K-mer Signatures Provide Accurate Means of Virus Contig Clustering

Geoffrey D Hannigan, Patrick D Schloss

Outline

Introduction

Virus studies are complex because they rely on evaluating entire genomes instead of amplicon sequences. Complications arise due to their diverse genome lengths and they are extremely modular.

Viruses are a crucial component to microbial ecosystems including the human microbiome. Their direct infections can cause disease by destroying cells or altering functionality. They can indirectly impact health by altering bacterial communities through predation, transduction (phage-mediated horizontal gene transfer, or both). Advances in shotgun sequencing have recently enabled the robust study of virus communities, oftern termed the virome.

The study of the virome is complicated by a lack of conserved genes that can be used in a manner analogous to 16S rRNA genes. To address this, the field has adopted whole shotgun sequencing techniques which randomly sequence the entire genomes of the viruses present, instead of focusing on a single gene. Just as sequencing technology is advancing, so too are analytical techniques for studying viruses. Unfortunately these techniques are less well developed than those in the amplicon sequencing field.

In these studies, virus sequences are assembled into contigs (i.e. genomic fragments). These are often used as Operational Taxonomic Units (OTUs) despite their incompleteness and variability. Viral contigs are often taxonomically annotated, used for gene prediction, and used for diversity. Ideally these contigs are clustered by similariy into OTUs based on a specified degree of similarity. Many techniques are being investigated to acheive such clustering.

Clustering by kmer frequencies have been used to evaluate the similarities of whole viral communities, as well as specific contigs, although this work has primarily been investigated in bacteria.

Here we present an evaluation of kmer frequency clustering that provides both technical and biological insights into the virome. From a technical perspective, we present new insights into the utility of kmer-based clustering techinuqes in the virome. Most interestingly, kmers are exceptionally adept at linking viral contigs from the same strain genomes but without genomic overlap. From a biological perspective, we provide further evidence of viral kmer signature conservation and thus glean information about viral genomic structure.

The point of this paper is to highlight kmer spectrum frequency throughout virus genomes.

Using an alignment-free technique is particularily benficial in viral systems because viruses are modular in nature and can often swap, gain, or lose genomic material.

Previous work has found that, although viruses within complex gut communities are highly dissimilar at the nucleotide level, contig alignments reveal conserved gene cassetes across diverse viruses. These genes are conserved in their functionality and orientation. Although informative, this is a reference-dependent approach that fails to consider the unknown, unannotated genes. Kmer sprectrum analysis is gaining popularity as an effective method to predict genomic conservation without the use of nucleotide/amino acid alignment or annotation. Although slightly abstract in concept, similar kmer spectra suggest similar nucleotide usage patterns, similar phylogenetic relationships, and similar functionality. Here we use kmer spectra to evaluate the conservation of otherwise unknown viral genomes and contigs in complex viral communities. This allows us to unviel conserved funcitonality in these complex environments, thus providing new insight into the uncharacterized, un-annotated virus dark matter.

Results

Kmer Spectrum Analysis Reflects Biological Relationships

It is important to confirm that the kmer spectrum analysis that are utilizing is informative by showing that it accurately reflects biological relationships. To accomplish this, we first calculated the kmer spectrums of a variety of bacteriophage reference genomes **(Figure 1)**.

Kmer Annotation to Reference Genomes

We began by confirming the ability of kmer spectra to provide for accurate contig annotations. We used the technique to compare whole reference genomes to each other, as well as to annotate reference genome contigs (fragments) using the reference database.

Kmer Contig Clustering

Blast and related alignment algorithms rely on sequence alignment, which is clear from the name. While these algorithms are efficient and effective, they are by definition unable to compare divergent or dissimilar sequences, even if they are biologically linked (e.g. from the same genome). This is the benefit of kmer spectra.

Discussion

Methods

Note to self: Can I make DMN clusters based on the kmer distances or counts (probably counts)? Or other clustering method?

