0/4	0.000/0.000	0.000/0.000
(((((FR,MS2),GA),NL95),SP,((M11,MX1),Qβ)),PP7)	0/4	0/0	0.000/0.011	0.000/0.000
(((((((FR,GA),MS2),NL95),SP),Qβ),M11),MX1),PP7)	0/4	0/0	0.000/0.000	0.000/0.000
((((FR,GA),MS2),(((SP,NL95),(M11,MX1)),Qβ)),PP7)	0/4	0/0	0.000/0.000	0.000/0.000
((((((FR,MS2),GA),NL95),SP),((M11,MX1),Qβ)),PP7)	0/2	0/0	0.000/0.033	0.000/0.000
(((((FR,MS2),GA),NL95),(SP,((M11,MX1),Qβ))),PP7)	0/2	0/0	0.000/0.000	0.000/0.000
(((((((M11,MX1),Qβ),SP),NL95),(MS2,FR)),GA),PP7)	0/0	0/0	0.000/0.073	0.000/0.000
(((((((M11,MX1),Qβ),SP),NL95),GA),(MS2,FR)),PP7)	0/0	0/0	0.000/0.060	0.000/0.000
((((((M11,MX1),Qβ),SP),NL95),((MS2,GA),FR)),PP7)	0/0	0/0	0.000/0.022	0.000/0.000
((((((M11,MX1),Qβ),(SP,NL95)),(MS2,FR)),GA),PP7)	0/0	0/0	0.000/0.013	0.000/0.000

[a] Support is reported as the proportion of bootstrap replicates a particular tree is recovered in the case of maximum likelihood and as the posterior probability for Bayesian inference. Only trees having a frequency of ≥1% under maximum likelihood or ≥0.05 probability under Bayesian inference are reported. Trees are listed in Newick format: MC, molecular clock; U, unconstrained.

Table 2. Maximum-likelihood bootstrap confidence intervals (CI) and mean values for κ and site-specific relative rates[a]

	Coat		Replicase	
Parameter	Molecular clock	Unconstrained	Molecular clock	Unconstrained
κ				
95% CI	(1.29, 2.53)	(1.60, 2.51)	(1.61, 2.23)	(1.56, 2.26)
Mean	2.01	1.99	1.88	1.91
MLE	2.00	1.97	1.88	1.89
First position rate				
95% CI	(0.46, 0.69)	(0.47, 0.69)	(0.54, 0.68)	(0.52, 0.68)
Mean	0.58	0.57	0.61	0.60
MLE	0.57	0.58	0.61	0.61
Second position rate				
95% CI	(0.25, 0.42)	(0.25, 0.43)	(0.38, 0.53)	(0.39, 0.52)
Mean	0.33	0.33	0.45	0.45
MLE	0.34	0.34	0.45	0.45
Third position rate				
95% CI	(1.91, 2.26)	(1.91, 2.24)	(1.82, 2.04)	(1.81, 2.08)
Mean	2.09	2.09	1.94	1.95
MLE	2.09	2.08	1.94	1.94

[a] Maximum-likelihood estimates (MLE) of these parameters from analyses of the original data set are included for comparison.

transfer origin of *Pseudomonas* phage. The basal position of PP7 in the maximum-likelihood clock analyses places its divergence before the origin of the extant coliphages and, presumably, the origin of F specificity. Yet there still remains a degree of uncertainty in this from the Bayesian unconstrained analyses of the coat gene ($P = 0.450$) and Bayesian clock analyses of the replicase gene ($P = 0.382$). In the coat gene analysis we do not place considerable weight on this result because the unconstrained model does not provide a significantly better fit

Table 3. Bayesian credibility regions (CR), mean, and median values for κ and site-specific relative rates

	Coat		Replicase	
Parameter	Molecular clock	Unconstrained	Molecular clock	Unconstrained
κ				
95% CR	(1.45, 2.28)	(1.40, 2.24)	(1.61, 2.16)	(1.53, 2.04)
Mean	1.83	1.79	1.88	1.77
Median	1.82	1.78	1.87	1.76
First position rate				
95% CR	(0 48, 0.70)	(0.47, 0.70)	(0.55, 0.69)	(0.52, 0.67)
Mean	0.59	0.58	0.62	0.59
Median	0.58	0.57	0.62	0.59
Second position rate				
95% CR	(0.25, 0.41)	(0.24, 0.40)	(0.39, 0.51)	(0.37, 0.49)
Mean	0.32	0.31	0.45	0.43
Median	0.32	0.31	0.45	0.43
Third position rate				
95% CR	(1.94, 2.24)	(1.95, 2.26)	(1.84, 2.03)	(1.87, 2.07)
Mean	2.09	2.11	1.93	1.96
Median	2.09	2.11	1.93	1.96

