RESEARCH ARTICLE

REVISED Tetranucleotide usage highlights genomic heterogeneity among mycobacteriophages [version 2; referees: 2 approved]

Benjamin Siranosian[1,2], Sudheesha Perera[2], Edward Williams[2], Chen Ye[2], Christopher de Graffenried[3], Peter Shank[3]

[1]Center for Computational Molecular Biology, Brown University, Providence, RI, 02912, USA
[2]Division of Biology and Medicine, Brown University, Providence, RI, 02912, USA
[3]Department of Molecular Microbiology and Immunology, Brown University, Providence, RI, 02912, USA

v2 **First published:** 04 Feb 2015, **4**:36 (doi: 10.12688/f1000research.6077.1)
Latest published: 30 Oct 2015, **4**:36 (doi: 10.12688/f1000research.6077.2)

Abstract

Background

The genomic sequences of mycobacteriophages, phages infecting mycobacterial hosts, are diverse and mosaic. Mycobacteriophages often share little nucleotide similarity, but most of them have been grouped into lettered clusters and further into subclusters. Traditionally, mycobacteriophage genomes are analyzed based on sequence alignment or knowledge of gene content. However, these approaches are computationally expensive and can be ineffective for significantly diverged sequences. As an alternative to alignment-based genome analysis, we evaluated tetranucleotide usage in mycobacteriophage genomes. These methods make it easier to characterize features of the mycobacteriophage population at many scales.

Description

We computed tetranucleotide usage deviation (TUD), the ratio of observed counts of 4-mers in a genome to the expected count under a null model. TUD values are comparable between members of a phage subcluster and distinct between subclusters. With few exceptions, neighbor joining phylogenetic trees and hierarchical clustering dendrograms constructed using TUD values place phages in a monophyletic clade with members of the same subcluster. Regions in a genome with exceptional TUD values can point to interesting features of genomic architecture. Finally, we found that subcluster B3 mycobacteriophages contain significantly overrepresented 4-mers and 6-mers that are atypical of phage genomes.

Conclusions

Statistics based on tetranucleotide usage support established clustering of mycobacteriophages and can uncover interesting relationships within and between sequenced phage genomes. These methods are efficient to compute and do not require sequence alignment or knowledge of gene content. The code to download mycobacteriophage genome sequences and reproduce our analysis is freely available at https://github.com/bsiranosian/tango_final.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 30 Oct 2015	✓ report	✓ report
version 1 published 04 Feb 2015	✗ report	

1 **Oliver Bonham-Carter**, University of Nebraska USA

2 **David Martin**, University of Dundee UK

Discuss this article

Comments (0)

Corresponding author: Benjamin Siranosian (benjamin_siranosian@alumni.brown.edu)

How to cite this article: Siranosian B, Perera S, Williams E *et al.* **Tetranucleotide usage highlights genomic heterogeneity among mycobacteriophages [version 2; referees: 2 approved]** *F1000Research* 2015, **4**:36 (doi: 10.12688/f1000research.6077.2)

Grant information: This work was funded by Brown University Biology Undergraduate Education and the HHMI SEA-PHAGES program.

Competing interests: No competing interests were disclosed.

First published: 04 Feb 2015, **4**:36 (doi: 10.12688/f1000research.6077.1)

REVISED **Amendments from Version 1**

This version addresses the review by Dr. Bonham-Carter. Changes have been made to make the methods section more clear, and I have included an example figure to show the calculation of TUD on a small sequence. The results from the paper remain unchanged.

See referee reports

Introduction

Mycobacteriophages, phages infecting mycobacterial hosts, are a subset of the estimated 10^{31} phage particles present globally. Mycobacteriophages infect a number of bacterial hosts from the genus *Mycobacterium*, and they are broadly classified into *Siphoviridae* and *Myoviridae*. Mycobacteriophages are present in both land and aquatic environments and play a large ecological role in the turnover and evolution of bacteria (Bohannan & Lenski, 2000; Chibani-Chennoufi *et al.*, 2004; Hendrix, 2002). The recent rise of antimicrobial-resistant pathogenic bacteria has renewed interest in mycobacteriophages and the potential for phage therapy of *Mycobacterium tuberculosis* infections. Although *in vivo* experiments have not yet yielded promising clinical results, mycobacteriophages are still powerful diagnostic tools for the investigation of mycobacterial pathogenesis (Danelishvili *et al.*, 2006; Hatfull, 2014; McNerney, 1999).

The genomic sequences of mycobacteriophages are mosaic and diverse. As of April 2014, 663 distinct mycobacteriophage genomes were available on the database PhagesDB.org; most were isolated on *Mycobacterium smegmatis* MC^2155. Global Guanine + Cytosine (GC) content ranges from 50.3% to 70% (mean of 63.9%), and genome lengths range from 41kb to 165kb (mean of 67kb). In total, more than 50,000 distinct genes are found within the population. The majority of these genes are of unknown function and do not have homologs in other types of phages or bacteria (Hatfull *et al.*, 2010). However, many genes are shared between closely related mycobacteriophages. Similar genes have been grouped into almost 4,000 phamilies (or phams, a play on gene families) based on shared amino acid sequence. Phams have been used to investigate horizontal gene transfer within the mycobacteriophage population and to create phylogenetic trees.

Despite the high levels of diversity, mycobacteriophages can be grouped into distinct clusters based on their morphologic and genetic features. Some clusters are large and further divided into subclusters (cluster A, for example, with 11 subclusters and 246 members), while other are small and undivided (cluster S with two members and no subclusters). Some phages have no nearest neighbor to establish a cluster and are classified as singletons. Clusters are defined using four methods: dot-plot comparisons, pairwise average nucleotide identities, pairwise genome map comparisons and gene content analysis (Hatfull *et al.*, 2010). However, it should be noted that the clustering scheme proposed for mycobacteriophages mainly serves to identify similarities in genome architecture. This clustering scheme, and our proposed methods of grouping based on tetranucleotide usage described below, are not true taxonomic representations of the mycobacteriophage population. Extensive horizontal gene transfer prevents accurate reconstruction of evolutionary history from purely phylogenetic information (Lawrence *et al.*, 2002).

Methods traditionally used to analyze mycobacteriophage genomes require sequence alignment or genome annotation. These analytical tasks can be effective, but they are not without drawbacks. Alignment-based methods can be biased by the choice of score parameters (Frith *et al.*, 2010), and genome annotation may require significant manual input, including by-hand verification of automated gene calls before a mycobacteriophage genome is submitted to GenBank. It is especially difficult to build multiple-sequence alignment based phylogenetic trees from mycobacteriophage genomes because phages lack a common genetic element, such as 16S rRNA in bacteria (Doolittle, 1999). Alignment-free methods avoid many of the disadvantages associated with alignment-based inference. These methods typically use statistics based on the oligonucleotide composition of a sequence and are completely independent of alignment or annotation. Several methods have been developed for different applications; most are covered in the excellent review by Vinga (2007). Alignment-free methods are also less computationally intensive than multiple sequence alignment. While the complexity of sequence alignment algorithms scales at least as fast as the square of the number of sequences (at least $O(n^2)$ complexity), alignment free methods typically fall below $O(n^2)$ (Chan & Ragan, 2013).

Even so, there are drawbacks to alignment-free methods for analyzing genomes, mostly related to the interpretation of statistics in an evolutionary context. It can be difficult to understand how oligonucleotide frequencies are modified in a population over time when selection usually takes place at the level of genes. Oligonucleotide frequencies can also be subject to convergent evolution: if two distantly related phages slowly converge to similar usage frequencies, these methods can give a false indication of common ancestry (Pride *et al.*, 2003).

Alignment-free methods have been used to study phage and bacterial genomes in a variety of contexts. For example, Pride *et al.* (2006) found tetranucleotide usage to carry a strong phylogenetic signal in bacteriophages and showed that tetranucleotide composition was similar among phages with common hosts. More recently, Ogilvie *et al.* (2013) surveyed metagenomic sequencing datasets using a tetranucleotide usage-based method and discovered several novel *Bacteroidales*-like phages which could not be identified with alignment-based methods. Oligonucleotide composition vectors have also been proposed as a method to root viral phylogenies (Simmons, 2008).

