Paula Berry

BI621 – Problem Set 6

Part 2, question 3.

a) 50 fosmids * 40 Kb = 2,000,000 nucleotides

b) C = N * L / G
G = 2,000,000 nt

Unmatched file is 62089591 nucleotides long.
62089591 / 2000000 = 31.0447955 expected coverage

fq_1 file is 68009851 nucleotides long.
68009851 / 2000000 = 34.0049255 expected coverage

fq_2 file is 67083157 nucleotides long.
67083157 / 2000000 = 33.5415785 expected coverage

c)$C_k = C * (L - K + 1) / L$
L = mean length of reads
K = 49 nt long

Unmatched file C_k = 106.495235
fq_1 file C_k = 133.891895
fq_2 file C_k = 129.258425

d)See assay results textfiles for contig.py results
kmer length 31, no cut off specified: kmer31_coveragenull.txt
kmer length 41, no cut off specified: kmer41_coveragenull.txt
kmer length 49, no cut off specified: kmer49_coveragenull.txt
kmer length 49, coverage cut off 20x: kmer49_cutoff20.txt
kmer length 49, coverage cut off 60x: kmer49_cutoff60.txt
kmer length 49, coverage cut off auto: kmer49_cutoffauto.txt
kmer length 49, coverage cut off 60x: kmer49_cutoffauto_mincontig500.txt

Part 3:

1. As the k-mer length increases, the total number of contigs decreases. This could indicate that the contigs are becoming longer and therefore there are fewer of them, however the maximum contig length is largest in the middle kmer value of 41. But the average contig length is largest for the largest kmer value of 49, indicating that even though the max contig length is smaller there are fewer very small contigs. Also, the total length of the assembled genome goes down as the k-mer length increases, but the coverage value goes up, and the N50 value also increases.

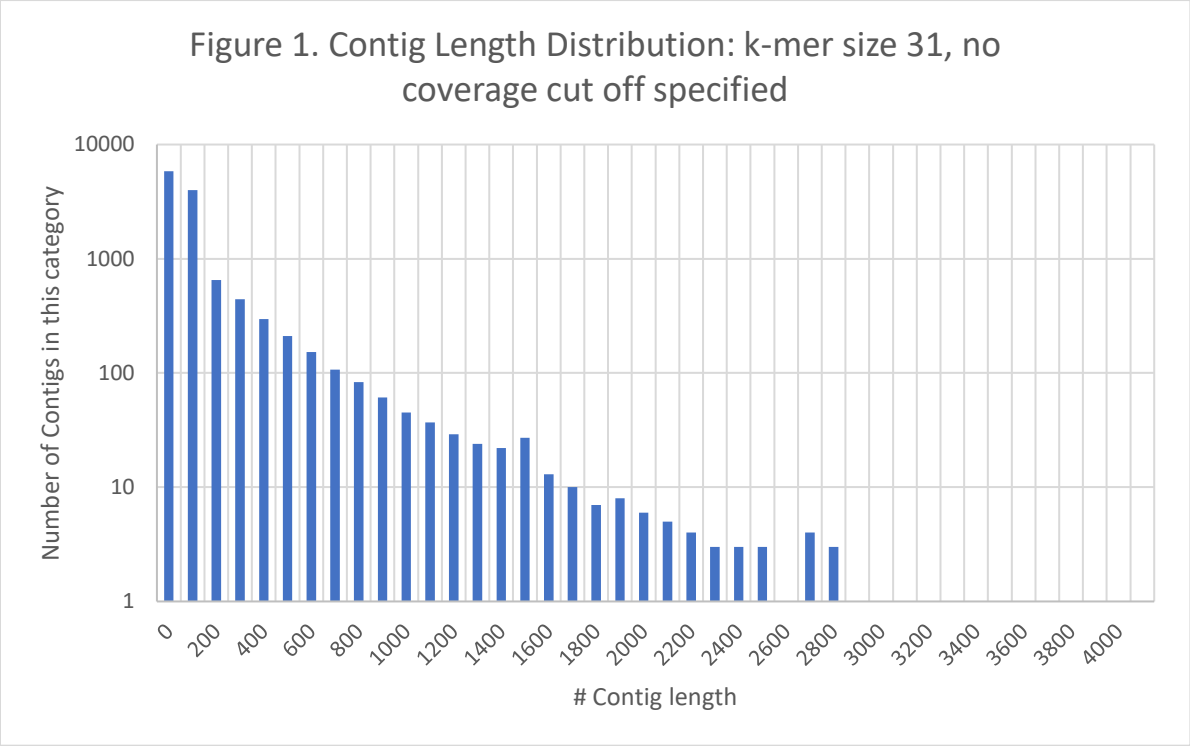
2. An increased coverage cut off affects the assembly in several ways – it sets a minimum amount of coverage and eliminates the short, low coverage nodes from the assembly. When it is set to "auto" Velvet sets the cutoff to half the weighted median contig coverage depth. The number of nodes in the de Bruijin graph is being reduced, reducing the possible contigs that can be made, but the contigs that are made have a higher degree of specificity.