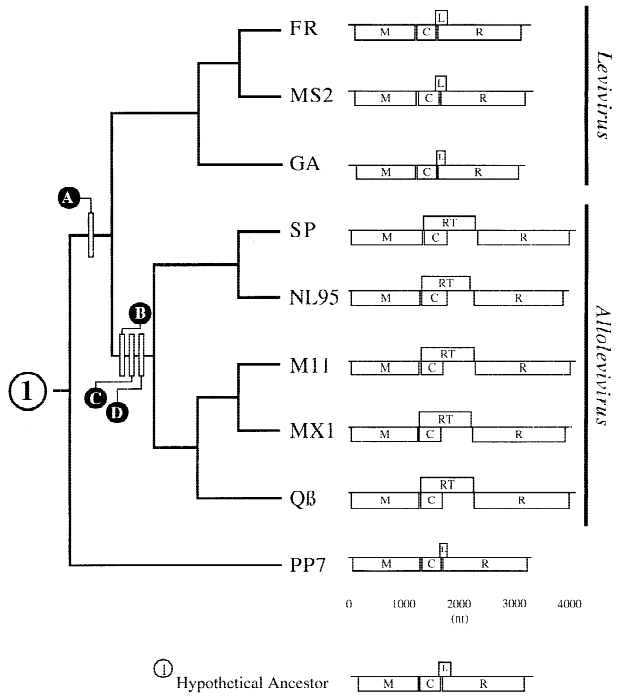

Fig. 4. Patterns of genome evolution, genome structure of the coding regions, reading frame and genome size are mapped onto the most probable topology using parsimony. Inferred changes are **(A)** evolution of F-plasmid specificity, **(B)** an increase in genome size through either a recombination or a duplication event(s), **(C)** loss of the lysis coding region and a shift in lysis function to the maturation gene, and **(D)** evolution of a novel readthrough protein involved in host infection. Coding regions are denoted as follows: M, maturation protein; C, coat or capsid protein; L, lysis protein; RT, readthrough protein; R, replicase protein.

Table 4. Summary of posterior probabilities associated with ancestral trait mapping shown in Fig. 4[a]

	Reconstruction			
	A	B	C	D
Coat Protein	0.990	1.000	1.000	1.000
Replicase	0.299	1.000	1.000	1.000
Total *P*	0.296	1.000	1.000	1.000

[a] Probabilities are the sum of the posterior probabilities associated with all trees that are the consistent with the most parsimonious placement of changes. The total probability (*P*) from analyses of the two genes is the product of the individual posterior probabilities.

over the clock model for the coat gene ($P = 0.723$). Moreover, analysis of the replicase gene under the clock model recovered three topologies of nearly equal probability. Differences among these topologies involve only the placement of PP7 (see Table 1). The similarity in the posterior probabilities assigned to these three topologies reflects uncertainty due to the very short branch present in the reconstruction placing PP7 basal to the coliphage. The inclusion of replicase sequence data for additional taxa in the future, such as the *Pseudomonas* phage 7S, *Caulobacter* phage, and additional coliphage species may provide additional support for the hypothesis of coliphage monophyly. Specifically, the inclusion of other coliphage sequences may help resolve the relationship of PP7 to the coliphage by breaking up the long branches leading to the *Levivirus* and *Allolevivirus* groups. Placement of the root may be further resolved with the inclusion of *Caulobacter* phage, PRR1 (P-plasmid phage), and additional *Pseudomonas* phage sequences.