Statistics based on nucleotide composition in a sliding window can theoretically be used to uncover horizontal gene transfer (HGT), based on the assumption that genomes have self-similar nucleotide composition and outlier regions could represent recent horizontal transfer events (Lawrence & Ochman, 1997). Guanine + Cytosine

(GC) content in a sliding window was first used to look for pathogenicity islands within a genome (Hacker & Kaper, 2000). More recent methods have used nucleotide composition and Naïve Bayesian classifiers (Sandberg *et al.*, 2001) or hidden Markov models (Waack *et al.*, 2006). However, if horizontally transferred segments change in oligonucleotide composition to be more similar to the resident genome, a process known as amelioration, it can obscure truly horizontally transferred segments (Koski *et al.*, 2001).

The number of sequenced mycobacteriophages has grown immensely in the past few years thanks to the Howard Hughes Medical Institute (HHMI) Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) course (Jordan *et al.*, 2014). This program allows first year undergraduate students to isolate and characterize novel mycobacteriophages from the environment. It has also provided excellent opportunities for collaborative projects between undergraduates, resulting in the work presented here Siranosian *et al.* (2015a).

As the number of sequenced mycobacteriophages continues to increase, researchers need new methods to quickly make comparisons at many scales. Alignment-free methods are one possibility: they are independent of sequence alignment or genome annotation, less computationally complex than alignment-based methods and applicable to genomes without a common subsequence. We investigated tetranucleotide usage in mycobacteriophage genomes as an alignment-free alternative to traditional methods for genome comparison. Our findings support what is known about mycobacteriophage biology: phages form identifiable groups and subgroups, known as clusters, but have extensive differences between clusters. Tetranucleotide usage also highlights outliers in the population and can describe unique genomic features. All of the analyses here can be done in minutes on a personal laptop. Tetranucleotide usage is a powerful tool to quickly investigate features of the growing mycobacteriophage population.

Methods

We obtained the genomic sequences of all 663 sequenced mycobacteriophages publicly available on the website PhagesDB.org as of April 2014. This dataset contains both unpublished genomes and genomes available on GenBank. There is not an easy way to download the mycobacteriophage database in its entirety, so we automated the process with a Python script available in the code accompanying this manuscript.

To compare mycobacteriophage genomes independently of sequence alignment, we investigated the usage of *k*-mers, substrings of DNA of length *k,* in each genome. Given a value for *k*, there are 4^k possible substrings. For example, the 16 possible ways to combine {A, T, C, G} in substrings of length two are {AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC, GG}. Different values for *k* are used throughout this paper, but we focus mainly on results from *k*=4 and *k*=6. In the following section, a substring of length *k* is called a word, abbreviated by *W*. Before computing *k*-mer usage, each genome is extended by the reverse complement to account for biases from transcriptional start orientation.

With a chosen value of *k*, we first compute the number of times each substring occurs in the genome. This gives a vector *N* of length 4^k, where each entry *N(W)* is the number of times word *W* occurs in the genome sequence. Next, we normalized the *k*-mer frequencies using a zero-order Markov model, which removes biases from the background nucleotide composition and can be effective for analysis of prokaryotic genomes (Pride *et al.*, 2003; Pride *et al.*, 2006). Normalization accounts for the fact that GC-rich genomes are expected to have more GC-rich *k*-mers simply because of the available nucleotide composition. Dividing the observed counts of *k*-mers by the expected counts highlights *k*-mer usage that can differentiate between mycobacteriophage genomes.

The expected number of a *k*-mer *W* given the background nucleotide distribution is calculated by:

$$E(W) = [(A^a * T^t * C^c * G^g) * N]$$

where *A,T,C,G* are the frequency of each nucleotide in the genome, *a,t,c,g* are the number of each nucleotide in the *k*-mer *W,* and *N* is the length of the genome.

The normalized value for a word *W* is calculated by dividing the observed counts by the expected counts. This is the usage deviation vector for a genome, and in the case of *k*=4, tetranucleotide usage deviation (TUD):

$$TUD(W) = N(W)/E(W)$$

An example of calculating TUD values for a short sequence is given in Figure 1. This is equivalent to the "tetranucleotide usage departures from expectation" measure proposed by Pride *et al.* (2003). For a given 4-mer, a TUD value of one corresponds to the expected usage, while a value of two corresponds to usage twice as frequently as expected.

Data filtering

Phage genomic sequences are extended by the reverse complement before calculation, leading to redundant values for a given tetranucleotide and its reverse complement. One of the redundant tetranucleotides was removed before distance calculations and Principal Components Analysis (PCA). We also removed tetranucleotides that were not present at least once in all phage genomes. Only ATAT and AATT were removed by this filter.

Comparison of phage genomes

To compare phage genomes in an alignment-free way, we calculated the Euclidean distance between usage deviation vectors. In the case of *k*=4 for a pair of TUD vectors from genomes *x* and *y*:

$$d_{x,y} = \sqrt{\sum_{W=1}^{4^4} \left(TUD_x(W) - TUD_y(W)\right)^2}$$

Where individual 4-mers are indexed by integers ranging from 1 to 4^4.

Sequence: **A A T T G C A A T T**

Step 1: Count occurrence of 4-mers in a sliding window of size 4, step size of 1. This gives the number of each word *W, N(W).* Compute the frequency of each nucleotide in the sequence.

4-mer	Count
A A T T	2
A T T G	1
T T G C	1
T G C A	1
G C A A	1
C A A T	1

Nucleotide	Frequency
A	0.4
T	0.4
C	0.1
G	0.1

Step 2: Compute expected count of each word *W* using equation *E(W)*.

$E(W) = [(A^a * T^t * C^c * G^g) * (N]$

4-mer	Expected count
A A T T	$[(0.4)^2 \times (0.4)^2 \times (0.1)^0 \times (0.1)^0 \times 7] = 0.1792$
A T T G	$[(0.4)^1 \times (0.4)^2 \times (0.1)^0 \times (0.1)^1 \times 7] = 0.0448$
T T G C	$[(0.4)^0 \times (0.4)^2 \times (0.1)^1 \times (0.1)^1 \times 7] = 0.0112$
T G C A	$[(0.4)^1 \times (0.4)^1 \times (0.1)^1 \times (0.1)^1 \times 7] = 0.0112$
G C A A	$[(0.4)^2 \times (0.4)^0 \times (0.1)^1 \times (0.1)^1 \times 7] = 0.0112$
C A A T	$[(0.4)^2 \times (0.4)^1 \times (0.1)^1 \times (0.1)^0 \times 7] = 0.0448$

Step 3: Compute normalized usage of each word *TUD(W)*.

TUD(W) = N(W) / E(W)

4-mer	TUD
A A T T	2 / 0.1792 = 11.16
A T T G	1 / 0.0448 = 22.33
T T G C	1 / 0.0112 = 89.29
T G C A	1 / 0.0112 = 89.29
G C A A	1 / 0.0112 = 89.29
C A A T	1 / 0.0448 = 22.33

Note: this example uses a very short sequence, leading to high TUD values for any 4-mers that do occur. With sequences in range of mycobacteriophage genomes (length 40k – 150k), it is rare to see TUD values above 6.

Figure 1. Example of calculating TUD for an input sequence of 10 bases.

Computing pairwise distances between all usage deviation vectors produced a distance matrix used for tree building. For analysis of the subset of 60 phage in Hatfull *et al.* (2010), we used the SplitsTree program (Huson & Bryant, 2006) to construct neighbor joining phylogenetic trees. This was done to facilitate easy comparisons between previously published figures and our alignment-free trees. Hierarchical clustering using the "average" method within the statistical programming language R (version 3.1.0) was used to construct dendrograms for analyzing the entire phage database.

Principal components analysis

PCA was used to visualize relationships between phage genomes in lower- dimensional space. PCA was done on log-transformed data in R using the 'prcomp' function and results were plotted using the 'ggbiplot' package.

