BMC Genomics

BioMed Central

Research article

Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses

David T Pride*[1], Trudy M Wassenaar[2], Chandrabali Ghose[3] and Martin J Blaser[4]

Address: [1]Department of Medicine, Division of Infectious Diseases And Geographic Medicine, Stanford University School of Medicine, Stanford, CA, USA, [2]Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany, [3]Department of Medicine, Division of Infectious Diseases, Harvard Medical School, Boston, MA, USA and [4]Departments of Medicine and Microbiology, New York University School of Medicine and VA Medical Center, New York, NY4, USA

Email: David T Pride* - dpride@stanford.edu; Trudy M Wassenaar - wassenaar_t@yahoo.co.uk; Chandrabali Ghose - cghose@partners.org; Martin J Blaser - Martin.blaser@med.nyu.edu

* Corresponding author

Published: 18 January 2006

BMC Genomics 2006, **7**:8 doi:10.1186/1471-2164-7-8

Received: 27 September 2005

Accepted: 18 January 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/8

Abstract

Background: Virus taxonomy is based on morphologic characteristics, as there are no widely used non-phenotypic measures for comparison among virus families. We examined whether there is phylogenetic signal in virus nucleotide usage patterns that can be used to determine ancestral relationships. The well-studied model of tail morphology in bacteriophage classification was used for comparison with nucleotide usage patterns. Tetranucleotide usage deviation (TUD) patterns were chosen since they have previously been shown to contain phylogenetic signal similar to that of 16S rRNA.

Results: We found that bacteriophages have unique TUD patterns, representing genomic signatures that are relatively conserved among those with similar host range. Analysis of TUD-based phylogeny indicates that host influences are important in bacteriophage evolution, and phylogenies containing both phages and their hosts support their co-evolution. TUD-based phylogeny of eukaryotic viruses indicates that they cluster largely based on nucleic acid type and genome size. Similarities between eukaryotic virus phylogenies based on TUD and gene content substantiate the TUD methodology.

Conclusion: Differences between phenotypic and TUD analysis may provide clues to virus ancestry not previously inferred. As such, TUD analysis provides a complementary approach to morphology-based systems in analysis of virus evolution.

Background

Eukaryotic viruses and bacteriophages exist in numerous forms and are capable of infecting disparate hosts. The taxonomy of viruses is based upon morphological features, including capsid and tail structures, specific type of genetic material, and mechanism of replication and assembly [1,2]. Genetic comparison across virus species has been complicated by generally different rates of gene evolution, thus, their overall classification rests on phenotypic and morphologic characteristics [3]. Horizontal gene transfer has been substantial in virus evolution [4-7], complicating reproduction of ontogeny based on the cur-

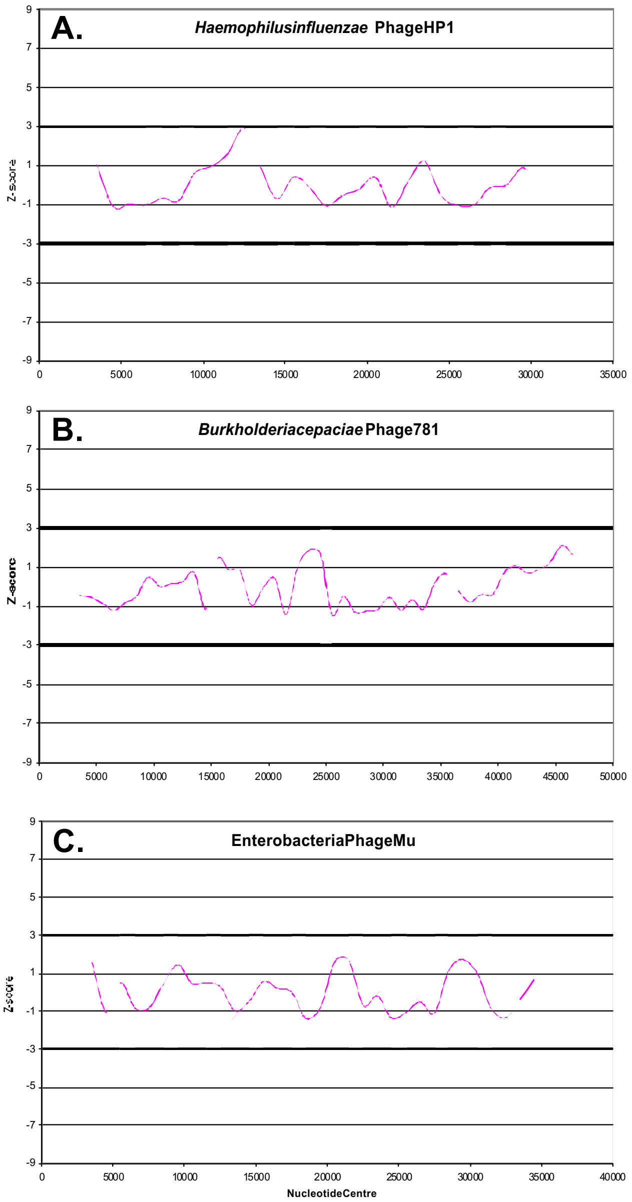

Figure 1
Tetranucleotide difference analysis of representative bacteriophage genomes (*H. influenzae* phage HP1, *B. cepaciae* phage 781, and Enterobacteria phage Mu). Tetranucleotide differences were determined with window and step sizes of 5,000 and 1,000, respectively, and Z-scores were determined as described in Materials and Methods. Solid black lines represent Z-scores of ± 3.

rent presence of particular loci. Analysis of phylogenies based on phenotypic systems is limited by convergent evolution, in which like characteristics are evolved by unrelated organisms to suit particular niches or evolutionary requirements [8]. Taxonomy of bacteriophages also has been based on morphologic characteristics [9,10]. Tail morphology forms the basis for bacteriophage classification into 3 separate families: Myoviridae (contractile tails), Podoviridae (short tail stubs), and Siphoviridae (long tails) [1]. Studies examining phage tail assemblies [11] have not ascertained whether the source of tail characteristics has phylogenetic significance, indicating the likely existence of polyphylogeny within these phage groups [12].

Unlike viruses, prokaryotes most commonly have been taxonomically classified according to a single locus, 16S rRNA [13-15]. Because of its relatively conservative rate of evolution, and presumed rarity of horizontal transfer due to its functional constraints, the 16S rRNA locus is believed to serve as an accurate marker of recent common ancestry [15].

Evaluation of prokaryotic ancestry based on shared gene content also has been proposed, guided by the principle that prokaryotes have cores of essential genes, whose presence or absence is evolutionarily significant [16-18]. Alternatively, there has been analysis of phylogenetic signal in whole-genome nucleotide usage patterns and is consistent with the predicted phylogenetic structure of prokaryotes based on 16S rRNA [19]. Whole-genome approaches are less biased by any single locus [20], with horizontal transfer being an intrinsic part of the signal, reflecting the current. Using Zero-Order Markov algorithms, we have previously demonstrated that patterns of tetranucleotide usage patterns retain phylogenetic signal among many related prokaryotes [19]. Differences between tetranucleotide usage and 16S rRNA in ancestral reconstruction likely are due to horizontal influences, such as extensive recombination, and/or presence of restriction/modification systems.

We sought to better understand the evolution of viruses by comparing methods for reproducing ancestry including tetranucleotide usage deviation, and shared gene content to construct a framework independent of phenotypic analysis. Our goals were to understand whether: 1) phylogenetic signal is retained in nucleotide usage patterns of viruses; 2) phylogenetic structures based on nucleotide usage patterns and tail morphology are similar; 3) nucleotide usage patterns in viruses are primarily determined by gene content; and 4) whether their prokaryotic hosts exert a substantial influence on bacteriophage nucleotide usage patterns.

Results

Conservation of tetranucleotide usage patterns across bacteriophage genomes

Nucleotide usage patterns are unique among different prokaryotes, providing distinct signatures [19,21,22] that are well-conserved across each genome, except for DNA

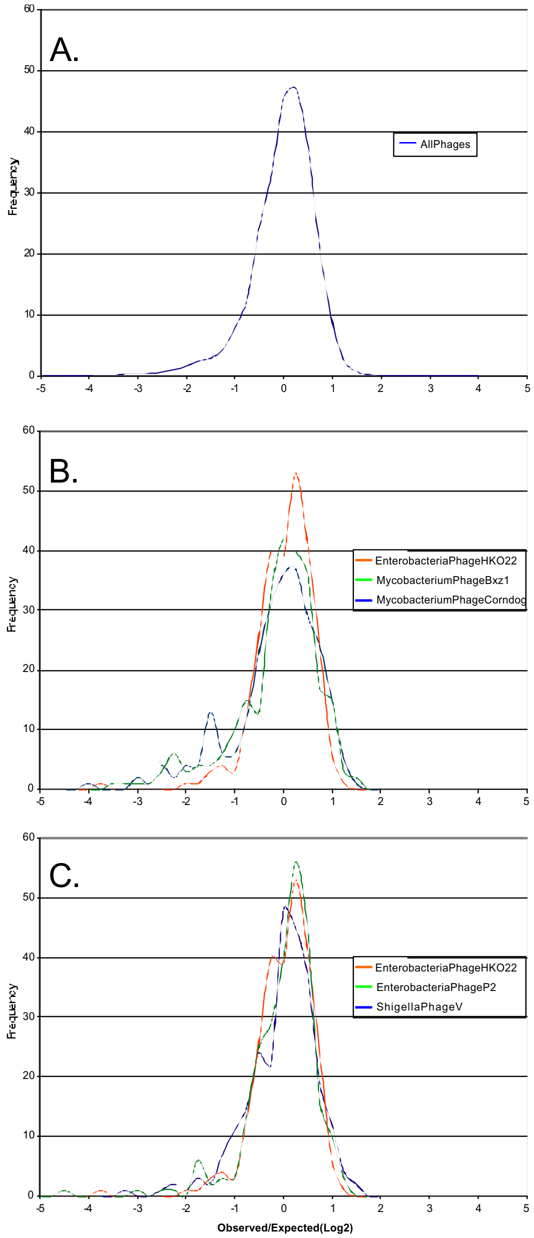

Figure 2
Frequency distribution of DNA tetranucleotide usage profiles of selected bacteriophages. The observed/expected TUD was determined for the 256 tetranucleotide combinations for each genome, as described in Materials and Methods. The resulting values were sorted within 0.25 intervals and the ordinate represents the number of tetranucleotide combinations within each interval. Panel A: Blue – All bacteriophages studied. Panel B: Red – Enterobacteria phage HK022 (Siphoviridae), Green – Enterobacteria phage P2 (Myoviridae), Blue – *Shigella* phage V (Podoviridae). Panel C: Red – Enterobacteria phage HK022, Green – *Mycobacterium* phage Bxz1 (Myoviridae), Blue – *Mycobacterium* phage Corndog (Siphoviridae).