Evolution of Host Specificity

Our results suggest that the *E. coli* F pilus receptor specific bacteriophages are monophyletic, and F receptor site specificity evolved after or at the time that polar specificity evolved (Fig. 2A) but not before. Taxonomic

restriction of the F plasmid to *E. coli* in nature suggests that an event of host switching occurred at this time (Fig. 2A). Thus host specificity and receptor specificity appear to have been nondependent evolutionary changes. The appearance of the F plasmid in the *E. coli* lineage may have made available a new host to a single-stranded RNA protophage and thus enabled further diversification within this group. A couple of lines of evidence support this hypothesis. First, our phylogenetic estimate exhibits strong support for the phage PP7 being basal to the F receptor specific phage species (Figs. 2A, B, and D and 3A–D), indicating the monophyletic nature of this group. Second, PP7 is a polar pilus specific phage and specializes on *Pseudomonas aeruginosa.* A primary assumption here is that all other non-F pilus specific phages would be placed outside the clade of F specific phage (they could be basal or form one or more monophyletic sister groups). The lack of additional polar specific phage species and *Caulobacter* species in our analysis could change the pattern apparent in our topology due to limited taxonomic sampling. Taxonomic sampling error would be problematic only if excluded taxa were to hold true phylogenetic positions within the ingroup and exhibit contradicting trait patterns to those observed here. This is not problematic for several reasons: (1) other *Pseudomonas* phages, such as 7S, exhibit biochemical similarities to PP7 (Shapiro and Bendis 1975), (2) both *Pseudomonas* and *Caulobacter* phage species utilize polar instead of F pili (Shapiro and Bendis 1975), and (3) biochemical measures such as the virion diameter and sedimentation weight of the viral particle are more similar within *Pseudomonas* and *Caulobacter* phage species than among them (Shapiro and Bendis 1975).

Single-stranded RNA bacteriophages have been observed to infect only three genera of bacteria (*Escherichia, Pseudomonas,* and *Caulobacter*) and phages from each group are unable to infect across these genera (Shapiro and Bendis 1975; Furuse 1987). The one exception to this rule of generic restriction is those phages specializing on bacteria with RP plasmid-mediated pili (PRR1). This phage appears to have a very broad host range (Shapiro and Bendis 1975), which may be the result of the broad occurrence of P-type plasmids in gram-negative bacteria. This high degree of host specificity has been suggested to be the result of biochemical differences in the surface pili structure found in these genera (Shapiro and Bendis 1975), suggesting that plasmids play a significant role in restricting bacteriophage host range. Given the high degree of host specificity among single-stranded RNA bacteriophage, it seems reasonable that the bacterial genera demarcate true monophyletic groups, except in the case of P-type drug resistance plasmid specialists. The question of the placement of PRR1 on the family tree is intriguing and, once known, may enhance our understanding of the evolutionary patterns of host specificity within the Leviviridae and to what extent plasmid evolution in bacteria has influenced bacteriophage evolution in the family.

Evolution of Genome Organization and Composition

Phages in the genus *Levivirus* and those specializing on *Pseudomonas, Caulobacter,* and P-type plasmids (PRR1) have a different genome organization than phages of the genus *Allolevivirus.* A number of differences are immediately apparent between these two groups: (1) the lack of a readthrough protein in *Levivirus* phage and PP7 (Olsthoorn et al. 1995; van Duin 1988), (2) the presence of a lysis gene in *Levivirus* phage and PP7 (Olsthoorn et al. 1995; van Duin 1988), and (3) the smaller genome size in *Levivirus* phage and PP7. One hypothesis of genome evolution is that the ancestral condition was similar to the *Allolevivirus* phage, and the smaller genome, lack of a readthrough protein, and presence of a lysis gene occurred by an ancestral deletion in the readthrough region of roughly 600 nucleotides (Furuse 1987). This hypothesis makes two predictions: (1) the *Allolevivirus* phage should occupy a more basal position on the phylogeny and (2) all phage species with an MS2-PP7-type genome should share a more recent common ancestor to each other than either do to the *Allolevivirus* phage.