Figures

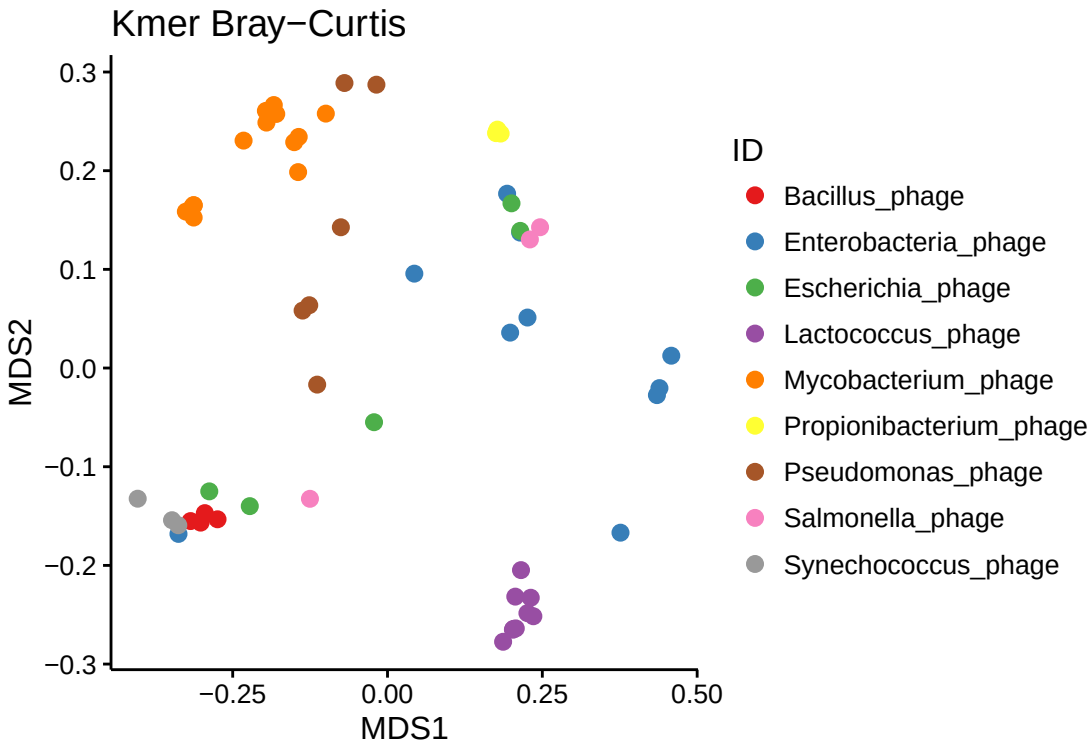

Figure 1: Comparison of reference genome kmer spectra.

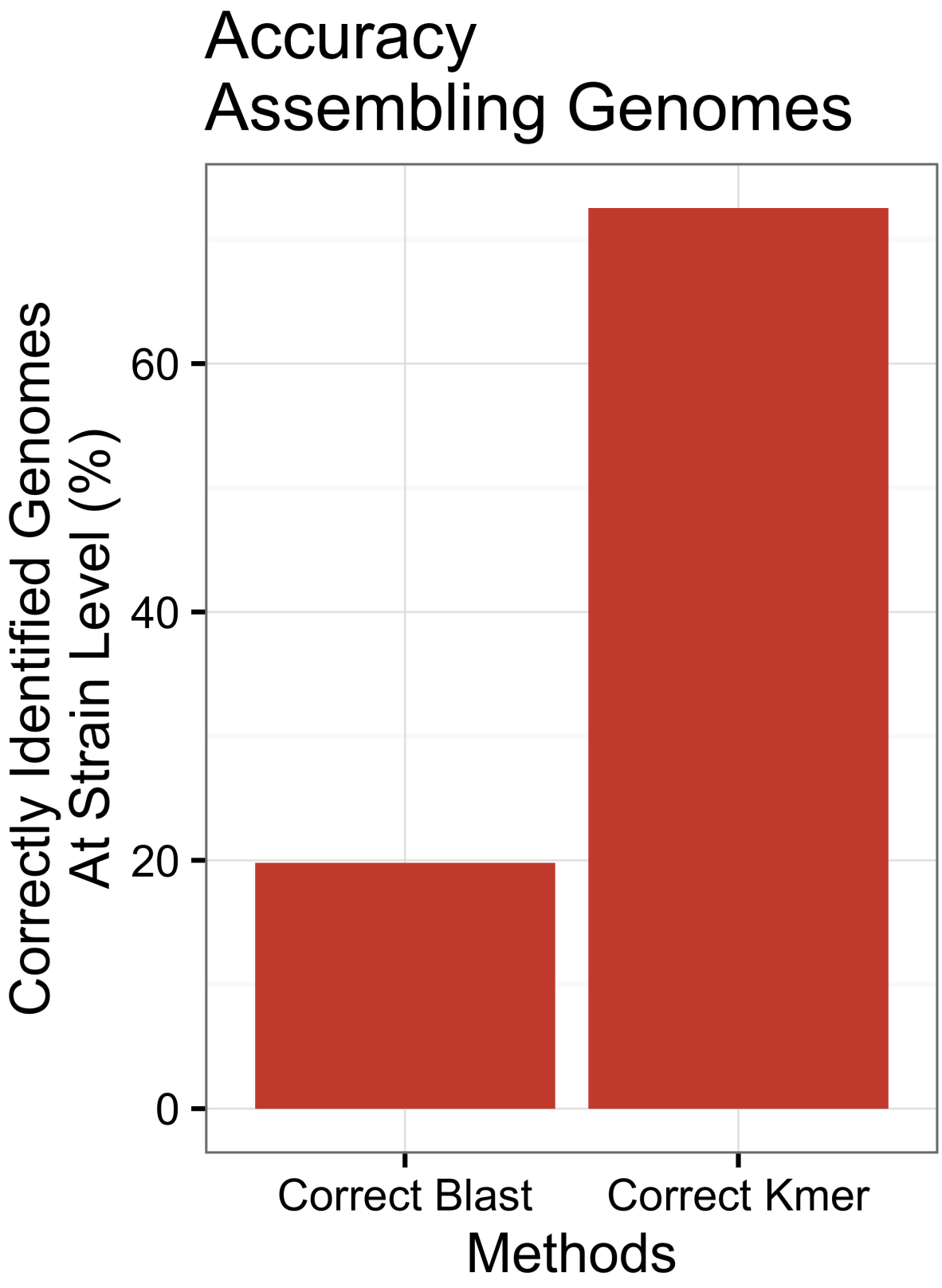

Figure 2: Comparison of kmer spectrum and blast approach to recunstructing reference phage genomes.

Bibliography