hypothesized to be acquired through lateral gene transfer [23,24]. To determine whether bacteriophages have unique genomic signatures, we employed a method based on tetranucleotide usage deviations (TUD) from expected. TUD patterns have substantially more phylogenetic signal than codon usage biases when compared to 16S rRNA [19]. We generated TUD based on Zero-Order Markov algorithms, which determine how patterns of tetranucleotide usage in each genome deviate from those expected based on overall nucleotide content, representing the genomic signature [19,25]. Zero-order Markov algorithms were chosen, as removal of constituent oligonucleotide biases (e.g. dinucleotide and trinucleotide biases) through Markov chain analysis results in a substantial loss of TUD phylogenetic signal [19].

To determine whether TUD patterns are conserved, we measured TUD profiles across representative bacteriophage genomes, comparing each portion tested against the genome mean [24]. The patterns of tetranucleotide usage are relatively well-conserved across all bacteriophages studied [see Additional file 1], with examples provided for *Haemophilus* phage HP1 (Figure 1a), *Burkholderia* phage 781 (Figure 1b), and Enterobacteriaphage Mu (Figure 1c). The relatively small variation in patterns of nucleotide usage across phage genomes indicates that mean genome nucleotide usage patterns can be considered representative as a first-order approximation. TUD profiles of all phage genomes combined follows a normal distribution (Figure 2a), which further supports that tetranucleotide usage patterns are well-conserved across each bacteriophage studied.

We next sought to determine whether the TUD distribution across each bacteriophage genome is unique, representing a genomic signature. Highly similar yet distinct TUD profiles were shown for Enterobacteria phage P2 (Myoviridae), Enterobacteria phage HK022 (Siphoviridae), and *Shigella* phage V (Podoviridae) (Figure 2b). However, within a bacteriophage family, TUD patterns are not strictly conserved, as demonstrated by the lack of similarity between *Mycobacterium* phage Bxz1 and Enterobacteria phage HK022 (both Siphoviridae) (Figure 2c). These data are an indication that each phage genome has unique patterns of nucleotide usage that are not strictly determined by tail characteristics.

Comparison of tetranucleotide usage patterns between bacteriophage genomes

Comparison of TUD profiles by linear regression analysis reveals that phages with comparable host range have similar patterns. Enterobacteria phage HK022 is more closely related in TUD to Enterobacteria phage P2 (R^2 = 0.649; Figure 3a) than to *Mycobacterium* phage Bxz1 (R^2 = 0.067; Figure 3b). A parallel relationship is shown by comparing

Burkholderia cepaciae phage 1 with *Burkholderia cepaciae* phage 781 ($R^2 = 0.980$), and with "closely related" *Bordetella* phage BIP-1 ($R^2 = 0.561$). The two Enterobacteria phages (Figure 3a) share high level TUD similarity despite different tail morphology; results are parallel for *Streptococcus* phages ($R^2 = 0.559$; Figure 3c) and for *Mycobacterium* phages ($R^2 = 0.741$; Figure 3d). These representative examples show that TUD patterns may not be predicted by tail morphology, suggesting that family classification in bacteriophages is not phylogenetically robust.

Bacteriophage tetranucleotide phylogeny

TUD patterns in prokaryotes have phylogenetic content similar to that of 16S rRNA [19]. Since our initial analysis indicated that TUD patterns are shared among bacteriophages with similar host range, we examined TUD-based phage phylogenetic structure to determine host-range influence on evolutionary relationships. Phylogenies based on TUD patterns were generated assuming that phylogenetic structure would be predicted by host range, tail morphology, a combination of both, or essentially be random. Compared with randomly generated phylogenies, TUD phylogenies (Figure 4) show a distinct, non-random pattern (Figure 5). With few exceptions, of the 83 test phages, TUD phylogenies have the following structure: 1) phages that parasitize gram-negative Enterobacteria (including *Escherichia*, *Shigella*, *Klebsiella*, *Yersinia*, *Salmonella*, and *Vibrio*) cluster together; 2) phages related to gram-positive cocci (including *Staphyloccus*, *Streptococcus*, and *Lactococcus* phages) cluster together, 3) phages related to gram-negative non-enterobacteria (including *Burkholderia*, *Pseudomonas*, and *Bordetella* phages) cluster together, and 4) phages related to gram-positive bacilli (including *Bacillus*, *Lactobacillus*, *Mycobacterium*, and *Streptomyces*) do not cluster. The observed clustering does not associate with tail morphology.

Host-bacteriophage tetranucleotide phylogeny

Since TUD-based phage phylogeny appears closely associated with host range, we determined the phylogenetic structure of bacteriophages together with their host bacteria. Nucleotide usage patterns in both bacteriophages and prokaryotes represent genomic signatures that can be compared across organisms in genome size-independent analysis [19,20]. With few exceptions, phages cluster near their host organisms, as is demonstrated by the close relationships with the Enterobacteria, *Staphylococcus*, *Streptococcus*, *Lactococcus*, *Lactobacillus*, *Listeria*, *Bordetella*, *Pseudomonas*, *Burkholderia*, *Streptomyces*, and *Mycobacteria* (Figure 6). The phylogeny is consistent with a model of host-phage co-evolution, but is inconsistent with phage evolution based on tail morphology. Exceptions to the model include *Bacillus* phages PZA and GA-1, *Pseudomonas* phage Phi KMV, and several Enterobacteria phages that cluster independent of presumed host (Figure 6).

Eukaryotic virus TUD phylogeny

Eukaryotic virus taxonomy also has been based on morphologic characteristics [2]. Since phage TUD phylogeny demonstrates substantial host influences (Figure 6), we determined eukaryotic virus TUD-based phylogeny using representative species to determine phylogenetic structure. Based on TUD patterns, eukaryotic viruses are associated based on family, size, and type of genetic material (Figure 7). Important trends include: 1) the single stranded RNA viruses cluster together, with the exception of togaviruses and coronaviruses; 2) within the RNA viruses, picornoviruses cluster with the exception of foot and mouth virus, and paramyxoviruses cluster with the exception of RSV; 3) segmented RNA viruses including orthomyxoviruses and arenaviruses cluster; 4) small double stranded DNA viruses cluster, but are separate from large double stranded DNA viruses; 5) polyomaviruses cluster with retroviruses in the group of small double stranded DNA viruses; 6) large double stranded DNA viruses, including bacteriophages cluster; and 7) single stranded DNA viruses including parvoviruses cluster variably. In summary, these data suggest that phylogenetic signal exists in the virus TUD patterns, but to variable extents in different virus groups.

Gene content phylogeny of eukaryotic viruses

For prokaryotes, phylogeny based on gene content also approximates that of 16S rRNA [17]. We determined the phylogeny of the representative eukaryotic viruses based on gene content as an independent measure of virus evolution, hypothesizing that closely related viruses share more gene content [16,26]. The eukaryotic viruses tested cluster in a pattern similar to those based on TUD (Figure 7), but with important differences (Figure 8). DNA viruses cluster on the top portion of the derived phylogeny, retroviruses share a more central position, and RNA viruses cluster on the lower portion with the exception of the coronaviruses. Phylogeny based on gene content appears more robust than TUD phylogeny for the positive-sense RNA viruses, both large and small DNA viruses, single stranded DNA viruses, and retroviruses, which cluster separately from the small DNA viruses.

Discussion

As greater numbers of prokaryote and virus genomes are solved, genomic signatures have become better defined. Prokaryotes have genomic signatures that contain phylogenetic signal at both the dinucleotide [25] and tetranucleotide [19] levels. Our data indicate the existence of genomic signature in viruses parasitizing prokaryotic and eukaryotic hosts. Because genomic signature analysis is based on whole-genomes and is independent of multiple

