K-mer Signatures Provide Accurate Means of Virus Contig Clustering

Geoffrey D Hannigan, Patrick D Schloss

Introduction

Recent advances in shotgun sequencing technology has allowed for an unprecidented evaluation of virus communities. Shotgun sequencing is required for proper study of the virome because there are no conserved genes analogous to the 16S rRNA genes that can be specifically amplified and used for community studies. Shotgun sequencing techniques provide large sequencing datasets whose analysis is complicated by the divergent virus genomes. The majority of virus genomes are unable to be annotated, and their genomes are modular, with gene cassettes being transferred across viruses[1]. K-mer spectrum analyses have been used as a reference-independent approach to drawing similarities between metagenomic contigs that are otherwise uncomparable by global or local alignment. While mostly described in bacterial environments, it is particularily informative in viral communities whose members share functionality and genome structure with minimal conservation at the nucleotide level. Here we use k-mer spectra to calculate the dissimilarity of virus contigs in these complex communities, and thus provide informative contig classifications that reduce the viral dark matter and provide new insights into their role in human disease.

It is important to note that we are using k-mer spectrum analysis as a tool to understand a biological system, and do not intend this as a software announcement.

Results

Calculation of K-mer Spectrum Dissimilarity

K-mer spectrum analyses have been increasingly utilized in recent years as researchers attempt to classify the unknown components of microbial metagenomes. For this study we built our own k-mer spectrum analysis workflow so that we can maintain the most control over the approach as possible. Similarities between genomes/contigs was calculated using the Bray-Curtis dissimilarity metric (perhaps try other metrics here). To account for genomes in different directions, or contigs with inverted regions, our k-mer spectra were calculated as a composite of k-mers in both forward and reverse. Because dissimilarity metrics like Bray-Curtis are sensitive to uneven sampling, the distances are based on equal sampling depths that were normalized by subsampling the contig with the greater number of k-mers down to an equal amount. When considering this approach as a clustering algorith, it is analogous to the *de novo* OTU clustering approach used in 16S rRNA gene analysis[2]. The processing was made to run in parallel so as to maximize efficiency.

K-mer Clusters Accurately Reflect Virus Biological Relationships

We first confirmed that k-mer spectra provided informative clustering that accurately reflects known biological properties. To acocmplish this, we collected all of the known bacteriophage reference genomes and calculated the degree of dissimilarity between themselves and all other reference phages. Indeed the reference phage genome k-mer classificaiton corresponded to the original phage annotations, which were based on the bacterial host that the phage infects (**Figure 1**). Some phages we more closely related than others, although this suggest that k-mer spectra do in fact accurately represent biological linkages.

We went on to evaluate the ability of k-mer spectra to accurately identify contigs, which in our benchmarking stage were randomly generated fragments of reference genomes. We calculated the k-mer spectrum dissimilarity between the contigs and whole reference genomes to determine whether the contigs could be accurately identified by this method. We found that, given a suffienctly long k-mer window, the spectrum approach performed as accurately as alignment based methods (**Figure 2**). This is largely to be expected because this approach is essentially performing an alignment by matching contigs to reference genomes to which it shares long, unique k-mers. This pseudo-alignment approach to k-mer analyses has been utilized in pervious work, but has failed to be completely evaluated for its alignment-free potential[3].

As mentioned above, the unique utility of a k-mer spectrum analysis is not in its ability to align, but rather in its ability to infer functional and genomic similarities between biologically related but sequenctially dissimilar

genomes. To confirm this benefit over an alignment approach, we assessed the ability of a k-mer spectrum and alignment algoorithm to pair the first and second half of reference genomes. In other words, given the first half of a genome, how accurately can the algorithm identify the matching second half of the genome. We found that alignment performs very poorly at this task and only accurately pairs approximately 20% of the genomes, while the k-mer spectrum algorithm accurately pairs approximately 70% (**Figure 3**). From this we conclude that k-mer spectrum analyses are able to link genome fragments almost four times more accurately than alignment-based approaches. Together with our biological clustering described above, the data suggest that k-mer spectra do in fact correlate with biological linkages, even when the nucleic acids diverge. This is beneficial when linking contigs (genome fragments) that may have minimal nucleotide similarity despite being biologically linked.

K-mer Spectra Allow Classification Within Viral Dark Matter

Now that we have demonstrated the utility of k-mer spectra in linking biologically-related virus genomic sequences, we used the technique to shed new light into the dark matter of the human virome. The virome dark matter is the large set of virome sequences and contigs that cannot be identified due to insufficient reference datasets.