Within-genome comparisons

To compare tetranucleotide usage within a phage genome, we used a sliding window of 2000bp (500bp step size). This window size was selected to balance two factors: a short window can detect differences in small regions, while a longer window is necessary to encounter the majority of tetranucleotides. 4-mers were counted and normalized to the nucleotide composition of a given window. A distance matrix was constructed from pairwise Euclidean distances of all windows and used to build heatmaps. Parts of the heatmap where windows overlapped were removed before plotting, leading to the white section along the diagonal in Figure 5.

Results

Mycobacteriophage genomes have heterogeneous, yet clustered tetranucleotide usage

First, we investigated if TUD reflected relationships described from alignment-based analysis of phage genomes. In particular, does a grouping scheme based on tetranucleotide usage agree with previously assigned phage clusters? To test this hypothesis, we examined a subset of 60 mycobacteriophages first analyzed by Hatfull *et al.* (2010), where the authors propose a clustering scheme based on dot-plot comparisons, pairwise average nucleotide identities, pairwise genome maps and gene content analysis. We calculated the pairwise Euclidean distances between TUD vectors for the subset of 60 phages and used the SplitsTree program (Huson & Bryant, 2006) to construct a neighbor joining tree (Figure 2a).

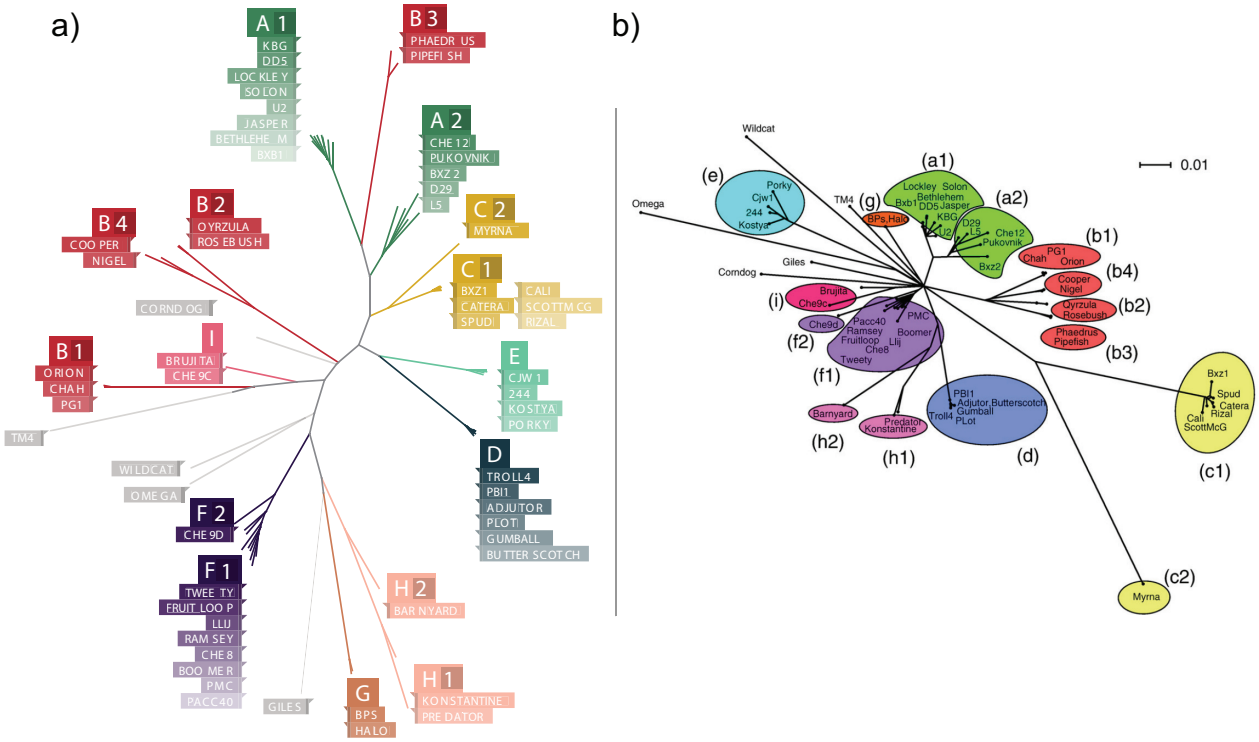


Figure 2. TUD captures similarity within mycobacteriophage subclusters. a) Neighbor joining phylogenetic tree constructed from pairwise Euclidean distances between TUD vectors for 60 mycobacteriophage genomes. Phage names are colored based on previously assigned cluster information. **b**) Neighbor joining phylogenetic tree constructed from gene presence data in mycobacteriophage genomes. Reproduced with permission from Figure 3 in Hatfull *et al.* (2010). The TUD tree is similar to the alignment-based tree. Phages from the same subcluster form monophyletic clades. In clusters C, F and H, subclusters from the same parent cluster form monophyletic clades.

Our alignment-free tree has a striking resemblance to the tree from Hatfull *et al.* (2010), which is constructed from similarities in genomic architecture (Figure 2b). In every case, phages are placed in a monophyletic clade with members of their subcluster.

Hierarchically grouping phages into clusters and subclusters represents heterogeneity within the mycobacteriophage population. In the alignment-free tree, subclusters from parent clusters C, H and F are placed in a monophyletic clade. However, in some cases, tetranucleotide usage was vastly different between subclusters of a parent cluster. For example, subcluster B3 phages are most similar to cluster A phages in terms of tetranucleotide usage, but they are similar to other cluster B genomes when compared on genetic elements (Figure 2). We investigate this relationship further in a following section. Importantly, the relationships between the subset of 60 phages are consistent for varying values of *k* (Figure 3).

Hundreds of mycobacteriophages have been sequenced in the past few years, bringing the total to 663 genomes (PhagesDB.org as of April 2014), 21 clusters and 48 subclusters. We next examined TUD patterns in the entire database to see if the relationships observed for the subset of 60 phages were conserved. We used hierarchical clustering within R to analyze this larger dataset (see Methods). As observed for the subset of 60, almost all phages are grouped closely with members of their subcluster. Subclusters of cluster F, C, D, M and L form a monophyletic clade (Supplementary Figure 1). The relationships for cluster B genomes are also conserved – genomes within a given B subcluster are similar, but the subclusters themselves are different and placed in separate sections of the dendrogram.

Principal components analysis captures variation in tetranucleotide usage

We further investigated the ability of TUD to differentiate between predetermined phage clusters using PCA. PCA is useful for visualizing TUD, a 256-dimensional vector, in intuitive 2D space. PCA was applied to log-transformed TUD vectors for all 663 genomes. The first three principal components captured 29.3%, 15.6% and 12.9% of the variance, respectively. Comparing PC1 and PC2 highlighted groups of phage that corresponded well with assigned clusters (Figure 4a). Clusters that were similar in PC1/PC2 space could be separated further by including additional PCs.

PCA was also useful to compare phages within a single cluster. When comparing cluster B phages, the first three components captured 44.6%, 31.4% and 10.3% of the variance present, respectively. Phage subclusters typically group tightly with each other in PC-space, which makes it easy to detect outliers in terms of TUD. A single member of B4, KayaCho, is placed far from the other genomes of that subcluster (Figure 4b). This indicates that KayaCho is dissimilar from other B4 phages, a finding that is supported through other methods of comparison. For example, KayaCho has a similar global genome architecture to other members of B4, but pairwise nucleotide identity is low in relation to other comparisons within the subcluster. TUD provides a quick and alignment-free way to detect genomes that are outliers within a subcluster.