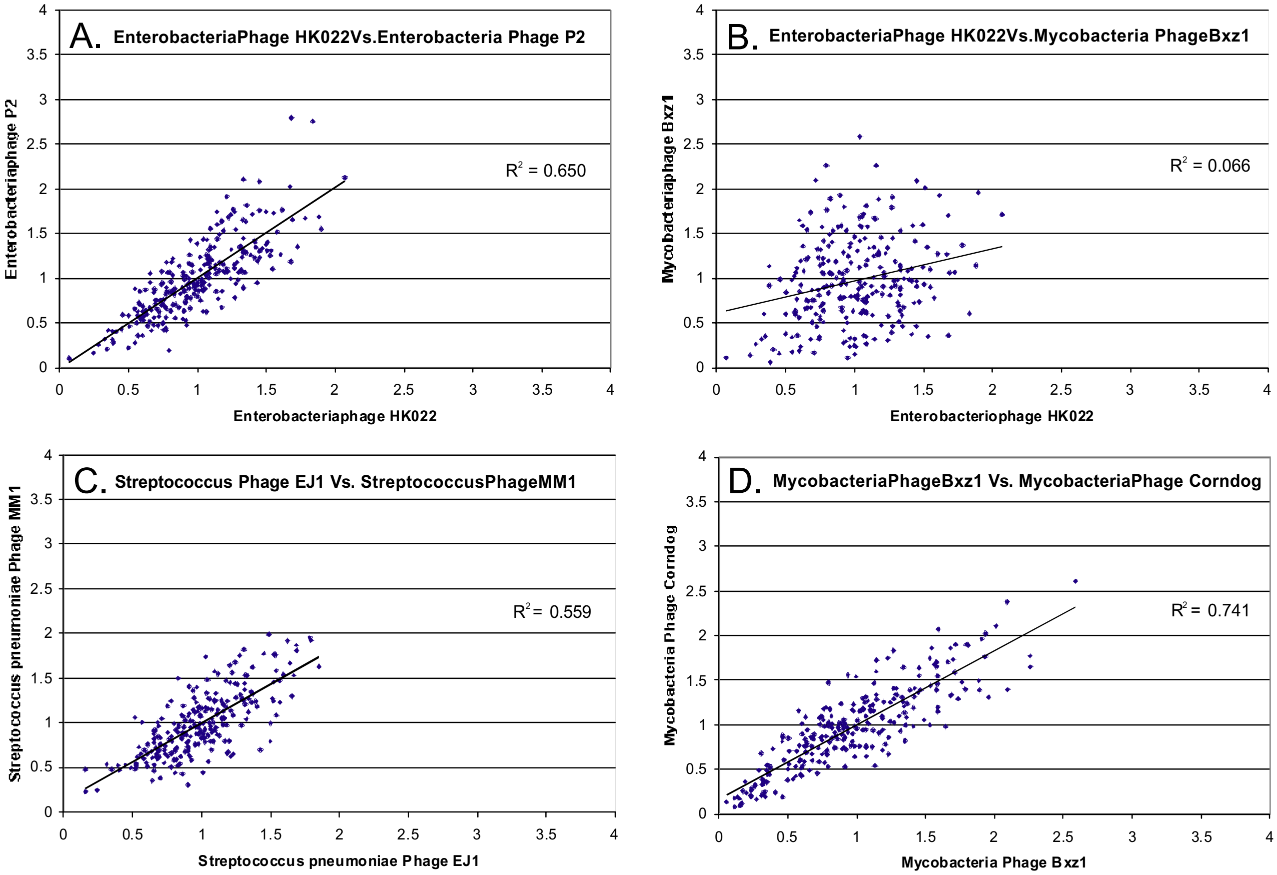

Figure 3

Linear regression analysis of DNA tetranucleotide usage profiles among selected genomes. Each of the 256 tetranucleotide combinations were determined for each genome as described in Materials and Methods, and the profiles compared by linear regression analysis. Panels A: Enterobacteria phage HK022 (Siphoviridae) vs. P2 (Myoviridae). Panel B: Enterobacteria phage HK022 vs. *Mycobacterium* phage Bxz1 (Myoviridae). Panel C: *Streptococcus pneumoniae* phage EJ1 (Myoviridae) vs. MM1 (Siphoviridae). Panel D: *Mycobacterium* phage Bxz1 vs. Corndog (Siphoviridae).

alignments, it provides a robust methodology for comparison across and between prokaryotes and viruses.

We show that bacteriophage genomic signatures are associated with their host organisms with few notable exceptions (Figure 6). This is most likely the effect of co-evolution between host and parasite, in which the host influences phage nucleotide usage patterns and possibly vice versa [25]. Indeed, both phage and host avoid use of certain tetranucleotides recognized by host restriction/ modification systems [19]. The close approximation, but not complete identity of phage TUD patterns to those of host organisms supports a co-evolution model (Figure 6). We hypothesize that the differences reflect limitations in the ability of phages to adapt to host TUD patterns, and/ or that phages need to retain particular TUD patterns to maintain host range.

An alternative explanation for bacteriophage TUD patterns is that their approximation to their hosts results from amelioration [27], and does not reflect recent common ancestry. Obligate parasitism likely necessitates that phages ameliorate to host nucleotide usage patterns, which is why host range may be critical. Bacteriophage phylogenetics based on shared gene content is dissimilar to that based on TUD, but not consistent with taxonomy based on morphologic features [26]. The differences between the methods likely reflect biases, with TUD phy-

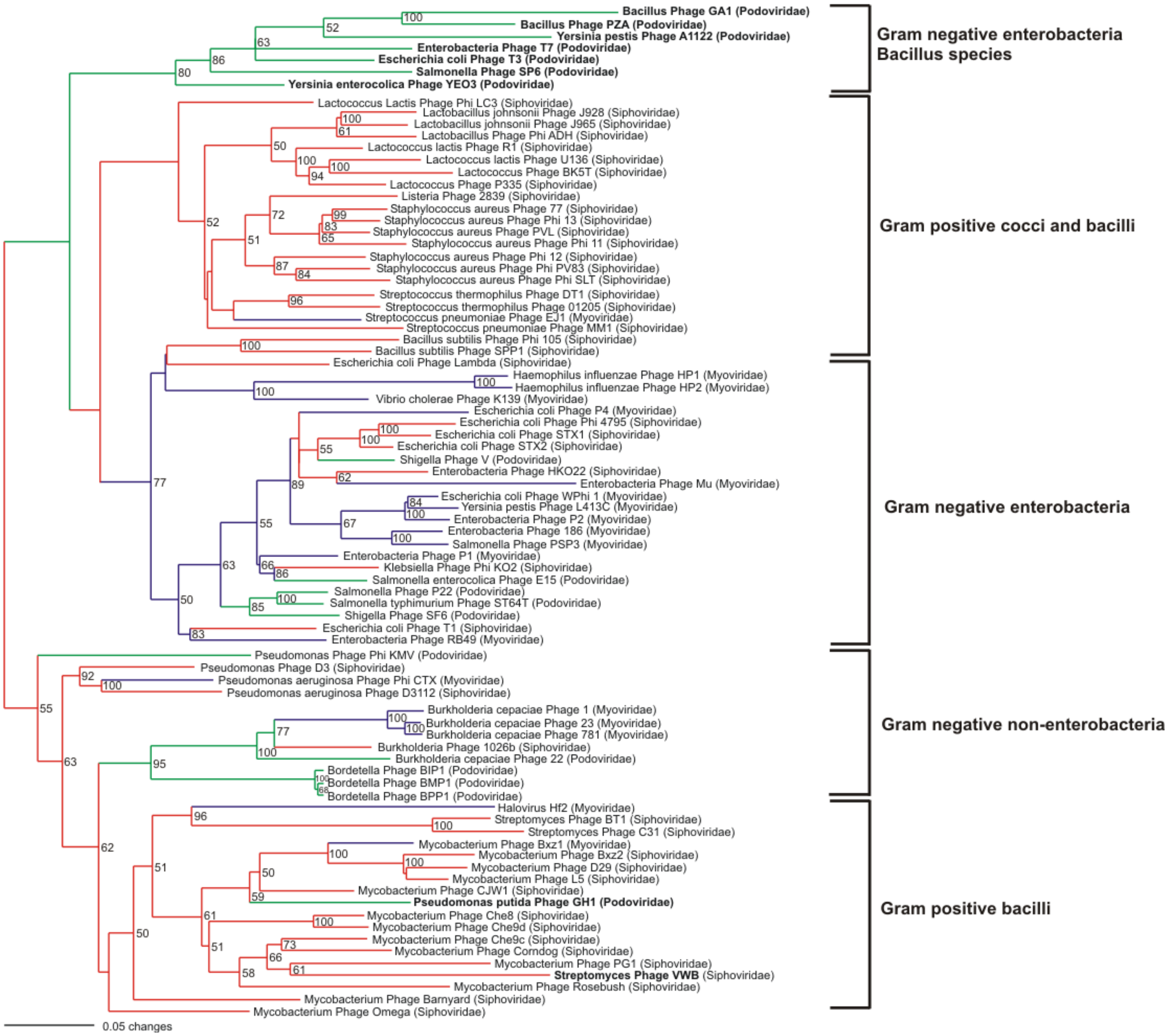
Bacillus Phage GA1 (Podoviridae)
100
Bacillus Phage PZA (Podoviridae)
52
Yersinia pestis Phage A1122 (Podoviridae)
63
Enterobacteria Phage T7 (Podoviridae)
86
Escherichia coli Phage T3 (Podoviridae)
80
Salmonella Phage SP6 (Podoviridae)
Yersinia enterocolica Phage YEO3 (Podoviridae)
Gram negative enterobacteria
Bacillus species
Lactococcus Lactis Phage Phi LC3 (Siphoviridae)
Lactobacillus johnsonii Phage J928 (Siphoviridae)
100
Lactobacillus johnsonii Phage J965 (Siphoviridae)
61
Lactobacillus Phage Phi ADH (Siphoviridae)
50
Lactococcus lactis Phage R1 (Siphoviridae)
100
Lactococcus lactis Phage U136 (Siphoviridae)
100
Lactococcus Phage BK5T (Siphoviridae)
94
Lactococcus Phage P335 (Siphoviridae)
Listeria Phage 2839 (Siphoviridae)
72
Staphylococcus aureus Phage 77 (Siphoviridae)
99
Staphylococcus aureus Phage Phi 13 (Siphoviridae)
83
52
Staphylococcus aureus Phage PVL (Siphoviridae)
65
51
Staphylococcus aureus Phage Phi 11 (Siphoviridae)
Gram positive cocci and bacilli
Staphylococcus aureus Phage Phi 12 (Siphoviridae)
87
Staphylococcus aureus Phage Phi PV83 (Siphoviridae)
84
Staphylococcus aureus Phage Phi SLT (Siphoviridae)
Streptococcus thermophilus Phage DT1 (Siphoviridae)
96
Streptococcus thermophilus Phage 01205 (Siphoviridae)
Streptococcus pneumoniae Phage EJ1 (Myoviridae)
Streptococcus pneumoniae Phage MM1 (Siphoviridae)
Bacillus subtilis Phage Phi 105 (Siphoviridae)
100
Bacillus subtilis Phage SPP1 (Siphoviridae)
Escherichia coli Phage Lambda (Siphoviridae)
Haemophilus influenzae Phage HP1 (Myoviridae)
100
Haemophilus influenzae Phage HP2 (Myoviridae)
100
Vibrio cholerae Phage K139 (Myoviridae)
Escherichia coli Phage P4 (Myoviridae)
Escherichia coli Phage Phi 4795 (Siphoviridae)
100
100
Escherichia coli Phage STX1 (Siphoviridae)
Escherichia coli Phage STX2 (Siphoviridae)
55
Shigella Phage V (Podoviridae)
Enterobacteria Phage HKO22 (Siphoviridae)
62
77
Enterobacteria Phage Mu (Myoviridae)
89
Escherichia coli Phage WPhi 1 (Myoviridae)
84
Yersinia pestis Phage L413C (Myoviridae)
100
Gram negative enterobacteria
Enterobacteria Phage P2 (Myoviridae)
55
67
Enterobacteria Phage 186 (Myoviridae)
100
Salmonella Phage PSP3 (Myoviridae)
Enterobacteria Phage P1 (Myoviridae)
66
Klebsiella Phage Phi KO2 (Siphoviridae)
63
86
Salmonella enterocolica Phage E15 (Podoviridae)
Salmonella Phage P22 (Podoviridae)
100
50
Salmonella typhimurium Phage ST64T (Podoviridae)
85
Shigella Phage SF6 (Podoviridae)
Escherichia coli Phage T1 (Siphoviridae)
83
Enterobacteria Phage RB49 (Myoviridae)
Pseudomonas Phage Phi KMV (Podoviridae)
Pseudomonas Phage D3 (Siphoviridae)
92
Pseudomonas aeruginosa Phage Phi CTX (Myoviridae)
100
Pseudomonas aeruginosa Phage D3112 (Siphoviridae)
55
Burkholderia cepaciae Phage 1 (Myoviridae)
100
Burkholderia cepaciae Phage 23 (Myoviridae)
100
77
Burkholderia cepaciae Phage 781 (Myoviridae)
Gram negative non-enterobacteria
Burkholderia Phage 1026b (Siphoviridae)
100
63
Burkholderia cepaciae Phage 22 (Podoviridae)
95
Bordetella Phage BIP1 (Podoviridae)
100
Bordetella Phage BMP1 (Podoviridae)
68
Bordetella Phage BPP1 (Podoviridae)
Halovirus Hf2 (Myoviridae)
96
Streptomyces Phage BT1 (Siphoviridae)
100
Streptomyces Phage C31 (Siphoviridae)
Mycobacterium Phage Bxz1 (Myoviridae)
100
Mycobacterium Phage Bxz2 (Siphoviridae)
62
100
Mycobacterium Phage D29 (Siphoviridae)
51
50
Mycobacterium Phage L5 (Siphoviridae)
Mycobacterium Phage CJW1 (Siphoviridae)
59
Pseudomonas putida Phage GH1 (Podoviridae)
Gram positive bacilli
61
Mycobacterium Phage Che8 (Siphoviridae)
100
Mycobacterium Phage Che9d (Siphoviridae)
50
Mycobacterium Phage Che9c (Siphoviridae)
73
51
Mycobacterium Phage Corndog (Siphoviridae)
66
Mycobacterium Phage PG1 (Siphoviridae)
61
58
Streptomyces Phage VWB (Siphoviridae)
Mycobacterium Phage Rosebush (Siphoviridae)
Mycobacterium Phage Barnyard (Siphoviridae)
Mycobacterium Phage Omega (Siphoviridae)
0.05 changes

