Michelle S. Massaquoi
PS6 Questions
7-19-19

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?

Based on the assembly stats we collected the trends are as follows:

N50 increased with kmer size.

The maximum contig length stayed around the same.

The mean contig length increases with kmer size.

The length of the genome decreases with kmer size.

The number of contigs decreases dramatically with kmer size.

The mean depth of coverage doesn't change.

The contig length distribution within the plots (all pdfs), shifts from a frequency primary at smaller contig lengths for smaller kmer sizes to an increased frequency at larger contig lengths from increased kmer sizes

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'?

Based on the assembly stats we collected the trends are as follows:

~the numbers for auto coverage and coverage of 20 were about the same for all stats

The N50 increased with increased coverage but not by a lot.

The max contig length decreases with increased coverage cutoff.

The mean contig length also stays about the same between 20x and 60x cutoffs.

The length of the genome decreases with increased coverage cutoff.

The number of contigs decreases dramatically with increased coverage cutoff.

The mean depth of coverage almost doubles with increased coverage cutoff.

The number of paths in the de Bruijin graph change when adjusting the coverage cutoff parameter.

Velvet automatically calculates the coverage cutoff by using half the length weighted median contig coverage depth. "It allows you to quickly obtain a decent assembly in your first run." (Page 9 of Velvet manual)

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

Comparing kmer size 49: 'auto coverage cutoff' versus 'auto coverage cutoff & min contig 500'

The N50 increased when we increased minimum contig length.

The contig length distribution has a frequency mostly within the 100s size of contig lengths and this distribution changed to a shift in frequency mostly within the 1000s size of contig length when we increased to 500 minimum contig length.

To turn in your work for this assignment, do the following:

Be sure to turn in your unit tests and expected results, your code, your output (mean, max, N50, etc.) and plots, as well as the answers to the questions above to GitHub.