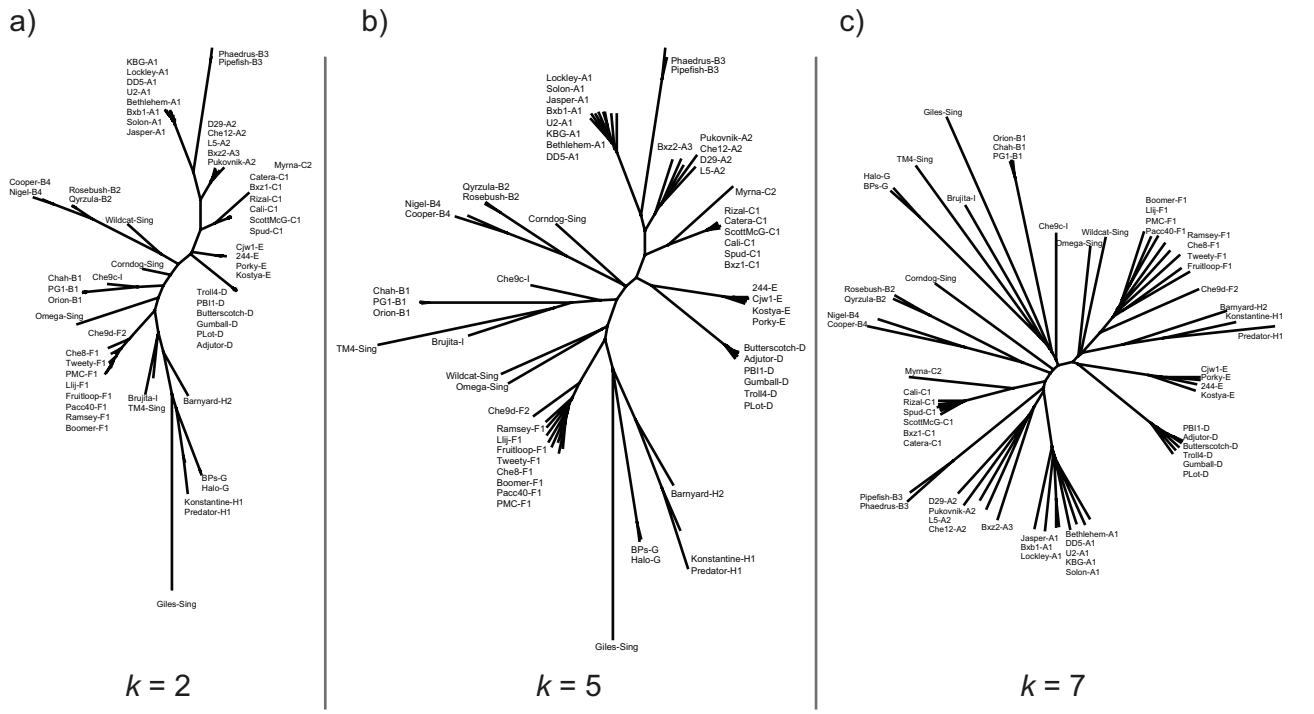


Figure 3. Changing k does not change the structure of the tree. Neighbor joining phylogenetic trees constructed from pairwise Euclidean distances between oligonucleotide usage deviation vectors for 60 mycobacteriophage genomes. Trees from k equal to two, five and seven are shown here. Trees show a high degree of similarity regardless of the k used. Trends observed in the tetranucleotide usage based tree (Figure 2), such as grouping of subcluster members into monophyletic clades, are conserved in these trees.

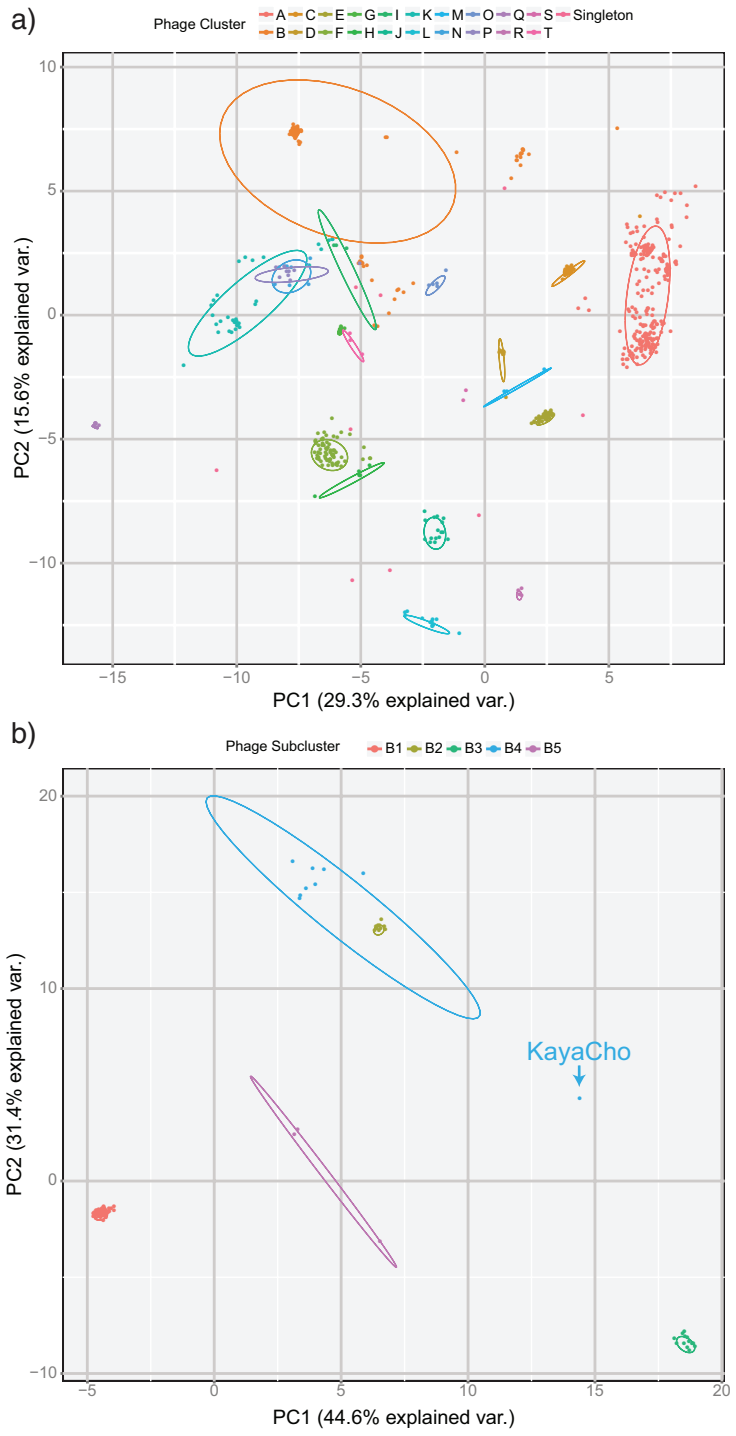


Figure 4. PCA differentiates between clusters and subclusters. a) Principal components analysis of all 663 mycobacteriophage genomes. Individual clusters of phages are well separated by PC1 and PC2 in most cases. Further separation can be achieved by incorporating additional principal components. **b**) Principal components analysis of cluster B phages. Individual subclusters are well separated. The outlier in B4 is KayaCho, a phage with different tetranucleotide usage but similar genome architecture when compared with other B4 phages.

Mycobacteriophage genomes have self-similar tetranucleotide usage, but some regions are outliers

Mycobacteriophage genomes are mosaic and heavily influenced by horizontal gene transfer (HGT) (Pedulla *et al.*, 2003). We looked for sections within a phage genome that stood out in TUD as potential candidates for HGT events. Tetranucleotide usage was calculated in a 2000bp window with a 500bp step size. Heatmaps of pairwise Euclidean distances between all windows were plotted.

Observation of these heatmaps revealed several interesting features. The last 5kb of cluster E phage "244" is self-similar, but different than the rest of the genome in terms of TUD (Figure 5a). This self-similar segment is present with >97% nucleotide identity in all cluster E phage and could represent a HGT event from a different phage cluster or organism. To search for potential transfer sources of this segment, we compared TUD in the region with other mycobacteriophages and searched for nucleotide similarity with BLAST (nr/nt database, blastn algorithm) (Altschul *et al.*, 1997). However, we were unable to find regions of considerable homology with either method.