Figure 4

Phylogram of 83 selected bacteriophages for which genomic sequences are available. The organisms were grouped by using distance matrices based on the sums of the differences from the other organisms for the 256 tetranucleotide combinations, as described in Materials and Methods. Phylogenies were created by neighbor-joining analysis. Colors indicate Myoviridae (contractile tails; blue), Podoviridae (short tail stubs; green), and Siphoviridae (long tails; red). Bootstrap values >50 based on 100 replicates are represented at each node, and the branch length index is indicated below the phylogeny. Bacteriophage source by host species – gram-positive bacilli, gram-positive cocci, gram-negative enterobacteria, and gram-negative non-enterobacteria are indicated by brackets.

logeny biased by amelioration, and gene content phylogeny biased by lateral gene transfer.

Horizontal gene transfer from host to phage produces apparent similarity to host nucleotide usage patterns that mechanistically does not represent recent common ancestry [28]. If horizontal transfer from host to phage served as the predominant mechanism of phage evolution, *Mycobacterium* phage TUD patterns would closely reflect their hosts. In the *Mycobacterium* phages, where horizontal acquisition is common [7], phage TUD patterns more closely approximate each other than their prokaryote hosts (Figure 6), which supports host-phage evolution in parallel, but not their evolution through horizontal gene

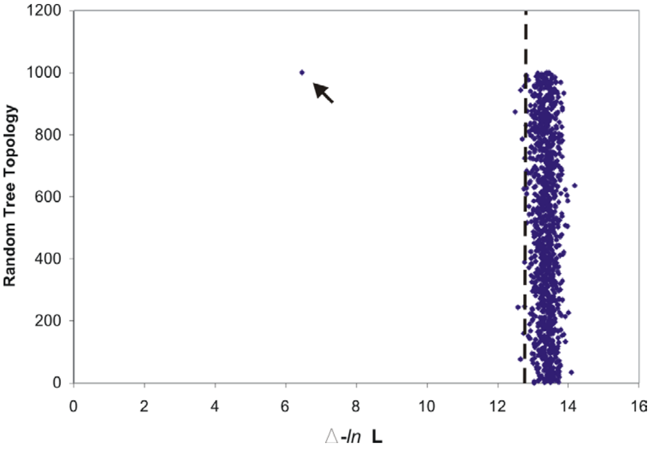

Figure 5
Likelihood analysis of phylogenetic congruence between the bacteriophage TUD phylogeny shown in Figure 4 and random trees. The 99th percentile of the likelihood differences between the TUD tree and the topologies from 1000 random trees is indicated by the vertical dashed line. The position of the TUD phylogeny (indicated by the arrow) is substantially outside of the null distribution.

transfer. That nucleotide usage patterns are relatively homogenous across each phage genome (Figure 1), suggests that the proportion of horizontally transferred genes in phage genomes may be relatively small or subject to rapid amelioration, and substantiates a parallel model of host-phage evolution. In phages with known transposons such as Enterobacteria phage Mu, these elements are not identified as anomalous in the phage genome (Figure 1c), suggesting that transposons are not responsible for the observed similarity in nucleotide usage patterns between host and phage.

Our data on bacteriophage phylogenies (Figure 5) directly oppose conceiving of morphological characteristics for understanding phage ancestry. Based on a TUD-based model of evolution, tail characteristics would have been horizontally transferred, subject to variable rates of evolution, or be continuously altered with little phylogenetic significance. Conversely, for a model of ancestry based on tail morphology to be correct, patterns of nucleotide usage would have to had shifted continuously throughout phage evolution independent of host. The substantial relationships between phage and host TUD (Figure 6) make independent evolution highly unlikely. Exceptions to the model of host-phage co-evolution presented by Podoviruses *Bacillus* phages PZA and GA-1, *Pseudomonas putida* phage GH-1, and a few Enterobacteria phages could be explained by broad host range, accelerated rates of change with loss of TUD phylogenetic signal, or alternate replication strategies. Each of these Enterobacteria phages is T7-like, with many known similarities at the genetic level [29], further suggesting their phylogenetic distinctness. The Bacillus phages are Phi29-like phages, which in addition to lack of shared ancestry with other Bacillus phages (Figure 4), have T7-like polymerases [30], supporting their recent ancestry with T7-like phages. Each of these phage groups has substantial strand bias [31], indicating their unique differences compared to other phage groups, and likely evolutionary distance.

TUD phylogenetics support co-evolution of host and bacteriophage, however, eukaryotic viruses cluster similar as expected based on recognized genetic and morphological features [1,2]. Bovine viruses (e.g. bovine RSV, papillomavirus, coronavirus, parvovirus, and polyomavirus) cluster independent of host, indicating that factors other than host influences determine their phylogenetic position (Figure 7). That coronaviruses do not belong to the major RNA virus cluster (Figures 7 and 8), contrary to previous studies [32], suggests that TUD may not be robust for certain RNA viruses. Most of the negative-sense RNA viruses cluster with the exception of RSV and the segmented RNA viruses. The clustering of RSV with rhinoviruses may represent convergent evolution or allelic exchange, as they occupy a similar ecological niche. The phylogenetic position of the segmented RNA viruses including orthomyxoviruses, arenaviruses, and reoviruses supports the concept of a common progenitor (Figure 7). The TUD analysis shows separate grouping of the large and small double stranded DNA viruses, which may be a limitation of the technique or reflect the occurrence of double stranded DNA more than once in the ancestral history of viruses. Bacteriophages cluster with the eukaryotic double stranded DNA viruses, further suggesting a single progenitor for the large double stranded DNA viruses. Whether the polyphylogeny observed among the large and small DNA viruses, the positive-sense RNA viruses, and the single stranded DNA viruses reflect methodologic limitations or evolutionarily significant phenomena remains to be determined. Phylogeny based on gene content, in which polyphylogeny among each group of viruses is diminished, supports the former hypothesis (Figure 8).