Alternatively, the ancestral condition could have been more similar to that found in *Levivirus* and PP7 phages. This hypothesis postulates a genome insertion due to either recombination or a duplication event within the genome of the *Allolevivirus* protophage increased the genome size from small to large (~600 nt). Evidence for homologous recombination in the family (Olsthoorn and van Duin 1996; Palasingham and Shaklee 1992; Chetverin et al. 1991) and intramolecular duplication(s) in Qβ (Mekler 1981) favor this hypothesis. In addition, this new intercistronic region between the major coat protein coding region and the replicase coding region evolved a novel function as the readthrough protein. The loss of function in the lysis gene could have occurred before or after the evolution of the readthrough protein's current function. The C terminus of the lysis amino acid sequence in MS2 and PP7 is composed of 30 mostly hydrophobic amino acids (Olsthoorn et al. 1995). This is typical of other known single-gene viral lysis proteins [e.g., ϕX174 (reviewed by Young et al. 2000)]. The lysis gene is thought to act in a similar manner to the E protein of ϕX174, which inhibits peptidoglycan synthesis by targeting the MraY protein of *E. coli.* This host protein is necessary in peptidoglycan synthesis (Bernhardt et al. 2000). In *Allolevivirus* species the lysis function is controlled by the maturation gene product (Karnik and Billeter 1983; Winter and Gold 1983). The lytic role of the maturation protein in causing lysis in *Allolevivirus* phage makes it reasonable, then, to postulate that loss of the lysis gene would have had little fitness cost to the pro-

tophage in which the lysis gene was lost. The gene expansion hypothesis makes the following predictions: phage species of the genus *Allolevivirus* should (1) be monophyletic, (2) not hold a basal position in the phylogeny, and (3) share a more recent common ancestor with other F specific coliphage.

Our results indicate that the most probable placement of PP7 is basal to the coliphage species. As discussed above the F specificity most likely evolved once along the lineage leading from the common ancestor with the *Pseudomonas* phage (Fig. 4A). The posterior probability associated with PP7 being basal to the coliphage is 0.296 compared to the posterior probability of PP7 being more closely related to either the *Allolevivirus* or the *Levivirus* 0.007. In addition, PP7 is never found nested within the clades representing these genera in the unconstrained branch length analyses.

The most probable hypothetical ancestral protophage genome architecture is shown in Fig. 4. The ancestral phage appears to have been small in size, similar to the *Levivirus* phage, lacked a readthrough protein, and contained a coding region for a lysis protein. This reconstruction holds regardless of the particular maximum-likelihood or Bayesian MAP phylogeny favored (Figs. 2 and 3). Even with the placement of PP7 sister to either the *Allolevivirus* or the *Levivirus*, the most parsimonious reconstruction is unchanged (see Table 1 for alternative topologies). The differences between these alternative placements of PP7 still place all of the genome changes (duplication/recombination, loss of lysis function, and evolution of novel readthrough function) along the internal branch leading to the *Allolevivirus* phage.

A reasonable scenario for these changes can be envisioned as follows: the ancestral lineage leading to the *Allolevivirus* (1) underwent either a duplication event(s) or recombination event(s) increasing the genome size. This increase in size may have decoupled translation between the coat protein and the lysis protein, possibly rendering the later unexpressed. This increase in size may have also compensated for the loss in lysis capabilities by upregulating the expression of the maturation protein which is involved in cell lysis in the *Allolevivirus.* Current research by Groenveld et al. (1995) has shown that expression of the maturation (A) protein in MS2 is regulated by the kinetics of RNA folding via a long-distance interaction—the ribosome binding site is exposed only for a brief period of time on the positive strand during replication and translation, resulting in lower levels of expression. In the *Allolevivirus* species the long-distance interaction is further downstream (~400 nt), presumably extending the amount of time that the ribosome binding site is exposed resulting in higher levels of expression and lysis (Karnik and Billeter 1983; Winter and Gold 1983). After increase in the genome size (2) evolution of the readthrough coding region evolved through substitutions in the ancestral coat stop codon (UAA or UAG) to a UGA codon, allowing for periodic misincorporation of tryptophan and subsequent ribosome readthrough. Ribosome readthrough of the coat protein is observed only in the *Allolevivirus* species. If the path to increase in size was via intramolecular duplication of the major coat region, as has been suggested by Mekler (1981), it is consistent with our understanding of the role that duplication plays in the origin and evolution of a novel protein—in this case involved in the virion structure and host infection. A more difficult aspect to explain is the need of the ancestral *Allolevivirus* phage for a gene specializing in infectivity when it presumably already contained a current mechanism. Initially it may have represented an additional source of a very similar coat protein and was not involved in the process of infection. Similarity in function with the major coat protein may have allowed its maintenance in the genome by imposing little or no cost to the protophage. Additionally, the increased genome length may have enabled upregulation of the maturation protein, necessary for cell lysis in the *Allolevivirus* phage. The role of the readthrough protein (3) in infection may then have evolved a considerable time after the duplication event, possibly in response to selection for increased adsorption efficiency or changes in host pilus structure. The latter explanation seems less likely because *E. coli* hosts sensitive to *Allolevivirus* phage are also sensitive to *Levivirus* phage. The former explanation seems more plausible since, once the maturation protein began to play the major role in cell lysis, it might be expected that such changes had negative effects on adsorption efficiencies. If this was the case, then it would have been selectively advantageous to establish a new pathway to infection—resulting in the evolution of the readthrough protein's current function in adsorption. Therefore, upregulation of the maturation protein coupled with evolution of the readthrough protein may have imposed a direct fitness benefit, preserving this new region through purifying selection against spontaneous deletion mutants.

