Representing Complex Viral Communities as Networks

Nick Predey, Jason W. Shapiro, and Catherine Putonti, Loyola University Chicago, USA

Abstract

Despite their ubiquity and obvious importance, our level of understanding of viruses generally is still equivalent to being in the dark ages. DNA sequencing of complex viral communities repeatedly finds an abundance of novel genes regardless of the environment being studied. In fact, the vast majority of sequences produced have no resemblance to any characterized known virus. Recently, network-based approaches for analyzing viral DNA sequences have gained popularity as they, unlike prior methods, can handle these "unknown" sequences. This new approach uses the well-studied area of graph and network theory in computer science and mathematics. Sequences generated from sequencing a complex community can thus be represented as a graph in which each node represents an entire genome, and an edge between two nodes indicates a weighted level of similarity between two genomes. While this genome-centric network can classify species within the context of taxonomy, it also can capture the diversity present in nature. Herein, we evaluate the network-based approach for virome analyses, evaluating metrics for assessing similarity and characteristics of the resulting networks, and thus laying the groundwork for future direct analyses of complex communities at a higher resolution than previously possible.

Methods

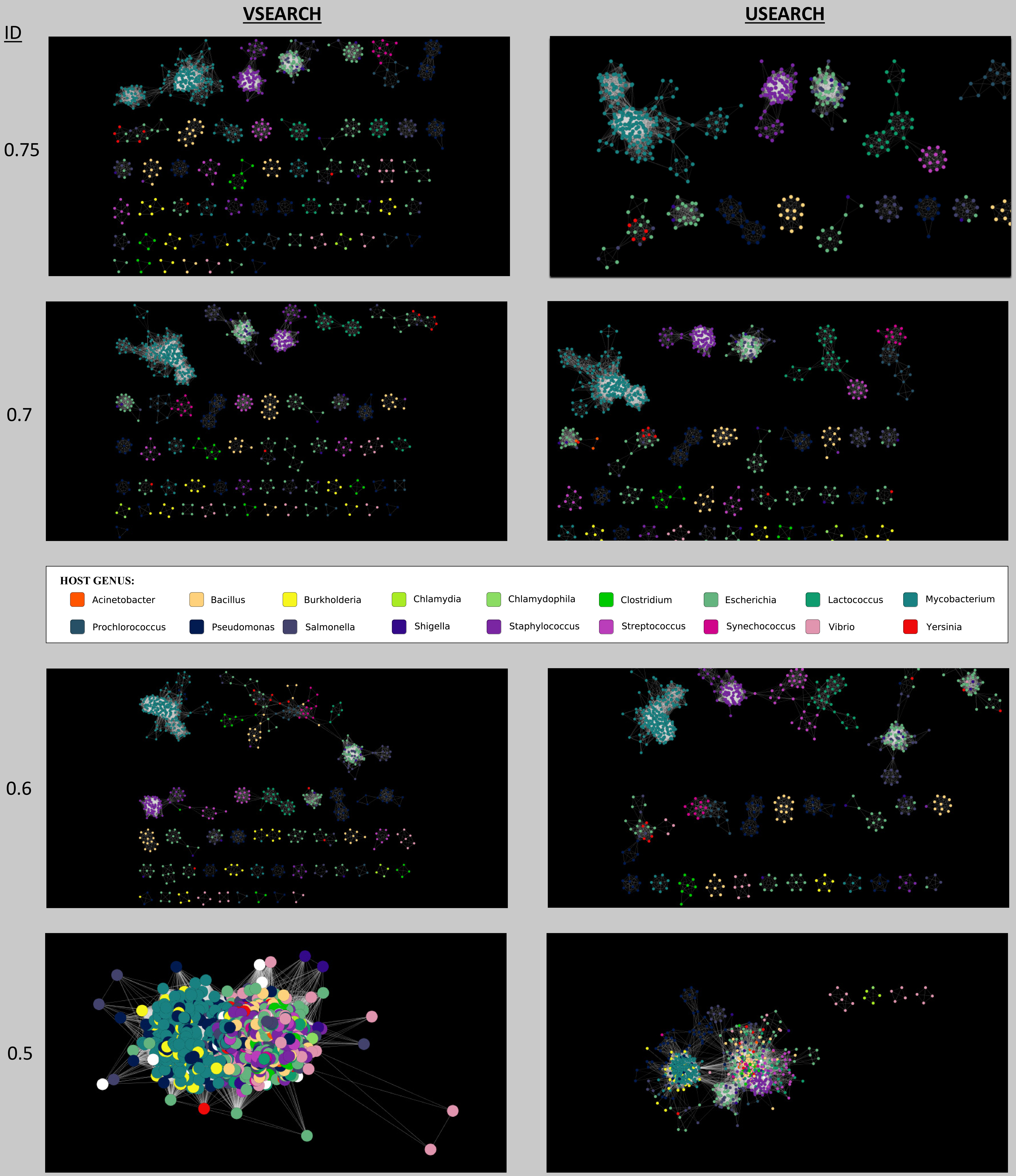
Software – use

USEARCH, VSEARCH – cluster phage genomes and host genes

Python – Create edge list from cluster output

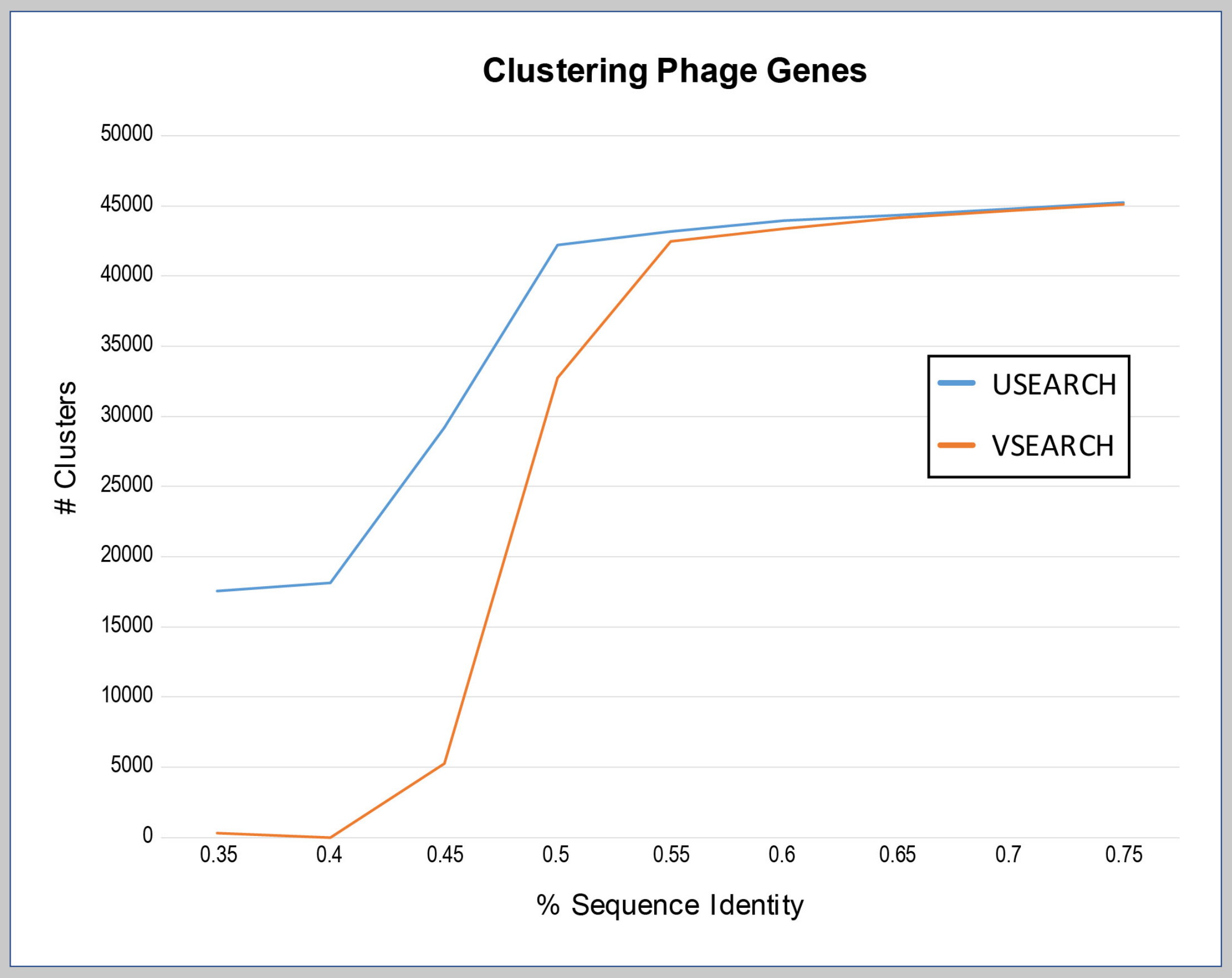
Cytoscape – draw genome networks from graph edge list

- **Use *cluster_fast* command in VSEARCH and USEARCH to cluster file of 82,301 unique genes between 0.35-0.75 identity threshold.**
- **Create an unweighted, undirected graph with clusters > 3 genes, based off genes in cluster outputs.**
- **Plot and color-coordinate graph by phage host genus in Cytoscape.**



Key Results

- **Both USEARCH and VSEARCH do not work well at and below a 0.45 homology threshold.**
- **There are discrepancies between clustering using USEARCH and VSEARCH.**
- **Genomes of the same host tend to infect the same genera.**



Future Directions

- Next, we'll be calculating the similarity & difference between the clusters derived by USEARCH, VSEARCH and a third tool CD-HIT.
- Based on our analysis here, however, we expect that USEARCH will provide the best resource for clustering homologous genes.
- We intend to expand this analysis to include all viral genes.

References

- Edgar (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics* 26(19), 2460-2461.
- Rognes et al. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.