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?

From my assembly statistics data, when increasing the k-mer length and keeping other parameters the same, the number of contigs decrease and the contig length increases. The mean contig length increases and the N50 increases therefore showing that the distribution of contig lengths skews towards contigs of longer lengths when the k-mer length is increased. Longer k-mers lead to longer contigs.

1. 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’?

By increasing the sequencing depth of coverage cutoff, the contigs having lower depths of coverage are no longer being analyzed because they do not meet that minimum threshold. Therefore, by increasing depth of coverage cutoff, the data becomes more stringent requiring higher amounts of reads for each nucleotide and throwing out reads that do not meet that minimum threshold. Because of this, the number of contigs being analyzed would be expected to decrease. This is evident in my velvetg data as when the coverage cutoff is increased, the number of contigs decrease and the average depth of coverage and k-mer coverage increase.

The de Bruijin graph looks more straight forward and contains less possible side loops. In velvet, when a coverage cutoff is specified it wipes out any data not meeting that coverage. This would create a much more concise and streamlined assembly with less ambiguity. If a k-mer has a small portion of reads and doesn’t meet that coverage cut-off it will not be considered whereas without that increased coverage cutoff it would be and the graph would detail that additional loop.

Velvet calculates auto by setting the expected coverage to the length weighted median contig coverage and then dividing this value by 2 to get the coverage cutoff value. When I used the auto cutoff value, I achieved similar results to my 20x result.

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

By increasing the minimum contig length, the N50 and the mean contig length increase. This is because the contigs of lower lengths are being filtered and thus causing the lowest contig length to increase and skewing statistics such that the mean and median move upwards. This is true in this case because there are contigs lower than the 500bp metric used here as a length cutoff. But if there were originally no contigs below the contig cutoff length then the results would not change.