Conclusion

Morphological features form one basis for virus taxonomy, however we provide data that suggests bacteriophage tail characteristics may not sufficiently reflect their evolution. Based on TUD patterns, phages are co-evolving with their hosts in a manner defined by their ability to achieve broad host range. That there are only few exceptions to the co-evolution model concerning the many phages analyzed, substantiates that phylogenetic signal exists in phage TUD patterns. The TUD methodology is easily reproducible, alignment-independent, affected by lateral gene transfer in proportion to the extents of trans-

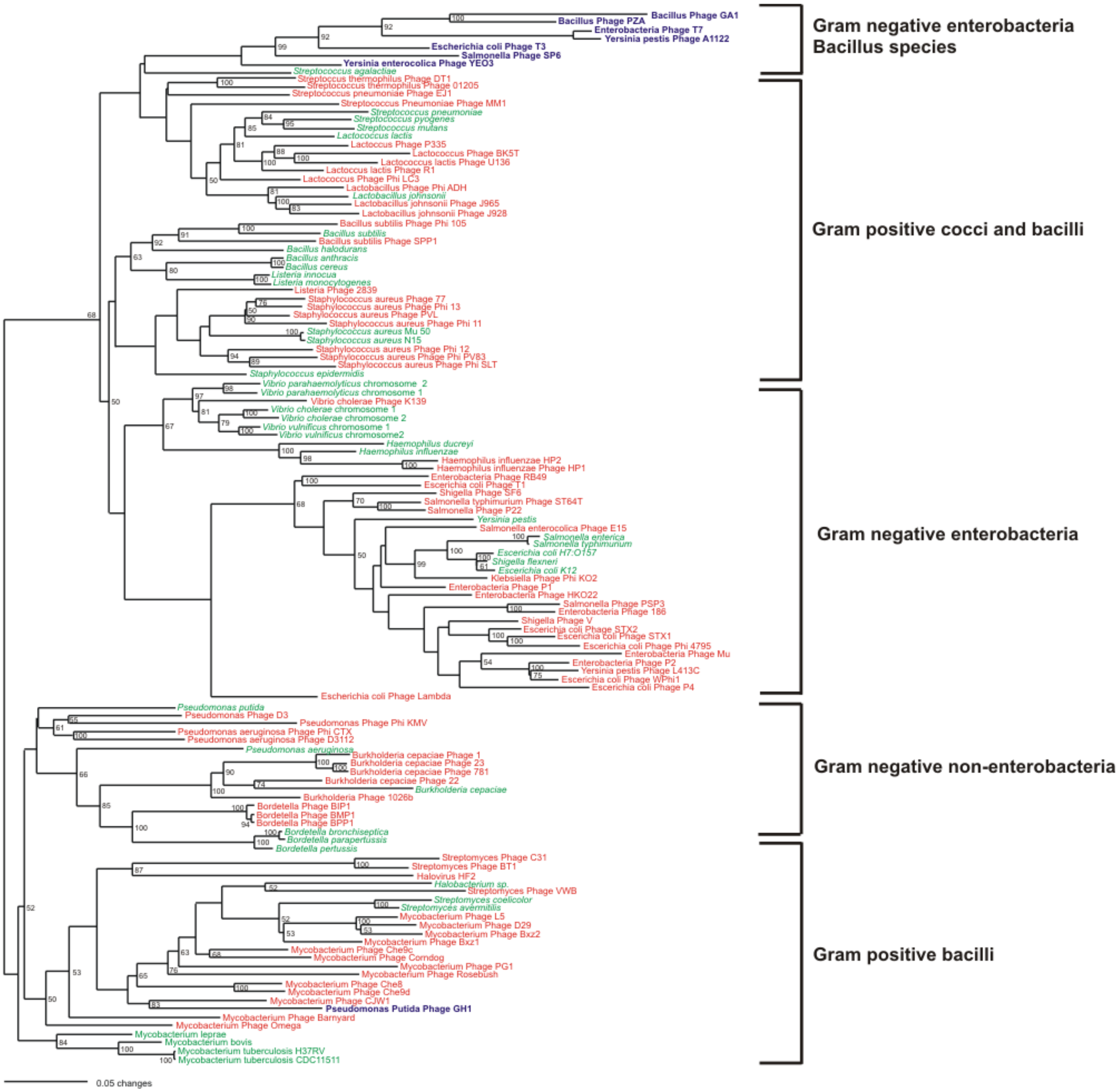
Bacillus Phage GA1
Bacillus Phage PZA
Enterobacteria Phage T7
Yersinia pestis Phage A1122
Escherichia coli Phage T3
Salmonella Phage SP6
Yersinia enterocolica Phage YEO3
Streptococcus agalactiae
Streptoccus thermophilus Phage DT1
Streptococcus thermophilus Phage 01205
Streptococcus pneumoniae Phage EJ1
Streptococcus Pneumoniae Phage MM1
Streptococcus pneumoniae
Streptococcus pyogenes
Streptococcus mutans
Lactococcus lactis
Lactoccus Phage P335
Lactococcus Phage BK5T
Lactococcus lactis Phage U136
Lactoccus lactis Phage R1
Lactococcus Phage Phi LC3
Lactobacillus Phage Phi ADH
Lactobacillus johnsonii
Lactobacillus johnsonii Phage J965
Lactobacillus johnsonii Phage J928
Bacillus subtilis Phage Phi 105
Bacillus subtilis
Bacillus subtilis Phage SPP1
Bacillus halodurans
Bacillus anthracis
Bacillus cereus
Listeria innocua
Listeria monocytogenes
Listeria Phage 2839
Staphylococcus aureus Phage 77
Staphylococcus aureus Phage Phi 13
Staphylococcus aureus Phage PVL
Staphylococcus aureus Phage Phi 11
Staphylococcus aureus Mu 50
Staphylococcus aureus N15
Staphylococcus aureus Phage Phi 12
Staphylococcus aureus Phage Phi PV83
Staphylococcus aureus Phage Phi SLT
Staphylococcus epidermidis
Vibrio parahaemolyticus chromosome 2
Vibrio parahaemolyticus chromosome 1
Vibrio cholerae Phage K139
Vibrio cholerae chromosome 1
Vibrio cholerae chromosome 2
Vibrio vulnificus chromosome 1
Vibrio vulnificus chromosome2
Haemophilus ducreyi
Haemophilus influenzae
Haemophilus influenzae HP2
Haemophilus influenzae Phage HP1
Enterobacteria Phage RB49
Escerichia coli Phage T1
Shigella Phage SF6
Salmonella typhimurium Phage ST64T
Salmonella Phage P22
Yersinia pestis
Salmonella enterocolica Phage E15
Salmonella enterica
Salmonella typhimurium
Escerichia coli H7:O157
Shigella flexneri
Escerichia coli K12
Klebsiella Phage Phi KO2
Enterobacteria Phage P1
Enterobacteria Phage HKO22
Salmonella Phage PSP3
Enterobacteria Phage 186
Shigella Phage V
Escerichia coli Phage STX2
Escerichia coli Phage STX1
Escerichia coli Phage Phi 4795
Enterobacteria Phage Mu
Enterobacteria Phage P2
Yersinia pestis Phage L413C
Escerichia coli Phage WPhi1
Escerichia coli Phage P4
Escherichia coli Phage Lambda
Pseudomonas putida
Pseudomonas Phage D3
Pseudomonas Phage Phi KMV
Pseudomonas aeruginosa Phage Phi CTX
Pseudomonas aeruginosa Phage D3112
Pseudomonas aeruginosa
Burkholderia cepaciae Phage 1
Burkholderia cepaciae Phage 23
Burkholderia cepaciae Phage 781
Burkholderia cepaciae Phage 22
Burkholderia cepaciae
Burkholderia Phage 1026b
Bordetella Phage BIP1
Bordetella Phage BMP1
Bordetella Phage BPP1
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Streptomyces Phage C31
Streptomyces Phage BT1
Halovirus HF2
Halobacterium sp.
Streptomyces Phage VWB
Streptomyces coelicolor
Streptomyces avermitilis
Mycobacterium Phage L5
Mycobacterium Phage D29
Mycobacterium Phage Bxz2
Mycobacterium Phage Bxz1
Mycobacterium Phage Che9c
Mycobacterium Phage Corndog
Mycobacterium Phage PG1
Mycobacterium Phage Rosebush
Mycobacterium Phage Che8
Mycobacterium Phage Che9d
Mycobacterium Phage CJW1
Pseudomonas Putida Phage GH1
Mycobacterium Phage Barnyard
Mycobacterium Phage Omega
Mycobacterium leprae
Mycobacterium bovis
Mycobacterium tuberculosis H37RV
Mycobacterium tuberculosis CDC11511
0.05 changes
Gram negative enterobacteria
Bacillus species
Gram positive cocci and bacilli
Gram negative enterobacteria
Gram negative non-enterobacteria
Gram positive bacilli

Figure 6

Phylogram of 39 selected bacteria and 83 selected bacteriophages for which genomic sequences are available. The organisms were grouped by using distance matrices based on the sums of the differences from the other organisms for the 256 tetranucleotide combinations, as described in Materials and Methods. Phylogenies were created by neighbor-joining analysis. Colors indicate bacteriophages (red), bacteria (green), and bacteriophages that are substantially beyond their presumed distribution (blue). Bootstrap values >50 based on 100 replicates are represented at each node, and branch length index is indicated below the phylogeny. Phages isolated in gram-positive bacilli, gram-positive cocci, gram-negative enterobacteria, and gram-negative non-enterobacteria are indicated by the brackets.

fer, and can be used for directly comparing both prokaryotes and viruses. Despite differences between TUD and morphology based classification, TUD phylogeny retains utility in understanding host-phage co-evolution and deviation in patterns of nucleotide usage in certain viruses since their divergence from recent common ancestors. As