Conclusions

RNA bacteriophages of the family Leviviridae have been the focus of intense study of the molecular mechanisms of replication, translation, gene regulation, and secondary structure over the last 30 years. Although this group has received a tremendous amount of attention, few researchers have addressed questions pertaining to evolutionary patterns and phylogenetic relationships. In this study we present the first statistical estimate of the phylogeny for the family based on an analysis of the replicase and major coat protein coding regions.

Our phylogenetic results are consistent with estimates of relatedness and classification schemes derived from serological cross-reactivity data (Furuse 1987). Group I

species FR and MS2 are found to be a monophyletic clade which shares a close relationship with the group II species GA. Together these two groups make up the genus *Levivirus*, which itself is indicated to be monophyletic. Group III species Qβ, M11, and MX1 are monophyletic and share a close affinity to SP and NL95 species of group IV. Together these two groups comprise the genus *Allolevivirus* and have also been found to be monophyletic. Based on the basal phylogenetic position of the *Pseudomonas* phage PP7, the coliphage species represent a monophyletic group sharing a common ancestor with the *Pseudomonas* bacteriophage species. Thus, F plasmid specificity (coliphagy) must have evolved once after the split with PP7. This split may have tracked the divergence of the ancestor of *Pseudomonas* and *Escherichia* or may represent a horizontal host switch from *Pseudomonas* or another gram-negative bacteria to *Escherichia.* Due to the very short internal branch between PP7 and the split between the *Levivirus* and the *Allolevivirus* phages, a degree of uncertainty in its placement still remains, although external evidence provides additional support for its basal position. As a result, *Pseudomonas* phages do not appear to have originated from a horizontal host transfer event from within the coliphage.

An understanding of phylogenetic relationships and the polarity of character change enables questions of genome evolution to be explicitly addressed. Marked differences in genome structure and organization have been observed between the *Allolevivirus* and the *Levivirus*/PP7 phages. The *Allolevivirus* have a larger genome, lack the lysis gene, and contain a readthrough protein involved in infection, and cell lysis is mediated by the maturation protein (see Fig. 4). It has been controversial whether the increase in genome size was the result of gene expansion (duplication/recombination) or gene contraction (deletion). We have found that the phylogeny for the family supports the gene expansion hypothesis. After the divergence of the *Allolevivirus* and *Levivirus* lineages, but before the diversification of *Allolevivirus*, the ancestral phage genome increased in size on the order of 600 nt. It remains unclear whether this expansion was the result of recombination or intramolecular duplication. Recombination has been demonstrated to occur in the family, and Mekler (1981) has identified sequence tracts that are repeated downstream of the major coat protein of Qβ, consistent with a duplication event(s). A number of additional changes appear to have been coupled with this genome expansion—notably the loss of the lysis coding region, the evolution of a novel protein (readthrough) involved during infection of the host, and an increase in translation of the maturation protein. Each of these changes also maps to the same branch as the genome expansion. Interestingly, an increase in genome size may be implicated in catalyzing all of these changes via changing the secondary structure of the RNA molecule. Secondary structure plays a predominant role in gene regulation and replication. An increase in the distance between long-distance RNA–RNA interactions may have upregulated the translation of the maturation protein, and an increase in the distance between the stop codon of the coat protein and the start codon of the lysis gene may have decoupled ribosome reinitiation and lysis translation, leading to its loss. The lack of a lysis gene in *Allolevivirus* is compensated for by the higher levels of the maturation protein which functions in phage release.

Genome evolution in viruses is typically believed to be characterized by an economization of the genome. However, our analyses of the family Leviviridae indicate that increases in genome size have played an important role in viral evolution.

