TETRANUCLEOTIDE USAGE IN MYCOBACTERIOPHAGE GENOMES

ALIGNMENT-FREE METHODS TO CLUSTER PHAGE AND INFER EVOLUTIONARY RELATIONSHIPS

Chen Ye | Benjamin Siranosian | Emma Herold | Minjae Kwon | Sudheesha Perera | Edward Williams | Sarah Taylor | Christopher de Graffenried

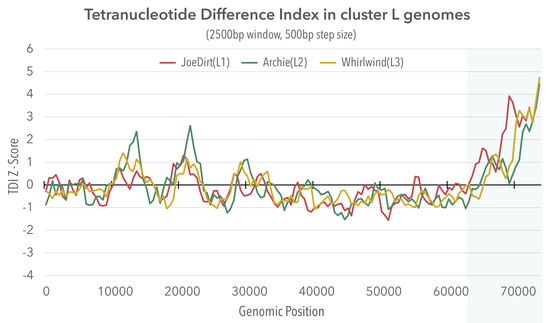
INTRODUCTION

Traditionally, phage genomes are compared using methods that require sequence alignment or gene annotation. These methods may be ineffective for populations with significant horizontal gene transfer and are computationally intensive for large datasets. Mycobacteriophages also lack a common genetic element, like ribosomal RNA in bacteria, from which to compute phylogenetic relationships. Alignment-free sequence analysis methods, such as measures that compute the usage of oligonucleotides in a genome, have the potential to infer relationships between significantly diverged sequences. We examined the usage of tetranucleotides in all 663 phage genomes available in the mycobacteriophage database as an alternative to alignment and annotation based methods.

We found tetranucleotide usage deviation (TUD), a normalized measure of tetranucleotide usage in a genome, to be comparable for members of the same phage subcluster and distinct between subclusters. We used TUD as a measure of distance between phage and were able to:

- Construct phylogenetic trees that place members of a subcluster in a monophyletic clade
- Accurately assign subclusters to phage with a nearest neighbor classifier
- Identify windows in a genome with significantly different tetranucleotide usage, possibly indicating horizontal gene transfer

GENOMIC SELF-SIMILARITY



HORIZONTAL GENE XFER?

investigate with

Phamerator

BLAST

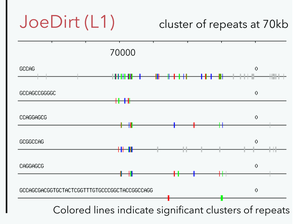
some homology

JoeDirt gp 130 @ 70,000 bp

- *Mycobacterium abscessus* E = 2e-48
- *Flavobacterium psychrophilum* E=2e-28
- *Opitutaceae bacterium* TAV1 ATPase E = 1e-23

more likely

- Cluster L genomes are very repetitive at the end.
- Repetitive regions have increased counts of specific 4-mers, contributing to the spike in TDI.



Colored lines indicate significant clusters of repeats

METHODS

k-mer counting

4-MERS ARE COUNTED USING A SLIDING WINDOW

GATGATGATCATG

GATGATGATCATG

GATGATGATCATG

GATGATGATCATG

HERE'S THE RESULT

GATG ×2

ATGA ×1

TGAT ×1

tetranucleotide usage deviation

To remove biases in tetranucleotide counts, we divided each observed count by the number of tetranucleotides expected under a model of random nucleotide distribution. This gives the TUD for a tetranucleotide w.

$$TUD(w) = \frac{observed}{expected}$$

$$Exp(w) = [(A^a * C^c * G^g * T^t) * N - 3]$$

A, C, G, T: genomic frequency of respective nucleotides
a, c, g, t: tetranucleotide frequency of nucleotides
N: length of genome

The TUD signal for a genome is a vector of $4^4 = 256$ values – one for each possible tetranucleotide

tetranucleotide difference index

Genomes are relatively self-similar in oligonucleotide usage. A region with a drastically different TUD signal can indicate horizontal transfer of genetic material. We computed the tetranucelotide difference index (TDI) in a sliding window to look for regions of interest in phage genomes.

Tetranucleotide differences are measured in each window s by the equation:

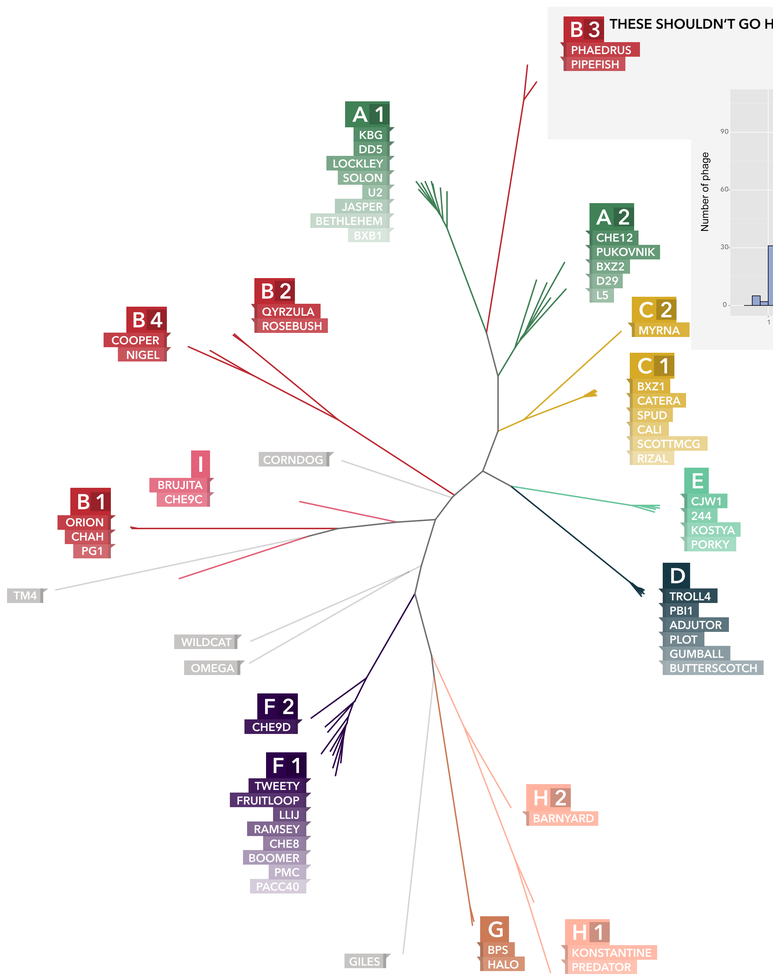
$$TD_s = \sum_{i=1}^{256} |TUD_s(w_i) - TUD_G(w_i)|$$

TUD_s: the TUD value for word w_i in the sliding window
TUD_g: the TUD value for the entire genome

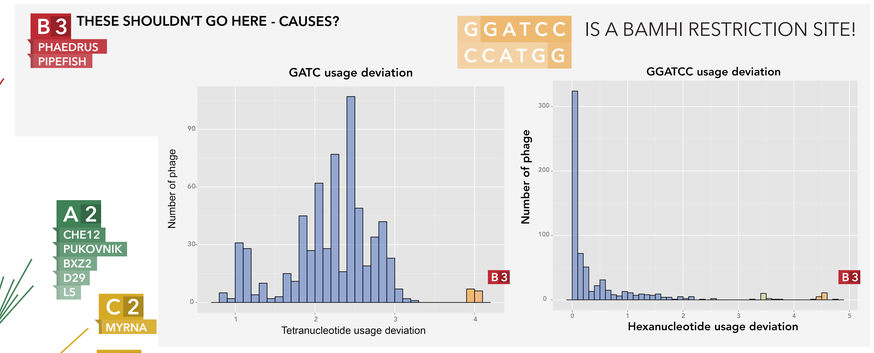
We compare the Z-score of tetranucleotide differences for each window to find regions of significant difference:

$$Z_s = \frac{TD_s - mean(TD)}{stdev(TD)}$$

NEIGHBOR-JOINING TREE FROM TUD DISTANCE



B3 EXCEPTIONAL K-MER MOTIFS



CONCLUSIONS

Tetranucleotide usage deviation and other alignment-free methods can investigate relationships within the diverse mycobacteriophage population. TUD accurately reconstructs phylogenetic trees and can highlight regions of particular interest in a genome. These methods can be applied in a high-throughput manner, take very small amounts of computational time, and serve as an excellent first pass in the comparative analysis of a mycobacteriophage genome. With some further work we hope to see these methods applied to every new phage sequence.

FUTURE DIRECTIONS

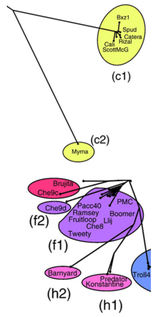
host-parasite coevolution

Hosts and parasites have similar oligonucleotide usage profiles. We will use data available on phage host preference to investigate this point further.

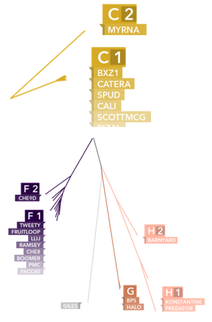
horizontal gene transfer

A naïve Bayesian classifier can use oligonucleotide counts to calculate the probability of a subsequence originating in a given genome. This can be used to find the most likely genome of origin for a possible HGT event. We plan to implement a naïve Bayesian classifier and further investigate leads uncovered with TDI.

Hatfull et. al (2010)



our alignment-free tree



LITERATURE CITED

Betley, J. N., Frith, M. C., Graber, J. H., Choo, S. & Deshler, J. O. A ubiquitous and conserved signal for RNA localization in chordates. Curr. Biol. 12, 1756–1761 (2002).

Hall, M. et al. The WEKA Data Mining Software: An Update. SIGKDD Explor. f. 11, 10–18 (2009).

Hatfull, G. F. et al. Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size. Journal of Molecular Biology 397, 119–143 (2010).

Sandberg, R. et al. Capturing Whole-Genome Characteristics in Short Sequences Using a Na?ve Bayesian Classifier. Genome Res 11, 1404–1409 (2001).

Pride, D. T., Wassenaar, T. M., Ghose, C. & Blaser, M. J. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8 (2006).

ACKNOWLEDGEMENTS

We are grateful to Dr. Peter Shank, Dr. Sorin Istrail, Dr. Zhijin Wu, HHMI's SEA program and the University of Pittsburgh.

additional information

Source code and processed data is available at

github.com/bsiranosian/tango

bsiranosian.com

A digital copy of this poster is available at

yeesus.com/tangoposter

HHMI
HOWARD HUGHES MEDICAL INSTITUTE