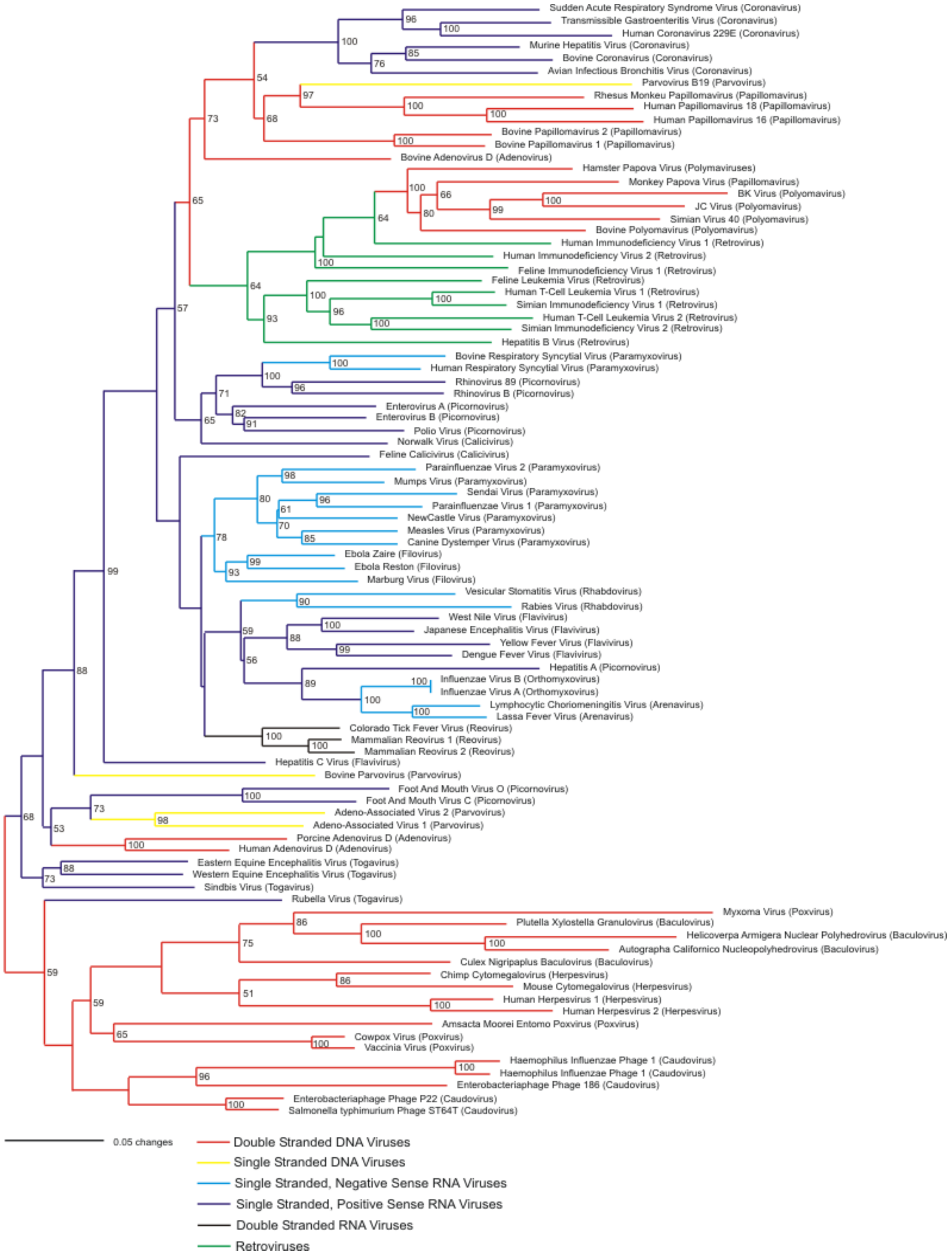
Sudden Acute Respiratory Syndrome Virus (Coronavirus)
Transmissible Gastroenteritis Virus (Coronavirus)
Human Coronavirus 229E (Coronavirus)
Murine Hepatitis Virus (Coronavirus)
Bovine Coronavirus (Coronavirus)
Avian Infectious Bronchitis Virus (Coronavirus)
Parvovirus B19 (Parvovirus)
Rhesus Monkeu Papillomavirus (Papillomavirus)
Human Papillomavirus 18 (Papillomavirus)
Human Papillomavirus 16 (Papillomavirus)
Bovine Papillomavirus 2 (Papillomavirus)
Bovine Papillomavirus 1 (Papillomavirus)
Bovine Adenovirus D (Adenovirus)
Hamster Papova Virus (Polymaviruses)
Monkey Papova Virus (Papillomavirus)
BK Virus (Polyomavirus)
JC Virus (Polyomavirus)
Simian Virus 40 (Polyomavirus)
Bovine Polyomavirus (Polyomavirus)
Human Immunodeficiency Virus 1 (Retrovirus)
Human Immunodeficiency Virus 2 (Retrovirus)
Feline Immunodeficiency Virus 1 (Retrovirus)
Feline Leukemia Virus (Retrovirus)
Human T-Cell Leukemia Virus 1 (Retrovirus)
Simian Immunodeficiency Virus 1 (Retrovirus)
Human T-Cell Leukemia Virus 2 (Retrovirus)
Simian Immunodeficiency Virus 2 (Retrovirus)
Hepatitis B Virus (Retrovirus)
Bovine Respiratory Syncytial Virus (Paramyxovirus)
Human Respiratory Syncytial Virus (Paramyxovirus)
Rhinovirus 89 (Picornovirus)
Rhinovirus B (Picornovirus)
Enterovirus A (Picornovirus)
Enterovirus B (Picornovirus)
Polio Virus (Picornovirus)
Norwalk Virus (Calicivirus)
Feline Calicivirus (Calicivirus)
Parainfluenzae Virus 2 (Paramyxovirus)
Mumps Virus (Paramyxovirus)
Sendai Virus (Paramyxovirus)
Parainfluenzae Virus 1 (Paramyxovirus)
NewCastle Virus (Paramyxovirus)
Measles Virus (Paramyxovirus)
Canine Dystemper Virus (Paramyxovirus)
Ebola Zaire (Filovirus)
Ebola Reston (Filovirus)
Marburg Virus (Filovirus)
Vesicular Stomatitis Virus (Rhabdovirus)
Rabies Virus (Rhabdovirus)
West Nile Virus (Flavivirus)
Japanese Encephalitis Virus (Flavivirus)
Yellow Fever Virus (Flavivirus)
Dengue Fever Virus (Flavivirus)
Hepatitis A (Picornovirus)
Influenzae Virus B (Orthomyxovirus)
Influenzae Virus A (Orthomyxovirus)
Lymphocytic Choriomeningitis Virus (Arenavirus)
Lassa Fever Virus (Arenavirus)
Colorado Tick Fever Virus (Reovirus)
Mammalian Reovirus 1 (Reovirus)
Mammalian Reovirus 2 (Reovirus)
Hepatitis C Virus (Flavivirus)
Bovine Parvovirus (Parvovirus)
Foot And Mouth Virus O (Picornovirus)
Foot And Mouth Virus C (Picornovirus)
Adeno-Associated Virus 2 (Parvovirus)
Adeno-Associated Virus 1 (Parvovirus)
Porcine Adenovirus D (Adenovirus)
Human Adenovirus D (Adenovirus)
Eastern Equine Encephalitis Virus (Togavirus)
Western Equine Encephalitis Virus (Togavirus)
Sindbis Virus (Togavirus)
Rubella Virus (Togavirus)
Myxoma Virus (Poxvirus)
Plutella Xylostella Granulovirus (Baculovirus)
Helicoverpa Armigera Nuclear Polyhedrovirus (Baculovirus)
Autographa Californico Nucleopolyhedrovirus (Baculovirus)
Culex Nigripaplus Baculovirus (Baculovirus)
Chimp Cytomegalovirus (Herpesvirus)
Mouse Cytomegalovirus (Herpesvirus)
Human Herpesvirus 1 (Herpesvirus)
Human Herpesvirus 2 (Herpesvirus)
Amsacta Moorei Entomo Poxvirus (Poxvirus)
Cowpox Virus (Poxvirus)
Vaccinia Virus (Poxvirus)
Haemophilus Influenzae Phage 1 (Caudovirus)
Haemophilus Influenzae Phage 1 (Caudovirus)
Enterobacteriaphage Phage 186 (Caudovirus)
Enterobacteriaphage Phage P22 (Caudovirus)
Salmonella typhimurium Phage ST64T (Caudovirus)
0.05 changes
Double Stranded DNA Viruses
Single Stranded DNA Viruses
Single Stranded, Negative Sense RNA Viruses
Single Stranded, Positive Sense RNA Viruses
Double Stranded RNA Viruses
Retroviruses

Figure 7

Phylogram of 90 selected viruses for which genomic sequences are available. The organisms were grouped by using distance matrices based on the sums differences from the other organisms for the 256 tetranucleotide combinations, as described in Materials and Methods. Phylogenies were created by neighbor-joining analysis. Colors indicate double stranded DNA viruses (red); single stranded DNA viruses (yellow); retroviruses (green); negative-sense single stranded RNA viruses (blue); positive-sense single stranded RNA viruses (sky blue); and double stranded RNA viruses (black). Bootstrap values >50 based on 100 replicates are represented at each node, and branch length index is indicated below the phylogeny.