3. Setting a minimum contig length drastically reduces the total number of contigs by eliminating the contigs below the set minimum. This has no effect on the max contig length, but the mean contig length is increased. The total length is decreased as there are fewer contigs, and the mean depth of coverage goes down slightly. However, the N50 increases also. As with most things in genome assembly, it seems to be a balancing act between setting parameters to having enough breadth to capture "true" rare sequences in a genome while eliminating bubbles from sequencing errors or repeated sequences.

Table 1. Values returned by contigs.py for various Velvet settings run on 3 fastq files of the same ~2000000 bp genome, one unmatched and two paired-end reads.

kmer size	cutoff	minimum contig	total contigs	max contig length	mean contig length	total length	mean depth of coverage	N50
31			12080	4059	182.4668046	2204199	37.25694452	296
41			5369	7451	334.8854535	1798000	37.2326769	789
49			3198	5032	503.627267	1610600	41.47958885	1086
49	20		1226	10467	910.8662316	1116722	62.27950573	1803
49	60		422	7268	939.4146919	396433	123.6534212	2105
49	auto		1344	10467	888.7113095	1194428	58.56039447	1784
49	auto	500	625	10467	1644.4064	1027754	49.78496699	2012

Contig Length distribution graphs with varying Velvet parameters. (note: logarithmic Y axis)



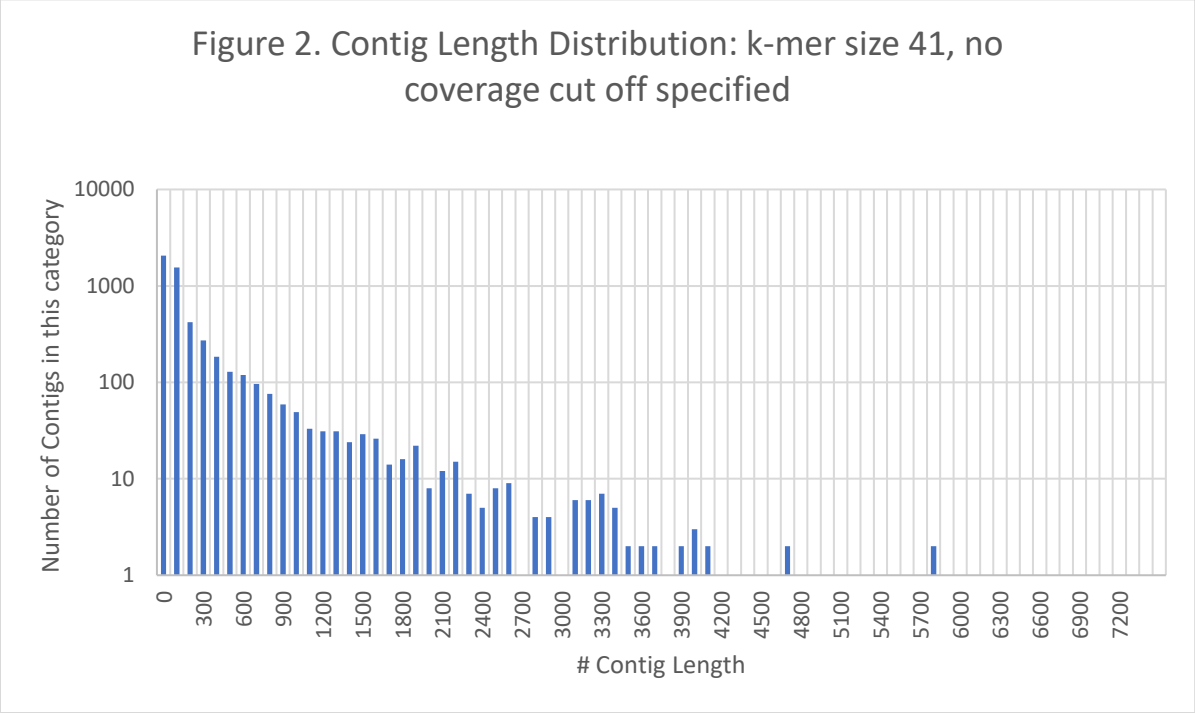


Figure 3. Contig Length Distribution: k-mer size 49, no coverage cut off specified