To this end, we classified contigs from human virome datasets. We assembled contigs from the skin virome that were unable to be annotated to a reference genome in the the NCBI non-redundant database by blastn (tblastx?). We used k-mer spectra to identify X% of the "dark matter" virus contigs. Virus were annotated to reference genomes by exhibiting a dissimilarity to the references no greater than their dissimilarity to each other. We also classified contigs by their overall cluster. We also observed that the unannotated contigs create cluster classes themselves (which is what I predict). This demonstrates the utility of classifying unannotatable contigs using k-mer spectra, but also provides some insight into the structure of these unknown viruses.

Human Disease is Associated with K-mer Spectrum Class

Here I want to compare previous assocaitions of viromes with human disease to what we can get by classifying the viruses using the k-mer spectrum method.

Discussion

Conclusions

K-mer spectra have long been an underappreciated approach to understanding genomic relationships and conservation. Our data suggest a significant benefit to incorporating k-mer spectrum analysis in human and environmental virome studies. It is important to stress that this should not replace alignment-based techniques as they are also useful under many circumstances, but supplementing the analysis with k-mers provides a new level to understanding the virome.

Methods

Acknowledgements

Conflicts of Interest

The authors declare no conflict of interest.

Figures

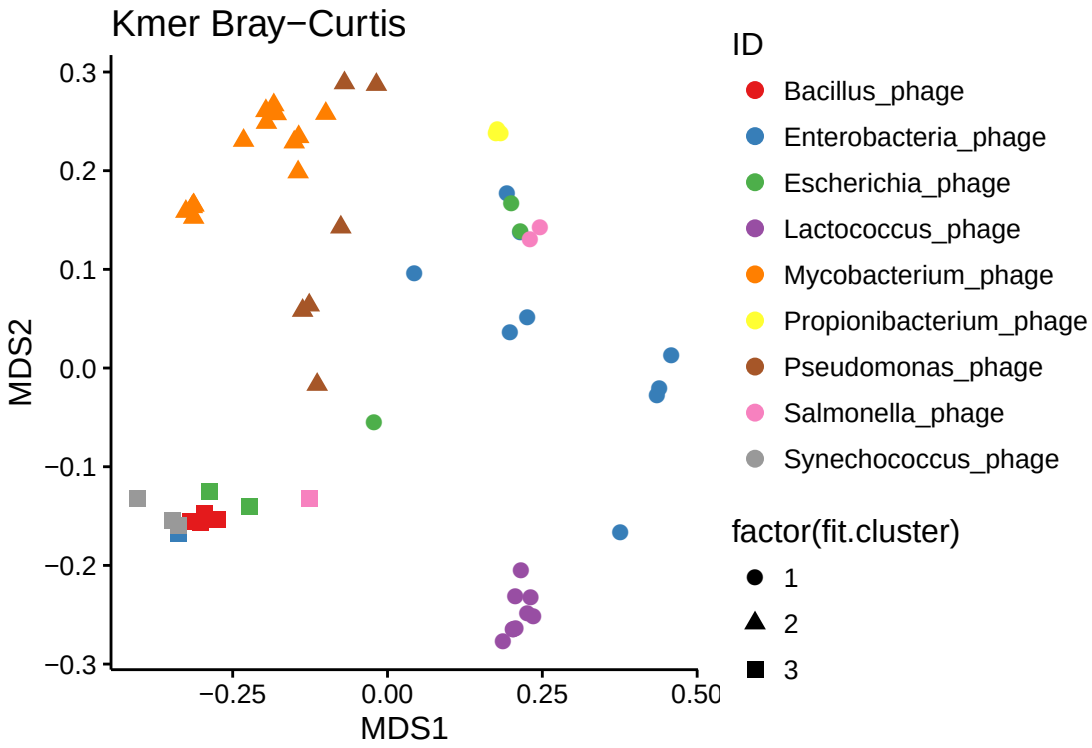

Figure 1: Comparison of reference genome kmer spectra. Only genomes with more than three strains are being shown. Points ideally clustered into 3 groups by k-means clustering. Kmer clustering (shape) is by taxonomic group (colour).