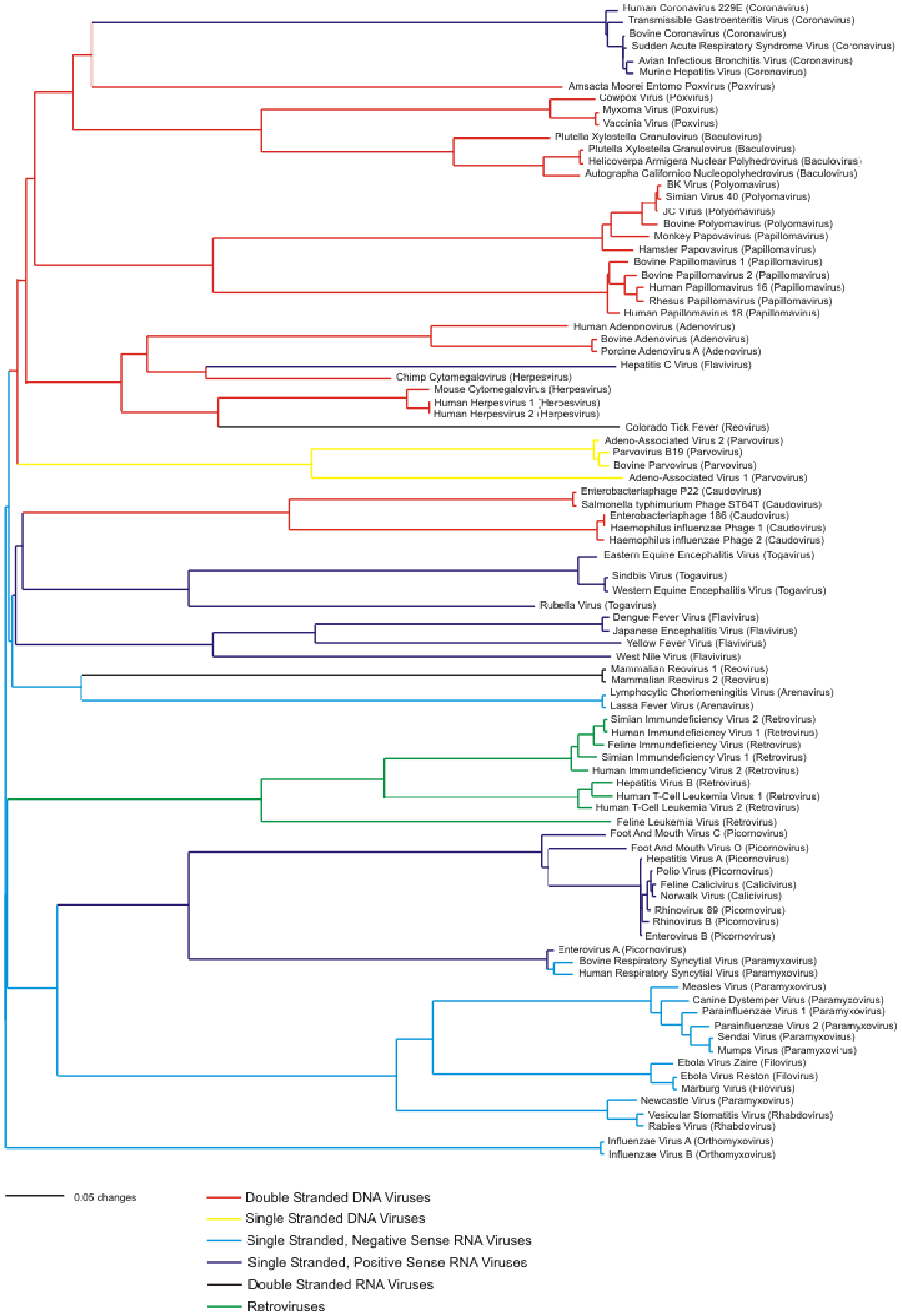
Human Coronavirus 229E (Coronavirus)
Transmissible Gastroenteritis Virus (Coronavirus)
Bovine Coronavirus (Coronavirus)
Sudden Acute Respiratory Syndrome Virus (Coronavirus)
Avian Infectious Bronchitis Virus (Coronavirus)
Murine Hepatitis Virus (Coronavirus)
Amsacta Moorei Entomo Poxvirus (Poxvirus)
Cowpox Virus (Poxvirus)
Myxoma Virus (Poxvirus)
Vaccinia Virus (Poxvirus)
Plutella Xylostella Granulovirus (Baculovirus)
Plutella Xylostella Granulovirus (Baculovirus)
Helicoverpa Armigera Nuclear Polyhedrovirus (Baculovirus)
Autographa Californico Nucleopolyhedrovirus (Baculovirus)
BK Virus (Polyomavirus)
Simian Virus 40 (Polyomavirus)
JC Virus (Polyomavirus)
Bovine Polyomavirus (Polyomavirus)
Monkey Papovavirus (Papillomavirus)
Hamster Papovavirus (Papillomavirus)
Bovine Papillomavirus 1 (Papillomavirus)
Bovine Papillomavirus 2 (Papillomavirus)
Human Papillomavirus 16 (Papillomavirus)
Rhesus Papillomavirus (Papillomavirus)
Human Papillomavirus 18 (Papillomavirus)
Human Adenonovirus (Adenovirus)
Bovine Adenovirus (Adenovirus)
Porcine Adenovirus A (Adenovirus)
Hepatitis C Virus (Flavivirus)
Chimp Cytomegalovirus (Herpesvirus)
Mouse Cytomegalovirus (Herpesvirus)
Human Herpesvirus 1 (Herpesvirus)
Human Herpesvirus 2 (Herpesvirus)
Colorado Tick Fever (Reovirus)
Adeno-Associated Virus 2 (Parvovirus)
Parvovirus B19 (Parvovirus)
Bovine Parvovirus (Parvovirus)
Adeno-Associated Virus 1 (Parvovirus)
Enterobacteriaphage P22 (Caudovirus)
Salmonella typhimurium Phage ST64T (Caudovirus)
Enterobacteriaphage 186 (Caudovirus)
Haemophilus influenzae Phage 1 (Caudovirus)
Haemophilus influenzae Phage 2 (Caudovirus)
Eastern Equine Encephalitis Virus (Togavirus)
Sindbis Virus (Togavirus)
Western Equine Encephalitis Virus (Togavirus)
Rubella Virus (Togavirus)
Dengue Fever Virus (Flavivirus)
Japanese Encephalitis Virus (Flavivirus)
Yellow Fever Virus (Flavivirus)
West Nile Virus (Flavivirus)
Mammalian Reovirus 1 (Reovirus)
Mammalian Reovirus 2 (Reovirus)
Lymphocytic Choriomeningitis Virus (Arenavirus)
Lassa Fever Virus (Arenavirus)
Simian Immundeficiency Virus 2 (Retrovirus)
Human Immundeficiency Virus 1 (Retrovirus)
Feline Immundeficiency Virus (Retrovirus)
Simian Immundeficiency Virus 1 (Retrovirus)
Human Immundeficiency Virus 2 (Retrovirus)
Hepatitis Virus B (Retrovirus)
Human T-Cell Leukemia Virus 1 (Retrovirus)
Human T-Cell Leukemia Virus 2 (Retrovirus)
Feline Leukemia Virus (Retrovirus)
Foot And Mouth Virus C (Picornovirus)
Foot And Mouth Virus O (Picornovirus)
Hepatitis Virus A (Picornovirus)
Polio Virus (Picornovirus)
Feline Calicivirus (Calicivirus)
Norwalk Virus (Calicivirus)
Rhinovirus 89 (Picornovirus)
Rhinovirus B (Picornovirus)
Enterovirus B (Picornovirus)
Enterovirus A (Picornovirus)
Bovine Respiratory Syncytial Virus (Paramyxovirus)
Human Respiratory Syncytial Virus (Paramyxovirus)
Measles Virus (Paramyxovirus)
Canine Dystemper Virus (Paramyxovirus)
Parainfluenzae Virus 1 (Paramyxovirus)
Parainfluenzae Virus 2 (Paramyxovirus)
Sendai Virus (Paramyxovirus)
Mumps Virus (Paramyxovirus)
Ebola Virus Zaire (Filovirus)
Ebola Virus Reston (Filovirus)
Marburg Virus (Filovirus)
Newcastle Virus (Paramyxovirus)
Vesicular Stomatitis Virus (Rhabdovirus)
Rabies Virus (Rhabdovirus)
Influenzae Virus A (Orthomyxovirus)
Influenzae Virus B (Orthomyxovirus)
0.05 changes
Double Stranded DNA Viruses
Single Stranded DNA Viruses
Single Stranded, Negative Sense RNA Viruses
Single Stranded, Positive Sense RNA Viruses
Double Stranded RNA Viruses
Retroviruses

Figure 8

Phylograms of 90 selected viruses for which genomic sequences are available. The organisms were grouped according to distance matrices based on the number of shared orthologues between each genome, as described in Materials and Methods. Phylogenies were created by neighbor-joining analysis. Color code as is for Figure 7. The branch length index is indicated below the phylogeny.

such, TUD phylogeny should be considered complementary to other systems for analysis of virus evolution.

Methods

Virus, phage, and microbial genomes

Complete genome sequences of the phages [see Additional file 1], viruses [see Additional file 2], and bacteria studied were obtained from Genbank [33,34].

Analysis of tetranucleotides

Tetranucleotides were selected for study because analysis of higher-order oligonucleotides was not possible given the limitation in virus genome sizes. We based our minimum genome sequence length on the assumption that 95% of tetranucleotide combinations should occur at least 10 times [31]. Our calculated minimum length was 5 kb based on analysis of concatenated genome strands designed to eliminate strand bias. The minimum genome length analyzed in this study was 4.7 kb (9.4 kb when analyzing both strands), while analysis of pentanucleotides would have required a minimum genome length of 20 kb. To determine the tetranucleotide usage deviations from expected (TUD) among prokaryotic genomes, a Zero-Order Markov algorithm [35] was used, which involves determining the expected number of tetranucleotides by removing biases in mononucleotide frequencies, as is determined by the equation: $E(W) = [(A^a * C^c * G^g * T^t) * N]$, where A, C, G, and T represent the frequency of the four nucleotides within the window being evaluated, respectively, a, c, g, and t represent the number of nucleotides A, C, G, and T in each tetranucleotide, respectively, and N represents the length of the genome being evaluated. The frequency of divergence for each tetranucleotide is expressed as the ratio of observed to expected, and the TUD profiles for all tetranucleotides determined for each organism studied using Swaap 1.0.1 [36], and their relative abundance between genomes compared by linear regression analysis.

