Joshua Schaaf

Week 4 Summary

For this week, a portion of each of Venter's 21st chromosome (100,000bp) were obtained with samtools faidx. Grep was then used to check N count, making sure it was not high. Art_illumina was then used to create a set of fastq 100bp long reads at 10X coverage from each chromosome. These two files were concatenated, and then velveth was used to create k51 and k21 genome assemblies using single reads. Velvetg was then used to create De Bruijn graphs. This was repeated for paired-end reads, and dual paired-end reads.

Assembly Statistics:

	Single 21	Single 51	Paired 21	Paired 51	Dual-Paired 21	Dual-Paired 51
# Nodes	969	101	99	5	110	5
N50	1621	3170	65939	100477	98704	100369
Max	9615	9103	65939	100477	98704	100369
Total	104320	100922	100539	100559	101735	100451
Reads	0/19974	0/19974	19904/19960	17781/19960	39688/39906	35948/39906

The assembly that seems to work the best is the paired k51 assembly, as it has the largest contig compared to the total BP of the assembly. However, the dual-paired k51 assembly is most likely the best assembly, compared to the actual 100,000 bp chromosome files. In general, a higher k-value provided less ambiguity, creating more accurate assemblies, and the more reads available for velveth, in general, the more accurate assembly. Below is a Gephard graph between the original 100,000bp file, and the contig file from the dual paired-end, as well as it's De Bruijn graph.