PS6 part 3

1. Describe how the assembly changes with different k-mer values using the assembly statistics you have collected. How does the contig length distribution change?

As kmer length increases the maximum values for the contigs become larger and larger. This is magnified when we add in coverage cut off. In addition, there are less reads overall in larger kmer files as smaller kmers produce more reads with smaller values which makes sense given the first observation. However, there is a general trend across all kmer lengths of a few very long IDs then a steep drop in length that quickly becomes a steady decrease going down the list.

2. How does an increased coverage cutoff affect the assembly? What is happening to the de Bruijin graph when you change the value of this parameter? How does velvet calculate its value for 'auto'?

Doing this sets 'exp cov' to the length weighted median contig coverage, and 'cov cutoff' to half that value according to the velvet manual. I believe that the auto values are calculated using only the reads velvet uses in the assembly. This means that if you try to calculate a value for coverage based on the expected genome length and the number of bases in all your reads, you will end up with different values.

3. How does increasing minimum contig length affect your contig length distribution and N50

It increases the N50 value by a fair margin ~300, which makes sense since N50 is a pseudo median. Removing the floor/lowest values would cause the median to rise.