Tetranucleotide difference index

Tetranucleotide differences in each genome were determined using Markov chain analysis [37], by determining the expected oligonucleotide word frequency by removing the biases existing in component oligonucleotides. $W = (w_1 w_2 ... w_m)$ denotes the word formed by the concatenation of m nucleotides, and $N(W)$ is its observed count in a sequence of length n, as described [24]. The expected count $E(W)$ of W is:

$$E(W) = N(w_1 w_2 ... w_{m-1}) N(w_2 w_3 ... w_m) / N(w_2 w_3 ... w_{m-1}).$$

The frequency of the word $F(W)$ is expressed as the ratio of the observed $O(W)$ to the expected $E(W)$. $F(W)$ can be determined for all windows $F_w(W)$ of specified size within each genome. Tetranucleotide differences for each window are measured by the expression: $\sum_j^i |F(W) - F_w(W)|$, where i - j represent all tetranucleotide combinations. The tetranucleotide difference index represents the difference between each window and the mean difference for all windows. Z-scores were determined for each window using the equation $Z = (x - \mu)/\sigma$, where x represents the tetranucleotide difference for the window being evaluated, μ represents the mean tetranucleotide differences for all windows, and σ represents the standard deviation of all windows. Those windows differing from the mean by ± 3 Z-scores were defined as significant, and were determined using Swaap 1.0.1 [36].

Phylogenetic analysis

Distances based on TUD were determined: $D_t = 1/4^4 * \sum |F_1(W) - F_2(W)|$, where $F_1(W)$and $F_2(W)$ represent $F(W)$ for each of the 256 tetranucleotides for any organisms 1 and 2 [19,38], which represents the Euclidean distance between 2 vectors in 256 space. Bootstraping was performed by sampling with replacement of each of the 256 tetranucleotide frequencies using Swaap PH 1.0.1 [39], and phylograms created based on distance matrices using Phylip 3.5 [40], reviewed via Treeview [41], and displayed using midpoint rooting with Paup 4.0b10 [42].

Phylogeny based on gene content was determined using methodology similar to that described [17]. Shared genes between genomes were determined using an operational definition of orthology. Briefly, all genes within each genome examined were added to a BLAST database, and each genome then was compared at the amino acid level against the database using a BLAST threshold value, E = 0.01. The resulting number of orthologues were tabulated for each pair of viruses compared, and distances were expressed as one minus the percent of shared genes between each genome. Data tabulation and distance matrices were generated using Swaap PH 1.0.1 [39]. Phylogenies were constructed based on distance matrices using Paup 4.0b10 [42].

Analysis of congruence among phylogenetic trees

Analysis of congruence among the gene phylograms was performed on consensus trees, and 1000 trees created with random topology. Differences in log likelihood (Δ-*ln* L) were computed between phylograms based on TUD phylogenies and 1000 random trees. Differences in Δ-*ln* L for random phylograms can be considered as the null distribution, obtained when there is no more similarity in topology than expected by chance. If the Δ-*ln* L values for comparisons among the phylograms are within the 99th percentile of the null distribution, then the topologies are significantly different, and thus incongruent [43].

List of Abbreviations

TUD – Tetranucleotide usage deviations from expected

Authors' contributions

DP – conceived of study, carried out genome analysis, revised software, and drafted manuscript

TW – participated in study design, data analysis, and manuscript revision

CG – participated in study design, literature review, data analysis, and manuscript revision

MB – participated in study design and manuscript revision

All authors read and approved the final manuscript

Acknowledgements

Supported in part by the Foundation for Bacteriology, and the National Institutes of Health (RO1GM63270). We thank Dr. Douglas Brutlag for his critical review of this manuscript.

References

1. Maniloff J, Ackermann H-W: **Taxonomy of bacterial viruses: establishment of tailed virus genera and the order *Caudovirales*.** *Arch Virol* 1998, **143:**2051-2063.
2. Van Regenmortel MHV, Fauquet CM, Bishop DHL, Carstens E, Estes M, Lemon S, Maniloff J, Mayo MA, McGeoch D, Pringle CR, Wickner WB: **Virus taxonomy. Seventh report of the International Committee on Taxonomy of Viruses.** Academic Press, New York; 2000.
3. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF: **Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage.** *Proc Natl Acad Sci USA* 1999, **96:**2192-2197.
4. Tetart F, Desplats C, Krisch HM: **Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: recombination between conserved motifs swaps adhesin specificity.** *J Mol Bio* 1998, **282:**543-556.
5. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW: **Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosiacism in the lamboid bacteriophages.** *J Mol Biol* 2000, **299:**27-51.
6. Nilsson AS, Haggard-Ljungquist E: **Detection of homologous recombination among bacteriophage P2 relatives.** *Mol Phylogenet Evol* 2001, **21:**259-269.
7. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR, Hendrix RW, Hatfull GF: **Origins of highly mosaic Mycobacteriophage genomes.** *Cell* 2003, **113:**171-182.
8. Li W-H, Graur D: **Fundamentals of Molecular Evolution.** Sinauer Associates, Sunderland, MA; 1991.
9. Brussow H, Desiere F: **Comparative phage genomics and the evolution of *Siphoviridae* : insights into dairy phages.** *Mol Microbiol* 2001, **39:**213-222.
10. Ackermann H-W: **Bacteriophage observations and evolution.** *Res Microbiol* 2003, **154:**245-251.
11. Steinbacher S, Seckler R, Miller S, Steipe B, Huber R, Reinemer P: **Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer.** *Science* 1994, **265:**383-386.
12. Lawrence JG, Hatful GF, Hendrix RW: **Imbroglios of viral taxonomy: genetic exchange and failure of phenetic approaches.** *J Bacteriol* 2002, **184:**4891-4905.
13. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74:**5088-5090.
14. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eukarya.** *Proc Natl Acad Sci USA* 1990, **87:**4576-4579.
15. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284:**2124-2128.
16. Fitz-Gibbon ST, House CH: **Whole-genome based phylogenetic analysis of free-living microorganisms.** *Nucleic Acids Res* 1999, **27:**4218-4222.
17. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nature Genetics* 1999, **21:**108-110.
18. House CH, Fitz-Gibbon ST: **Using homolog groups to create a whole-genomic tree of free-living organisms: an update.** *J Mol Evol* 2002, **54:**539-547.
19. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ: **Evolutionary implications of nucleotide usage patterns in prokaryotes.** *Genome Research* 2003, **13:**145-155.
20. Karlin S, Mrazek J, Campbell AM: **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriol* 1997, **179:**3899-913.
21. Karlin S, Campbell AM, Mrazek J: **Comparative DNA analysis across diverse genomes.** *Annu Rev Genet* 1998, **32:**185-225.
22. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T: **Informatics for unveiling hidden genomic signatures.** *Genome Research* 2003, **13:**693-702.
23. Karlin S: **Detecting anomalous gene clusters and pathogenecity islands in diverse bacterial genomes.** *Trends Microbiol* 2001, **9:**335-343.
24. Pride DT, Blaser MJ: **Identification of horizontally acquired genetic elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis.** *Genome Letters* 2002, **1:**2-15.
25. Burge C, Campbell AM, Karlin S: **Over- and under-representation of short oligonucleotides in DNA sequences.** *Proc Natl Acad Sci USA* 1992, **89:**1358-1362.
26. Rohwer F, Edwards R: **The phage proteomic tree: a genome-based taxonomy for phage.** *J Bacteriol* 2002, **184:**4529-4535.
27. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44:**383-397.
28. Nilsson AS, Karlsson JL, Haggard-Ljungquist E: **Site-specific recombination links the evolution of P2-like coliphages and pathogenic enterobacteria.** *Mol Biol Evol* 2004, **21:**1-13.
29. Garcia E, Elliot JM, Ramanculov E, Chain PSG, Chu CC, Molineux IJ: **The genome sequence of *Yersinia pestis* bacteriophage PhiA1122 reveals an intimate history with the coliphage T3 and T7 genomes.** *J Bacteriol* 2003, **185:**5248-5262.
30. Kamtekar S, Berman AJ, Wang J, Lazaro JM, de Vega M, Blanco L, Salas M, Steitz TA: **Insights into strand displacement and processivity from the crystal structure of the protein-primed DNA polymerase of bacteriophage Phi29.** *Molecular Cell* 2004, **16:**609-618.
31. Reva ON, Tummler B: **Global features of sequences of bacterial chromosomes, plasmids, and phages revealed by analysis of oligonucleotide usage patterns.** *BMC Bioinformatics* 2004, **5:**90.
32. Yap YL, Zhang XW, Danchin A: **Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling.** *BMC Bioinformatics* 2003, **4:**43.
33. [http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html].
34. [http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html].
35. Almagor H: **A Markov analysis of DNA sequences.** *J Theor Biol* 1983, **104:**633-645.
36. Pride DT: **Swaap 1.0.1: a tool for analyzing substitutions and similarity in multiple alignments.** 2004 [http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm].
37. Schbath S, Prum B, de Turckheim E: **Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences.** *J Comp Biol* 1995, **2:**417-437.
38. Cardon LR, Karlin S: **Computational DNA Sequence Analysis.** *Annu Rev Microbiol* 1994, **48:**619-654.
39. Pride DT: **Swaap PH 1.0.1: a tool for analyzing nucleotide usage patterns in coding and noncoding portions of microbial genomes.** 2004 [http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm].
40. Felsenstein J: **PHYLIP – Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
41. Page RDM: **TREEVIEW: an application to display phylogenetic trees on personal computers.** *Comp Appl Biosci* 1996, **12:**357-458.

42. Swofford DL: **Paup 4.0b10. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.** Sinauer Associates, Sunderland, Massachusetts; 1998.
43. Feil EJ, Holmes EC, Bessen DE, Chan M-S, Day NPJ, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG: **Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences.** *Proc Natl Acad Sci USA* 2001, **98:**182-187.