Acknowledgments. We thank Kenneth G. Karol, Andrea J. Betancourt, Daven C. Presgraves, and Bret Larget for helpful comments and suggestions. This work was supported by funding from the National Science Foundation (MCB-0075404 and DEB-0075406) to J.P.H.

References

Adhin MR, van Duin J (1990) Scanning model for translational reinitiation in eubacteria. J Mol Biol 213:811–818

Atkins JF, Steitz JA, Anderson CW, Model P (1979) Binding of mammalian ribosomes to MS2 phage RNA reveals an overlapping gene encoding a lysis function. Cell 18:247

Bayer M, Eferl R, Zelling G, Teferle K, Dijkstra A, Koraimann G, Högenauer G (1995) Gene 19 of plasmid R1 is required for both efficient conjugative DNA transfer and bacteriophage R17 infection. J Bacteriol 177:4279–4288

Beremand MW, Blumenthal T (1979) Overlapping genes in RNA phage: A new protein implicated in cell lysis. Cell 18:257

Bernhardt TG, Roof WD, Young R (2000) Genetic evidence that the bacteriophage ϕX174 lysis protein inhibits cell wall synthesis. Proc Natl Acad Sci USA 97:4297–4302

Bradley DE (1966) The structure and infective process of a *Pseudomonas aeruginosa* bacteriophage containing ribonucleic acid. J Gen Microbiol 45:83–96

Bradley DE (1972) Shortening of *Pseudomonas aeruginosa* pili after RNA-phage adsorption. J Gen Microbiol 72:303–319

Chetverin AB, Chetverina HV, Munishkin AV (1991) On the nature of spontaneous RNA synthesis by Qβ replicase. J Mol Biol 222:3–9

Crawford EM, Gesteland RF (1964) The adsorption of bacteriophage R17. Virology 22:165–167

de Smit M, van Duin J (1994) Another role for the Shine-Dalgarno interaction. J Mol Biol 235:173–184

Dhaese P, Vandeherckhove JS, van Montagu M (1979) The primary structure of the coat protein of the broad-host-range RNA bacteriophage PRR1. Eur J Biochem 94:375–386

Dhaese P, Lenaerts A, Gielen J, van Montagu M (1980) Complete amino acid sequence of the coat protein of the *Pseudomonas aeruginosa* bacteriophage PP7. Biochim Biophys Acta 94:1394–1400

Drake J (1993) Rates of spontaneous mutation rates among RNA viruses. Proc Natl Acad Sci USA 90:4171–4175

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791

Fiers W, Contreras R, Duerinck F, Haegman C, Iserentant D, Merregaert J, Min Jou J, Molemans W, Raeymakers A, Vandenberghe A, Volckaert C, Isebaert M (1976) Complete nucleotide sequence

of bacteriophage MS2-RNA: primary and secondary structure of the replicase gene. Nature 260:500–507

Furuse K (1987) Distribution of coliphages in the environment: general considerations. In: Goyal SM (ed) Phage ecology. Wiley, New York

Goldman N (1993) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198

Golmohammadi R, Vålegard K, Fridborg K, Liljas L (1993) The refined structure of bacteriophage MS2 at 2.8 Å resolution. J Mol Biol 234:620–639

Groenveld H, Thimon K, van Duin J (1995) Translational control of maturation-protein synthesis in phage MS2: A role for the kinetics of RNA folding? RNA 1:79–88

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by molecular clock of mitochondrial DNA. J Mol Evol 21:160–174

Hofstetter H, Monstein HJ, Weissmann C (1974) The read-through protein A1 is essential for the formation of viable Qβ particles. Biochim Biophys Acta 374:238

Inokuchi Y, Hirashima A, Watanabe I (1982) Comparison of the nucleotide sequences at the 3′-terminal region of RNAs from RNA coliphages. J Mol Biol 158:711–730

Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro H (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Karnik S, Billeter M (1983) The lysis function of RNA bacteriophage Qβ is mediated by the maturation (A2) protein. EMBO J 2:1521–1526

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Klovins J, Tsareva N, de Smit M, Berzins V, van Duin J (1997a) Rapid evolution of translational control mechanisms in RNA genomes. J Mol Biol 265:372–384

Klovins J, van Duin J, Olsthoorn CL (1997b) Rescue of the RNA phage genome from RNase III cleavage. Nucleic Acids Res 25:4201–4208

Klovins J, Berzins V, van Duin J (1998) A long range interaction in Qβ RNA that bridges the thousand nucleotides between the M-site and the 3′ end is required for replication. RNA 4:948–957

