CONSIDER FOR TALK

2021 SEA Symposium Abstract

University of California, Los Angeles

Los Angeles CA

Corresponding Faculty Member: Amanda Freise (amandafreise@gmail.com)

Novel application of dimensionality reduction and marker analysis enables comprehensive analysis of Mycobacteriophage diversity

Raneesh Ramarapu, Ryan Fang, Scott Chin, Andrew Kapinos, Amanda C Freise, Jordan Moberg Parker

Mycobacteriophage genomes are constantly evolving through horizontal gene transfer events, resulting in enormous genetic diversity and making comprehensive analysis of their relationships challenging with current methods. Traditional methods of characterizing phage relationships either rely on inefficient pairwise analyses or on broad, large-scale analyses which fall short in capturing the intimate relationships between phages. Here, we report the novel application of dimensionality reduction and marker analysis on mycobacteriophage genomes. Through the R package Seurat, dimensionality reduction was used to group phages based on pham presence. Phages were then plotted onto a two-dimensional space where each phage was placed near other similar phages in the dataset, facilitating identification of patterns across large genomic datasets while preserving small-scale relationships between individual phages. This analysis was assessed across multiple phage data subsets of incremental sizes. Our grouping analysis was fortified through identification of Seurat markers (genes with at least 25% presence in a specified group and no more than 50% presence in all non-specified groups) and more stringent Pham markers (genes with at least 95% presence in a specified group and no more than 5% presence in all non-specified groups). Dimensionality reduction allowed us to visualize subtle relationships between singletons and clustered phages. We explored such relationships further using Marker Analysis and found that singleton phage MooMoo was closely related to Cluster F phages, sharing a comparable number of Seurat and Pham markers with other cluster members. Verification with canonical analyses indicated that MooMoo shared over 35% GCS (the new clustering threshold) with subcluster F4 and F5 phages. In order to explore the spatial distribution of markers within phage genomes, we plotted the average position of each marker pham within every genome and observed two patterns of distribution. The two patterns identified were found to be associated with the level of pham marker conservation within a cluster. We plotted the intra- and inter-cluster conservation of markers across their relative positions on the genome, facilitating the identification of regions of high/low marker concentration in phage clusters as well as putative regions of horizontal gene transfer. Lastly, we found that minor tail proteins were the most called marker function (excluding NKF), exceeding the next most-called function by over 200% across Seurat groups, PhagesDB clusters and PhagesDB subclusters. Minor tail proteins likely play an influential role in phage diversity and play an integral role in phage classification according to our approach. Thus, we propose the application of dimensionality reduction and marker analysis to future phage research, enabling a more comprehensive understanding of the genomic diversity in bacteriophages.