Cluster L1 phages contain two small self-similar yet genome-different regions at the end of the genome (Figure 5b). We examined the genome of "UPIE" with the Repfind program (Betley *et al.*, 2002) to search for repetitive sequences that could be driving the change in TUD. There are two blocks of repetitive GC-rich *k*-mers, from 68650-69050bp and 71100-71900bp, which match the regions in the heatmap (Figure 6). As the sliding window moves through each of these blocks, the TUD signal becomes dominated by the repetitive sequence and makes the regions appear self-similar yet genome different. The repetitive features don't preclude the possibility of HGT in the region, but they do likely obscure a HGT signal carried by TUD. We found other self-similar yet genome-different repetitive regions in phages from clusters F1, H and O. Although the regions highlighted here have variations in GC content, TUD removes biases from the nucleotide composition using a zero-order Markov model (see Methods). Differences in TUD are not a result of variations in the underlying GC content.

B3 phages contain overrepresented 4-mers and 6-mers

Finally, we examined why B3 phages are not placed with other members of cluster B in the hierarchical clustering dendrogram, while most of the other clusters show this relationship. B3 genomes share greater than 60% average nucleotide identity with other members of cluster B. This is comparable with the relationship between B2 and B4 phages, which are placed close to each other in the dendrogram. The difference in TUD is not likely to be driven solely by differences in pairwise nucleotide identity. We investigated the individual *k*-mers making up the TUD vector to examine this relationship further.

B3 phages used the 4-mer GATC four times more than expected by chance, greater than all other B subclusters (Figure 7a). The high abundance of GATC could be driven by a global increase in

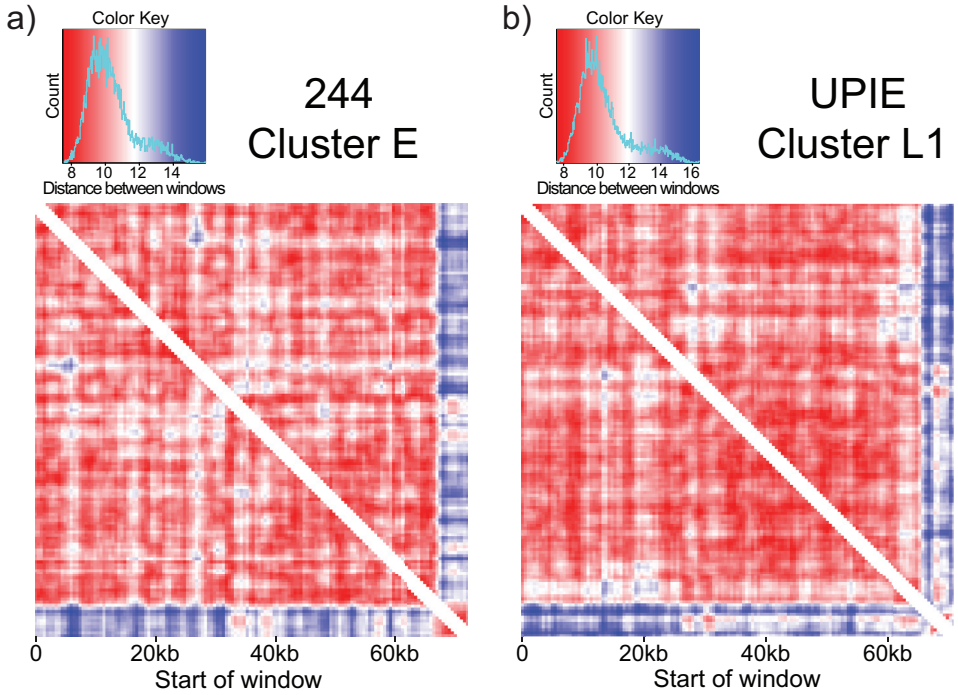


Figure 5. TUD highlights putative horizontally transferred segments. Comparing tetranucleotide usage in a sliding window (2000bp window, 500bp step size) across phage genomes. Each entry in the heatmap is the Euclidean distance between windows. **a**) 244, a cluster E phage, is relatively self-similar with low distance values (red) between most windows. The last 5kb of the genome is an exception: it is self-similar but different than the rest of the genome. This signature is not driven by repetitive sequences, and represents a putative HGT event. **b**) UPIE, a cluster L1 phage, also has a self-similar signature at the end of the genome. However, the difference in TUD in this window is driven by two cluster of repetitive k-mers (Figure 6).

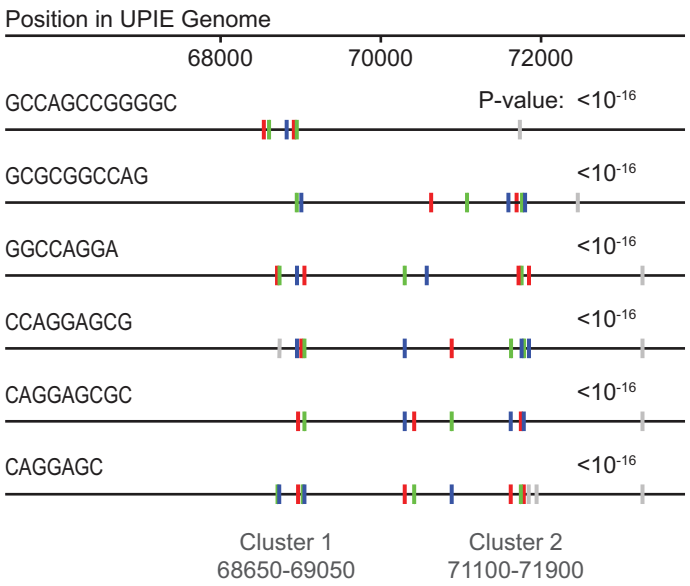


Figure 6. L1 phages contain two clusters of repetitive k-mers. Two clusters of GC-rich repetitive sequences at the end of the genome of UPIE (cluster L1). The repetitive sequences drive the differences in TUD and correspond with the self-similar yet genome-different sections in the within-genome heatmap (Figure 5). This image was reconstructed from the output of Repfind (Betley *et al.*, 2002).

