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 deGraffenried

Background

The genomic sequences of phages isolated on mycobacterial hosts are incredibly diverse and often share little nucleotide similarity. Many tools used for the analysis of mycobacteriophage genomes depend on sequence alignment or knowledge of gene content. These methods are computationally expensive, can require significant manual input (for example, gene annotation) and can be ineffective for significantly diverged sequences. We evaluated tetranucleotide usage in mycobacteriophages as an alternative to alignment-based methods for genome analysis.

Description

First, we computed tetranucleotide usage deviation, the ratio of observed counts of 4-mers in a genome to the expected count under a null model. Tetranucleotide usage deviation is comparable for members of the same phage subcluster and distinct between subclusters. Hierarchical clustering dendrograms and neighbor joining phylogenetic trees were constructed on pairwise Euclidean distances between all 663 genomes in the mycobacteriophage database. In almost every case, phage were placed in a monophyletic clade with members of the same subcluster. We found that tetranucleotide usage deviation is efficient at capturing relationships between subclusters of the same cluster, in contrast to previous findings that suggest tetranucleotide usage does not carry a strong phylogenetic signal. We also evaluated the possibility of assigning clusters to unknown phage based on tetranucleotide usage deviation. Under a simple nearest neighbor classifier, cluster assignments were recovered at a frequency greater than 98%.

In addition, we looked for evidence of horizontal gene transfer by using tetranucleotide difference index, a measure of the deviation in tetranucleotide usage from the genomic mean in a sliding window across the genome. Our tetranucleotide difference index plots showed a strong spike at the end of cluster L mycobacteriophages, which could indicate horizontal gene transfer in the region.

Conclusions

Genome analysis based on tetranucleotide usage shows promise for evaluating host-parasite coevolution and gene exchange within the mycobacteriophage population. These methods are computationally inexpensive and independent of gene annotation, making them optimal candidates for further research aimed at clustering phage and determining evolutionary relationships.