Lab Report 5

Zachary Foster

May 14, 2014

It this lab we sought to estimate the number of protein families in the core- and pan- genomes of *Bacteroidetes fragilis* using 12 genomes. The instructors supplied us with amino acid sequences of predicted proteins from the JGI's IMG database. First, I concatenated all of the protein sequences into a single file using `cat`. In order to group similar proteins into families, I used LAST to identify homologs, Linux commands to filter results, and `mcl` to cluster similar sequences into families. Last was used in order to minimize computational time. I used the commands `lastdb` to format a database from the concatenated protein sequence file and `lastal` to perform the pairwise comparison between the database and the concatenated protein sequence file. The resultant `*.m8` text file was parsed with `grep` and `awk` to remove alignments in which less than 80% of the query was aligned or the alignment score was less than 100. `mcl` was then used to cluster the proteins using the filtered alignment information. In total, 8875 protein families were found. The custom perl scripts `parse_mcl.pl` and `add_orphans_mcl_table.pl` were used to parse the mcl output in to a count table and include orphan sequences in the counts. Finally, an R script `mcl.full.sum` was used to plot rarefaction curves of the number of pan- and core- genome protein families.

The core genome rarefaction cure almost levels by the addition of 12 genomes (Figure 1). It appears that there are about 2,000 protein families in the genome core. The pan-genome curve does not level out and may converge on a positive linear correlation, as is predicted by theory (Figure 2). If this is the case than the 12 genomes could be sufficient the ascertain the slope of this relationship, but otherwise many more genomes would be needed. In the 12 genomes sampled there is about 10,000 protein families detected.

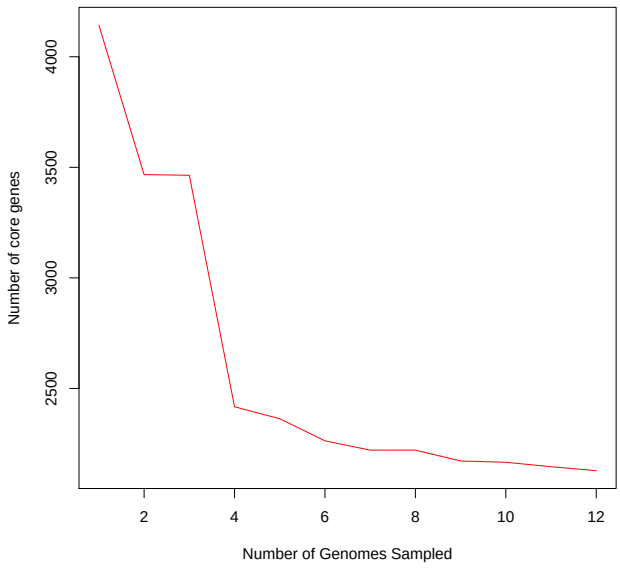

Figure 1: Core genome estimation

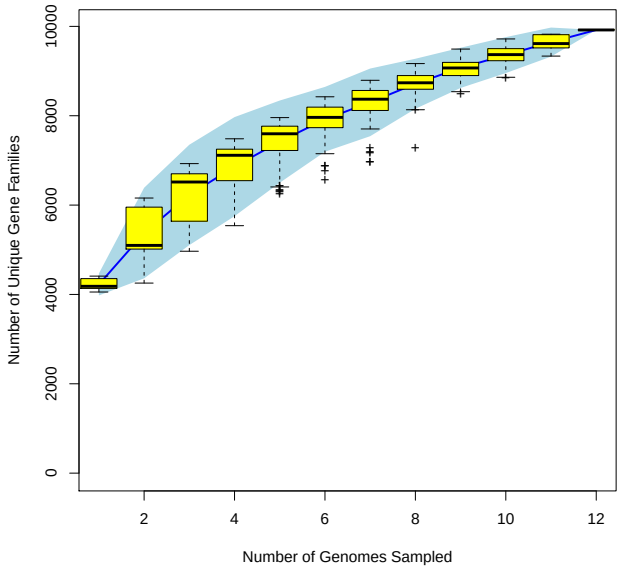

Figure 2: Pan genome estimation