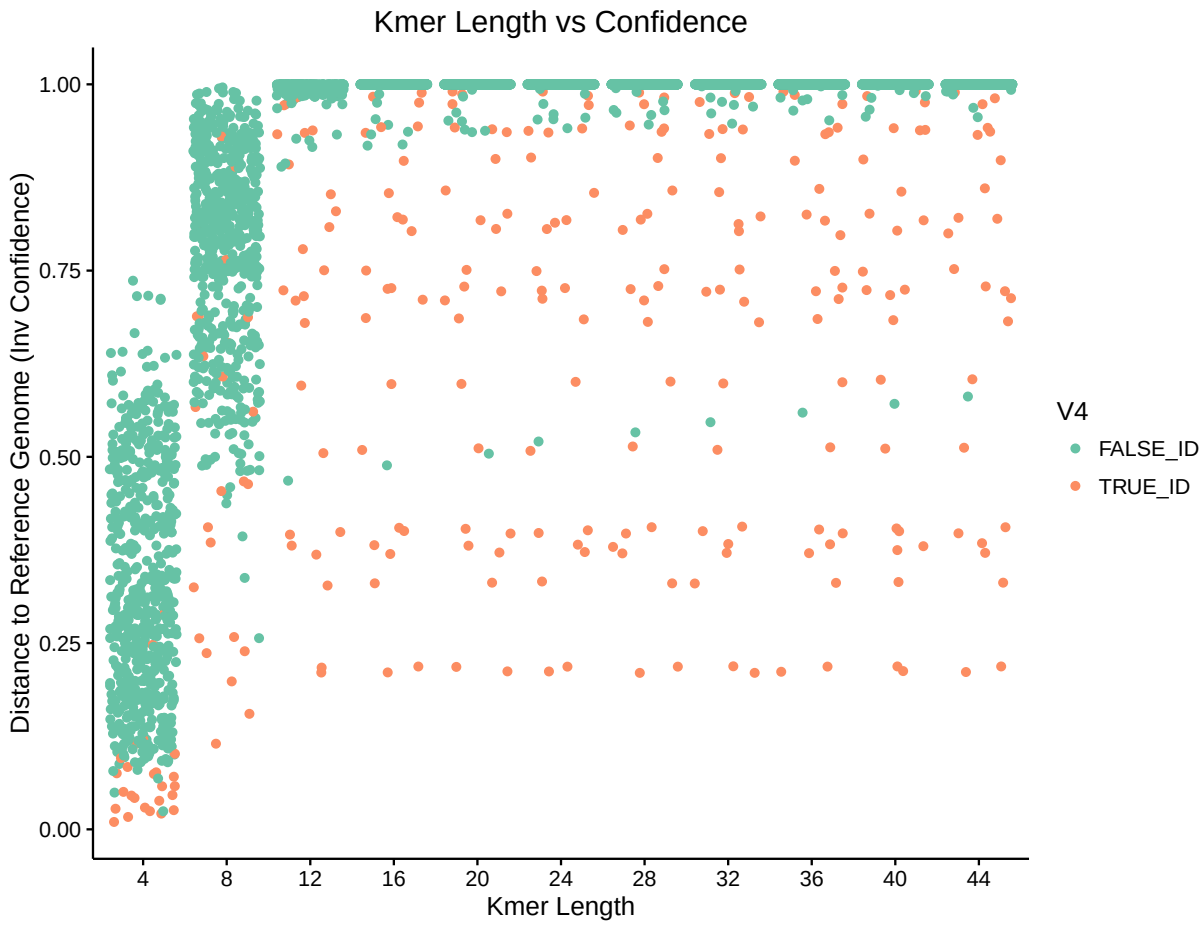

Figure 2: Accuracy of k-mer spectra in contig identification.

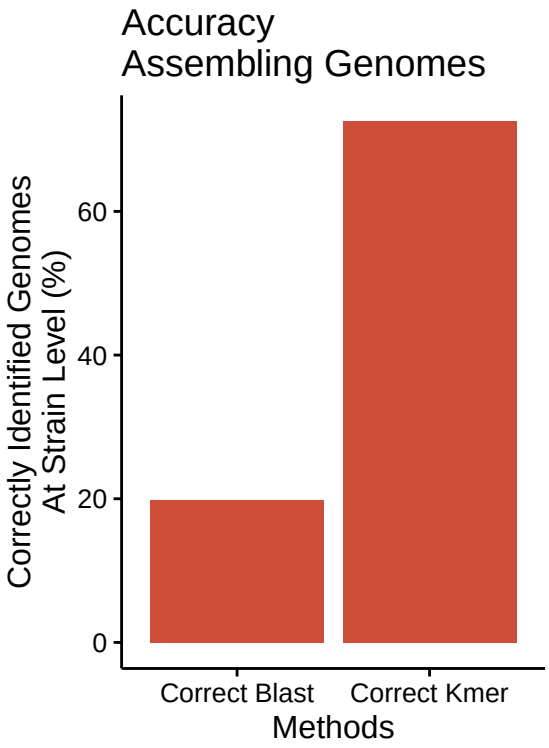

Figure 3: Comparison of kmer spectrum and blast approach to recunstructing reference phage genomes.

References

1. Minot, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Conservation of gene cassettes among diverse viruses of the human gut. *PLOS ONE* **7,** e42342 (2012).

2. Westcott, S. L. & Schloss, P. D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3,** e1487 (2015).

3. Koslicki, D. & Falush, D. *MetaPalette: A K-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation.* (2016).