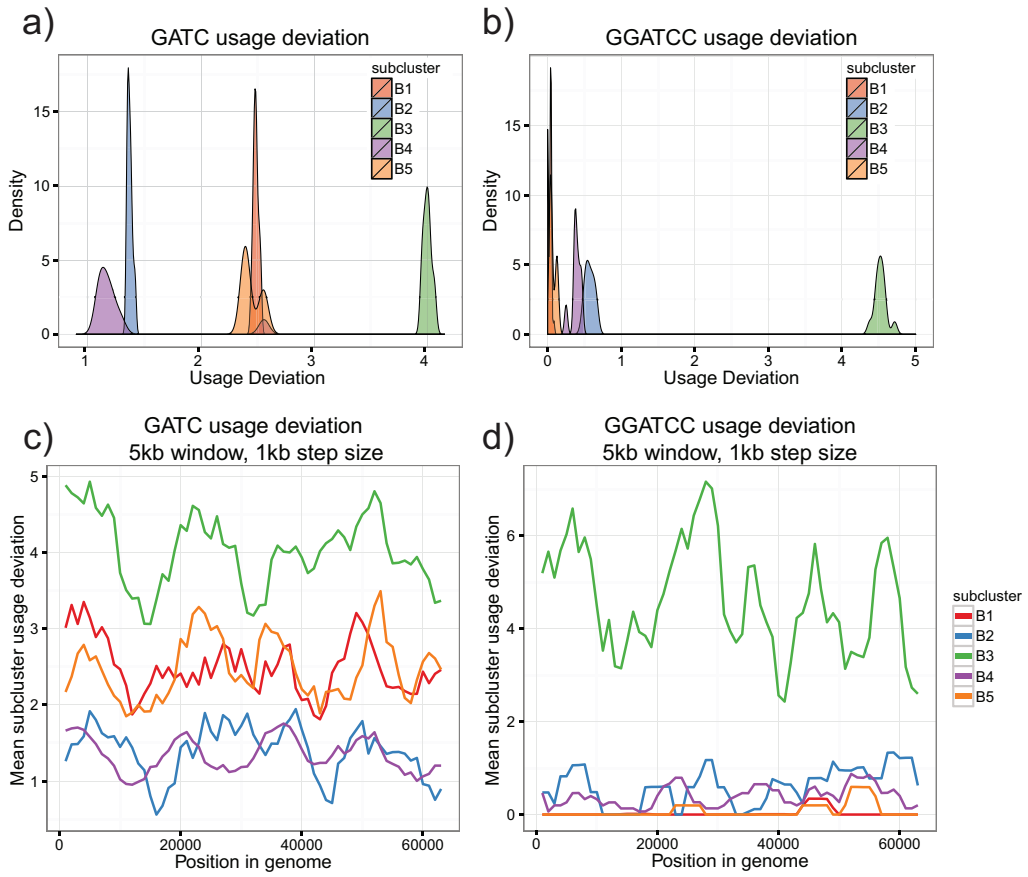


Figure 7. GATC and GGATCC are overrepresented in B3 phages. a) Density plot of TUD values for the 4-mer GATC. Individual subclusters form well-defined groups. B3 phages have GATC usage four times what is expected, much higher than other B subclusters. **b**) Repeat of (**a**) with the 6-mer GGATCC. B3 phages use this 6-mer greater than four times what is expected. **c**) GATC usage deviation in a sliding window (5kb, 1kb step size). Each line represents the mean value in the specified subcluster. The increase in GATC usage is genome-wide, indicative of a global change in usage frequency. **d**) Repeat of (**c**) with the 6-mer GGATCC. Increased usage is also genome-wide.

frequency or by discrete regions with very high usage of the 4-mer. To address this point, we compared normalized GATC usage in a sliding window across all cluster B genomes. GATC usage was increased genome-wide in B3 phages, refuting the hypothesis that the deviation was caused by a single genomic region (Figure 7c). This points to a genome-wide amelioration of GATC usage in cluster B3 genomes. Interestingly, some local peaks and valleys in GATC usage are persistent across all cluster B genomes, even though these genomes are unaligned.

Given the genome-wide increase in B3 GATC usage, it is possible that a higher-order signal could be driving the trend. We searched for highly used 6-mers in B3 phages and found GGATCC had a usage deviation value greater than four, while all other B genomes had a value less than one (Figure 7b). This increase was also genome-wide (Figure 7d). GATC and GGATCC are both palindromes, DNA sequences with identical reverse complements. Palindromes are typically underrepresented in bacteriophage and other prokaryotic genomes because they can be parts of recognition sites for restriction enzymes (Gelfand & Koonin, 1997; Karlin *et al.*, 1992; Sharp, 1986).

GATC is recognized by Dam methylase in *E. coli* (Marinus & Morris, 1973), but *Mycobacterium* species do not encode Dam methylase (Hemavathy & Nagaraja, 1995). If B3 phages recently accessed a host with an active Dam methylase, it could lead to a change in GATC frequency. Several restriction enzymes recognize GATC, like *Mgo*I in *Mycobacterium gordonae* (Shankar & Tyagi, 1993), while others recognize GGATCC, such as *Bam*HI in *Bacillus amyloliquefaciens*. However, the presence of a restriction/modification system in a host would theoretically lead to a decrease in usage of the recognized site. The finding that GATC and GGATCC occur in B3 genomes four times more than expected and significantly more frequently than in all other sequenced mycobacteriophages bears further investigation.

Discussion

In 2010, there were 60 sequenced mycobacteriophages. There are more than 660 as of April 2014. Alignment-based methods have been used to investigate the mycobacteriophage population, leading to interesting characterizations, such as hierarchical grouping into clusters and subclusters. However, as the number of published genomes continues to grow, there is a need for methods to quickly analyze the entire database of mycobacteriophage sequences.

Throughout this paper, we apply oligonucleotide usage methods to uncover relationships within the population of sequenced mycobacteriophages. These methods allow phage genomes to be compared independently of sequence alignment or genome annotation. The methods for counting *k*-mer usage and normalizing to expected counts are simple to implement and compute. A usage deviation value has a clear interpretation: a value of two corresponds to a *k*-mer occurring twice as frequently as expected in a randomized genome sequence. Usage deviation vectors are also well-suited to distance computation and PCA.

Our findings support what is known about mycobacteriophage biology. Neighbor joining and hierarchical clustering from TUD place closely related phage in well-defined groups that correspond with assigned phage subclusters. In most cases, TUD supports grouping into larger clusters, such as cluster A, where all 246 members form a monophyletic clade in the hierarchical clustering dendrogram. The fact that members of cluster B do not form a clade in TUD-based comparisons does not invalidate grouping of phage into clusters, but rather serves as a way to highlight phages where TUD and gene or sequence comparisons capture different relationships.

Comparing TUD in a sliding window can highlight regions with dissimilar tetranucleotide composition and identify genomic segments that could have been horizontally transferred. We found self-similar yet genome-different regions at the end of cluster E and L genomes. The new TUD 'space' occupied by these segments could be from HGT – a recently transferred genomic section that had not yet ameliorated to the average genome TUD profile. At least for cluster L, we can say that HGT is likely not the cause. Two groups of repetitive sequences at the end of the genome are driving the difference in TUD. However, we found neither repetitive sequences nor a putative transfer candidate for the segment in cluster E. An improvement on our method could potentially detect legitimate HGT events, but we note that the concept of phams (Hatful *et al.*, 2010) and the computer program Phamerator (Cresawn *et al.*, 2011) are already efficient for detecting and visualizing these features.

TUD vectors are similar between subcluster B3 phages but different from other members of cluster B. We found that the 4-mer GATC and 6-mer GGATCC were present over four times more than expected in B3 genomes. These sequences are palindromes and part recognition sites for restriction enzymes, two characteristics of sequences that are typically underrepresented in prokaryotic genomes. GATC and GGATCC are highly used in all sections of B3 genomes, pointing to genome-wide amelioration of usage frequencies.

Oligonucleotide composition methods do not require knowledge of sequence alignment or gene content. They are ideal to compare mycobacteriophage genomes, which lack a common subsequence on which to make alignment-based inference. Alignment-free methods are also valuable when a reference sequence is not available. Recently, methods based on tetranucleotide usage were used to investigate sequences from a gut microbiome and uncovered a population of *Bacteroidales*-like phage that was previously unrepresented in metagenomic sequencing datasets (Ogilive *et al.*, 2013). Statistics based on oligonucleotide usage are part of a broader class of alignment-free methods. These methods are easy to compute across large datasets: constructing the dendrogram in Supplementary Figure 1 from raw phage sequences takes less than two minutes on a personal laptop. Comparably, creating phylogenetic trees from pairwise global sequence alignment with the

Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) takes over 24 hours on a computing cluster. We envision oligonucleotide usage methods to be used alongside alignment-based techniques. Highlighting large trends and outliers is easy with these methods, but sequence alignment and gene annotation need to be applied to extract biological insights from the data.

Data and software availability

The genomic sequences of all 663 sequenced mycobacteriophages are publicly available on the website PhagesDB.org as of April 2014. The authors obtained permission to use the data.

Software access

The code to download mycobacteriophage genome sequences and reproduce our analysis is freely available at https://github.com/bsiranosian/tango_final. Mycobacteriophage genome sequences are available at http://phagesdb.org.

Latest source code

https://github.com/bsiranosian/tango_final

Source code as at the time of publication

https://github.com/F1000Research/tango_final

Archived source code as at the time of publication

http://dx.doi.org/10.5281/zenodo.14609 (Siranosian *et al.*, 2015b).

Author contributions

BS designed the study. BS, SP, EW and CY performed the analysis. BS and CY prepared the figures. BS, SP, EW, CDG and PS wrote the manuscript.

Competing interests

No competing interests were disclosed.