Lanavé C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Larget B, Simon D (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16:750–759

Maddison D (1994) Phylogenetic methods for inferring the evolutionary history and process of change in discretely values characters. Annu Rev Entomol 39:267–292

Mekler P (1981) Determination of nucleotide sequences of the bacteriophage Qβ genome: Organization and evolution of an RNA virus, PhD thesis. University of Zurich, Zurich

Mills D, Peterson R, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci USA 58:217–224

Mills D, Kramer FR, Dobkin C, Nishihara T, Spiegelman S (1975) Nucleotide sequence of microvariant RNA: Another small replicating molecule. Proc Natl Acad Sci USA 72:4252

Miyake T, Haruna I, Shiba T, Itoh YH, Yamane K, Watanabe I (1971) Grouping of RNA phages based on template specificity of their RNA replicases. Proc Natl Acad Sci USA 68:2022–2024

Munishkin AV, Voronin LA, Chetverin AB (1988) An *in vivo* recombinant RNA capable of autocatalytic synthesis by Qβ replicase. Nature 333:473–475

Murphy FA, Fauquet CM, Bishop DHL, Ghabrial SA, Jarvis AW, Martelli GP, Mayo MA, Summers MD (1995) Virus taxonomy: The classification and nomenclature of viruses. The sixth report of the International Committee on Taxonomy of Viruses. Springer-Verlag, Vienna

Olsthoorn RC, van Duin J (1996) Random removal of inserts from an RNA genome: Selection against single-stranded RNA. J Virol 70: 729–736

Olsthoorn RC, Garde G, Dayhuff T, Atkins JF, van Duin J (1995) Nucleotide sequence of a single-stranded RNA phage from *Pseudomonas aeruginosa:* Kinship to coliphages and conservation of regulatory RNA structures. Virology 206:611–625

Osawa S, Furuse K, Watanabe I (1981) Distribution of ribonucleic acid coliphages in animals. Appl Env Microbiol 41:909–911

Palasingam K, Shaklee PN (1992) Reversion of Qβ RNA phage mutants by homologous RNA recombination. J Virol 66:2435–2442

Rodríguez F, Oliver J, Marín A, Medina J (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142:485–501

Schaffner W, Ruegg KJ, Weissman C (1977) Nanovariant RNA's: Nucleotide sequence and interaction with bacteriophage Qβ replicase. J Mol Biol 117:877

Schmidt B, Berkhout B, Overbeek G, van Strien A, van Duin J (1987) Determination of the RNA secondary structure that regulates lysis gene expression in bacteriophage MS2. J Mol Biol 195:505–516

Shapiro L, Bendis I (1975) RNA phages of bacteria other than *E. coli.* In: N Zinder (ed) RNA phages. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Simon D, Larget B (1998) Bayesian analysis in molecular biology and evolution (BAMBE), Version 1.01 beta. Department of Mathematics and Computer Science. Duquesne University, Pittsburgh, PA

Stockley P, Stonehouse N, Vålegard K (1994) Molecular mechanism of RNA phage morphogenesis. Int J Biochem 26:1249–1260

Swofford D (1998) PAUP*: Phylogenetics analysis using parsimony and other methods. Sinauer Associates, Sunderland, MA

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tavaré S (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. Lect Math Life Sci 17:57–86

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

van Duin J (1988) Single stranded RNA bacteriophages. In: Fraenkel-Conrat H, Wagner R (eds) The viruses. Plenum, New York

van Himbergen J, van Geffen B, van Duin J (1993) Translational control by a long range RNA-RNA interaction; A base substitution analysis. Nucleic Acids Res 21:1713–1717

Weiner AM, Weber K (1971) Natural read-through at the UGA termination signal of the Qβ coat protein cistron. Nature New Biol 234:206

Wilks S (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9:554–560

Winter RB, Gold L (1983) Overproduction of bacteriophage Qβ maturation protein leads to cell lysis. Cell 33:877–885

Witherell GW, Gott JM, Uhlenbeck OC (1991) Specific interaction between RNA phage coat proteins and RNA. Progr Nucl Acid Res Mol Biol 40:185–220

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39:306–314

Young R, Ing-Nang W, Roof WD (2000) Phages will out: Strategies of host cell lysis. Trends Microbiol 8:120–128