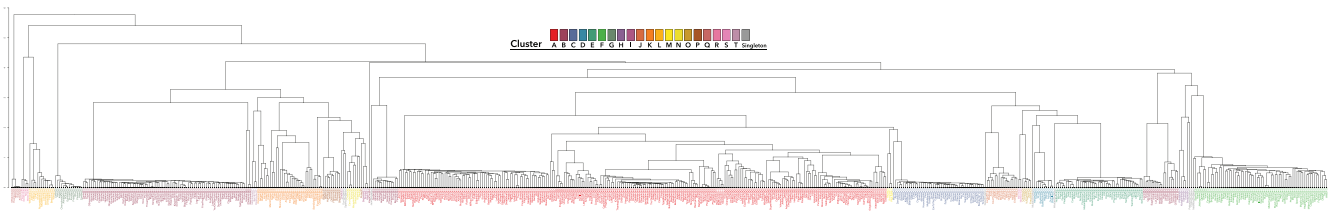
Grant information

This work was funded by Brown University Biology Undergraduate Education and the HHMI SEA-PHAGES program.

Acknowledgments

We would like to thank Sarah Taylor for instructing the Brown University Phage Hunters course and for her assistance during the development and presentation of this work. The present manuscript benefited from helpful comments by Dr. Graham Hatfull. We would also like to thank the hundreds of students from schools participating in the SEA-PHAGES program who have isolated, characterized and purified the mycobacteriophages we analyzed. Finally, we are deeply grateful to the SEA-PHAGES program and Howard Hughes Medical Institute for providing the resources to sequence hundreds of mycobacteriophage genomes, and PhagesDB.org for providing access to the unpublished material that formed the base of this work.

Supplementary material



Supplementary Figure 1. Hierarchical clustering of all 663 phage genomes. Hierarchical clustering dendrogram constructed on pairwise Euclidean distances between all 663 phages in the mycobacteriophage database. In almost every case, phages are placed in a monophyletic clade with members of their subcluster, highlighting the concordance between alignment-based and alignment-free methods for comparison for these genomes. Some clusters (F, C, D, M and L) form monophyletic clades, while others (B, for example) are grouped in different parts of the dendrogram. A larger version of this figure can be downloaded from here.

References

Altschul SF, Madden TL, Schäffer AA, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997; **25**(17): 3389–3402.
PubMed Abstract | Publisher Full Text | Free Full Text

Betley JN, Frith MC, Graber JH, *et al.*: **A ubiquitous and conserved signal for RNA localization in chordates.** *Curr Biol.* 2002; **12**(20): 1756–1761.
PubMed Abstract | Publisher Full Text

Bohannan BJM, Lenski RE: **Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage.** *Ecology Letters.* 2000; **3**(4): 362–377.
Publisher Full Text

Chan CX, Ragan MA: **Next-generation phylogenomics.** *Biol Direct.* 2013; **8**: 3.
PubMed Abstract | Publisher Full Text | Free Full Text

Chibani-Chennoufi S, Bruttin A, Dillmann ML, *et al.*: **Phage-host interaction: an ecological perspective.** *J Bacteriol.* 2004; **186**(12): 3677–3686.
PubMed Abstract | Publisher Full Text | Free Full Text

Cresawn SG, Bogel M, Day N, *et al.*: **Phamerator: a bioinformatic tool for comparative bacteriophage genomics.** *BMC Bioinformatics.* 2011; **12**: 395.
PubMed Abstract | Publisher Full Text | Free Full Text

Danelishvili L, Young LS, Bermudez LE: ***In vivo* efficacy of phage therapy for *Mycobacterium avium* infection as delivered by a nonvirulent mycobacterium.** *Microb Drug Resist.* 2006; **12**(1): 1–6.
PubMed Abstract | Publisher Full Text

Doolittle WF: **Phylogenetic classification and the universal tree.** *Science.* 1999; **284**(5423): 2124–2129.
PubMed Abstract | Publisher Full Text

Frith MC, Hamada M, Horton P: **Parameters for accurate genome alignment.** *BMC Bioinformatics.* 2010; **11**: 80.
PubMed Abstract | Publisher Full Text | Free Full Text

Gelfand MS, Koonin EV: **Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes.** *Nucleic Acids Res.* 1997; **25**(12): 2430–2439.
PubMed Abstract | Publisher Full Text | Free Full Text

Hacker J, Kaper JB: **Pathogenicity islands and the evolution of microbes.** *Annu Rev Microbiol.* 2000; **54**: 641–679.
PubMed Abstract | Publisher Full Text

Hatfull GF, Jacobs-Sera D, Lawrence JG, *et al.*: **Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size.** *J Mol Biol.* 2010; **397**(1): 119–143.
PubMed Abstract | Publisher Full Text | Free Full Text

Hatfull GF: **Mycobacteriophages: windows into tuberculosis.** *PLoS Pathog.* 2014; **10**(3): e1003953.
PubMed Abstract | Publisher Full Text

Hemavathy KC, Nagaraja V: **DNA methylation in mycobacteria: absence of methylation at GATC (Dam) and CCA/TGG (Dcm) sequences.** *FEMS Immunol Med Microbiol.* 1995; **11**(4): 291–296.
PubMed Abstract | Publisher Full Text

Hendrix RW: **Bacteriophages: evolution of the majority.** *Theor Popul Biol.* 2002; **61**(4): 471–480.
PubMed Abstract | Publisher Full Text

Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol.* 2006; **23**(2): 254–267.
PubMed Abstract | Publisher Full Text

Jordan TC, Burnett SH, Carson S, *et al.*: **A broadly implementable research course in phage discovery and genomics for first-year undergraduate students.** *MBio.* 2014; **5**(1): e01051–13.
PubMed Abstract | Publisher Full Text | Free Full Text

Karlin S, Burge C, Campbell AM: **Statistical analyses of counts and distributions of restriction sites in DNA sequences.** *Nucleic Acids Res.* 1992; **20**(6): 1363–1370.
PubMed Abstract | Publisher Full Text | Free Full Text

Koski LB, Morton RA, Golding GB: **Codon bias and base composition are poor indicators of horizontally transferred genes.** *Mol Biol Evol.* 2001; **18**(3): 404–412.
PubMed Abstract | Publisher Full Text

Lawrence JG, Hatfull GF, Hendrix RW: **Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches.** *J Bacteriol.* 2002; **184**(17): 4891–4905.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol.* 1997; **44**(4): 383–397.
PubMed Abstract | Publisher Full Text

Marinus MG, Morris NR: **Isolation of deoxyribonucleic acid methylase mutants of *Escherichia coli* K-12.** *J Bacteriol.* 1973; **114**(3): 1143–1150.
PubMed Abstract | Free Full Text

McNerney R: **TB: the return of the phage. A review of fifty years of mycobacteriophage research.** *Int J Tuberc Lung Dis.* 1999; **3**(3): 179–184.
PubMed Abstract

Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol.* 1970; **48**(3): 443–453.
PubMed Abstract | Publisher Full Text

Ogilvie LA, Bowler LD, Caplin J, *et al.*: **Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences.** *Nat Commun.* 2013; **4**: 2420.
PubMed Abstract | Publisher Full Text | Free Full Text

Pedulla ML, Ford ME, Houtz JM, *et al.*: **Origins of highly mosaic mycobacteriophage genomes.** *Cell.* 2003; **113**(2): 171–182.
PubMed Abstract | Publisher Full Text

Pride DT, Meinersmann RJ, Wassenaar TM, *et al.*: **Evolutionary implications of microbial genome tetranucleotide frequency biases.** *Genome Res.* 2003; **13**(2): 145–158.
PubMed Abstract | Publisher Full Text | Free Full Text

Pride DT, Wassenaar TM, Ghose C, *et al.*: **Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses.** *BMC Genomics.* 2006; **7**: 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Sandberg R, Winberg G, Bränden CI, *et al.*: **Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier.** *Genome Res.* 2001; **11**(8): 1404–1409.
PubMed Abstract | Publisher Full Text | Free Full Text

Shankar S, Tyagi AK: **Purification and characterization of restriction endonuclease *Mgo*I from *Mycobacterium gordonae*.** *Gene.* 1993; **131**(1): 153–154.
PubMed Abstract | Publisher Full Text

Sharp PM: **Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes.** *Mol Biol Evol.* 1986; **3**(1): 75–83.
PubMed Abstract

Simmons MP: **Potential use of host-derived genome signatures to root virus phylogenies.** *Mol Phylogenet Evol.* 2008; **49**(3): 969–978.
PubMed Abstract | Publisher Full Text

Siranosian B, Herold E, Williams E, *et al.*: **Tetranucleotide usage in mycobacteriophage genomes: alignment-free methods to cluster phage and infer evolutionary relationships.** *BMC Bioinformatics.* 2015a; **16**(Suppl 2):A7.
Publisher Full Text | Free Full Text

Siranosian B, Perera S, Williams E, *et al.*: **Code to download mycobacteriophage genome sequences.** *Zenodo.* 2015b.
Data Source

Vinga S: **Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification** in *Advanced Computational Methods for Biocomputing and Bioimaging.* (Nova Science Publishers). 2007; 71–107.
Reference Source

Waack S, Keller O, Asper R, *et al.*: **Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models.** *BMC Bioinformatics.* 2006; **7**: 142.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status:  

Version 2

Referee Report 20 April 2016

doi:10.5256/f1000research.7828.r13337

David Martin
Life and Biomedical Sciences Education, School of Life Sciences, University of Dundee, Dundee, UK

The study provides an interesting approach to the evaluation of divergence between the phage genomes. I'm not an expert in this area so come into it with a more general view. I found the revised paper clear and well explained in terms of approach. I agree with the first reviewer that the authors have perhaps been selective in just showing data from a select choice of *k*-mer values. Expanding the results to show the deviation across the full range of *k* tested, even if just in summary, would be interesting, though there would be a disparity between odd and even values of *k* as there are no palindormes with odd *k*.

A minor issue with regard to the present publication, but which might be worth consideration for future work, is over the TUD metric where the authors compare the observed frequencies to the expected. It is not clear from the study as to the variation one might see in a null model. If TUD is the test statistic of choice, a significance value for the deviation from expected should be deteminable empirically by modelling TUD, e.g.where there is a randomly assigned sequence of nucleotides corresponding to the genome of the organism. This could be done by shuffling the whole genome, taking a large sliding window and aggregating these scores (with or without shuffling etc.) A discussion of the significance of the deviation from expected (or the lack of appreciation of it) is worth including into the paper.

It is nice to see the distance measures, but without an estimate of the significance of the deviation from expected values, it becomes difficult to assess the significance of the deviation between genomes. It may be the case that using a significance measure as the distance (a Z-score or equivalent) may produce a different clustering.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 17 November 2015

doi:10.5256/f1000research.7828.r11005

Oliver Bonham-Carter
College of Information Science & Technology, School of Interdisciplinary Informatics, University of Nebraska, Omaha, NE, USA

My initial concerns have been addressed.

References

1. Bonham-Carter O, Steele J, Bastola D: Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis.*Brief Bioinform*. 2014; **15** (6): 890-905 PubMed Abstract | Publisher Full Text

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 10 March 2015

doi:10.5256/f1000research.6506.r7811

Oliver Bonham-Carter
College of Information Science & Technology, School of Interdisciplinary Informatics, University of Nebraska, Omaha, NE, USA

The article is nicely written but sadly, there are elements of discussion which are absent from the paper. If added, the paper's research on mycobacteriophages using alignment-free analysis would have much more support.

- The choice of TUD's as statistics for the alignment-free analysis is not fully explained /justified, nor is there much discussion about what algorithm or method is being employed by the analysis tools of the paper. Are TUD's frequencies? How do these software tools work?

- An simple example of how to calculate a TUD and apply it to a method is necessary to completely understand what they are and to see how they are different from any other motif frequency calculation applied to some other method.

- The assumptions of the methods are not discussed. Many methods from information theory, statistics and other kinds of mathematics require that the input data meets specific requirements (is normal, has a certain distribution, is a frequency, etc.). From the discussion in this paper, the function of analysis tool (the exact algorithm or method) is never clear and so we cannot be sure that the calculations from this work, as applied to these tools, is appropriate. For instance, many tools in information theory require that frequencies be used for their analysis. These frequencies must pass basic rules to be called as such (i.e., found on the scale of 0 to 1, all frequencies must sum to 1, 0 = false, 1 = true). This discussion is not mentioned and if it were, then the choice to used TUDs could be easily integrated into this discussion.

- The manuscript mentioned that k-mers in the range of two to seven were calculated (Methods Section). Where are the results for all these other values of k={2, 3, 5 and 7} which were not the k={4 and 6} results of the article?

- Although other sizes of motifs where apparently used in the analysis, the manuscript focuses on the length-4 motifs. The choice of k=4 for the size of motifs to study is not a very interesting statistic since the probability of a particular length-4 motif showing up randomly in a sequence not very high (1/(4^4) = 1/256). Given that the frequency of mutations, and all the evolutionary time during which to make changes to a sequence, these length=4 similar motifs are likely to randomly turn-up.

- The authors should consider using the occurrence of motifs which are at least seven since these frequencies begin to become less randomly placed. Length-4 words are already common in many many bacteria as restriction sites for restriction enzymes. The authors will also find that there are restriction sites of length-6 for the same purpose and so they will have to remove all restriction enzyme palindromes from their sets of k=4 or 6 sized motifs if they cannot continue with a longer motif length. However, if they are determining the level of conservation between organisms, then having longer motifs should not hurt their results.

Once these issues are addressed, the manuscript will be much stronger.

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 23 Oct 2015

Benjamin Siranosian,

-Thank you for reviewing the manuscript. I have considered the points you raised, and responded in order below. Changes to the manuscript are noted.

1. The usage deviation-based statistics chosen for this paper are similar to those based on the composition vector of a sequence (Bonham-Carter *et al*., 2013). Usage deviation (tetranucleotide usage deviation, TUD, in the case of k=4) is a vector of the counts of the possible k-mers, normalized to the expected counts in a randomized genome with the same nucleotide composition. I have made additions to the methods section and included a new figure that makes the calculation of usage deviation more clear. The software tools used to perform these calculations have a description at the github page linked in the paper.

2. I have added an example in the methods section that shows how to calculate TUD for a small sequence. Although this example outlines the method, the results are not very informative. The expected number of any 4-mer is very small in a short sequence, resulting in high TUD values for any 4-mers that do occur.

3. We do not make any assumptions about the input data when calculating usage deviation or performing statistics in the paper.

4. I showed trees constructed from other values of k in Figure 2. The relationships between phage genomes were consistent regardless of the value chosen for k. Other analyses mirrored this result, so we proceed exclusively with k={4, 6}.

5. I agree that length-4 motifs are not interesting to study in isolation. Usage deviation, where values represent deviations from expected frequencies, overcome this point. Single

occurrences or counts of any 4-mer are uninteresting. Only when counts are normalized and compared in aggregate do the trends that observed in the paper become meaningful.

6. 7-mers would be less randomly placed in the phage genomes analyzed. Similar to the point above, however, the occurrences of singular k-mers are not considered. As k increases, the resulting usage deviation vectors become sparse. Up to 43% of the (4^7=16384) 7-mers are absent from individual genome sequences, and no 7-mer occurs at least once in every genome analyzed. The sparse nature of the data for 7-mers would not be well-suited to some of the analyses presented in this paper (PCA, searching for horizontally transferred segments).

7. I acknowledge that many 4-mers and 6-mers are restriction sites. In fact, this makes the substrings more interesting. B3 mycobacteriophages have 4 times the expected usage of GATC, a restriction site in some bacteria. Biological sense dictates restriction sites would occur infrequently, but the results say the opposite. I do not feel it is necessary to remove restriction sites before the analysis, and doing so would be somewhat arbitrary. The set of restriction sites in mycobacteria species is not entirely characterized, and the host range for each mycobacteriophage has not been studied.

We hope you find the answers to the points you raised and the revisions to the paper acceptable.

References:

Bonham-Carter, O., Steele, J. & Bastola, D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief Bioinform* **15,** 890–905 (2014).

Competing Interests: No competing interests were disclosed.