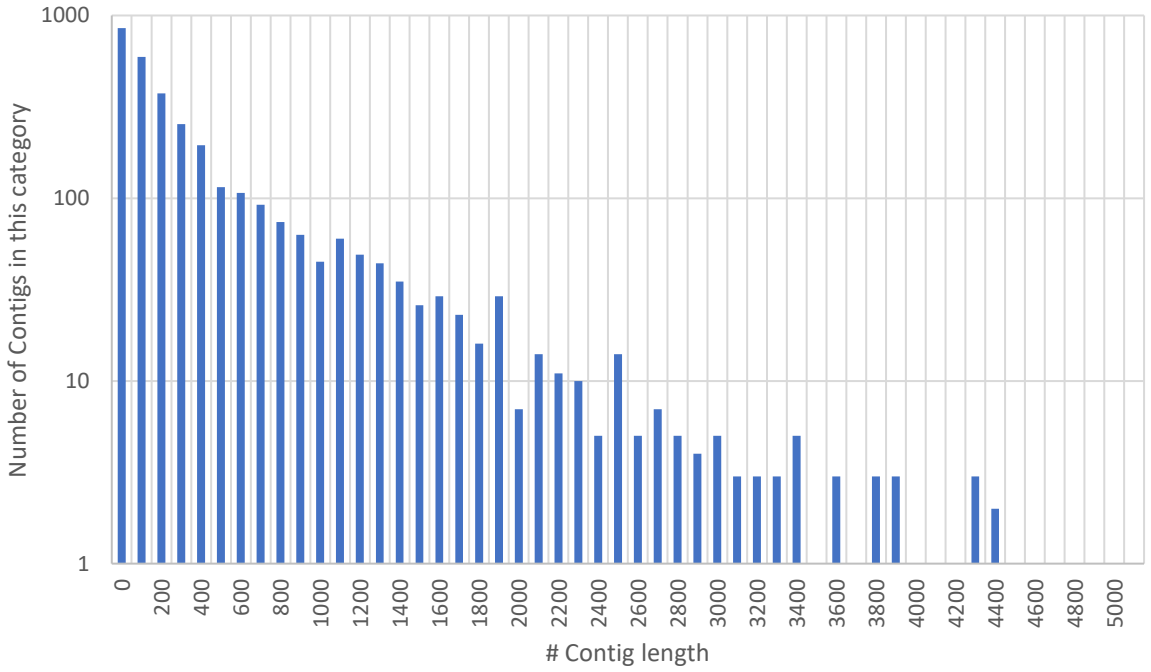


Figure 4. Contig Length Distribution: k-mer size 49, 20x coverage cut off

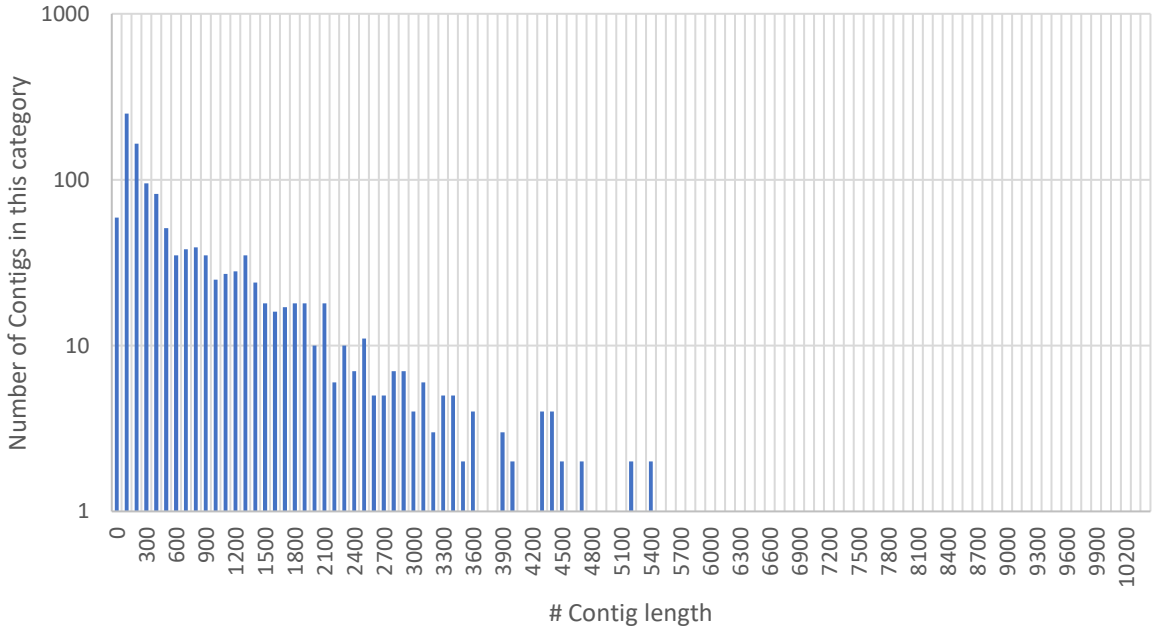


Figure 5. Contig Length Distribution: k-mer size 49, 60x coverage cut off

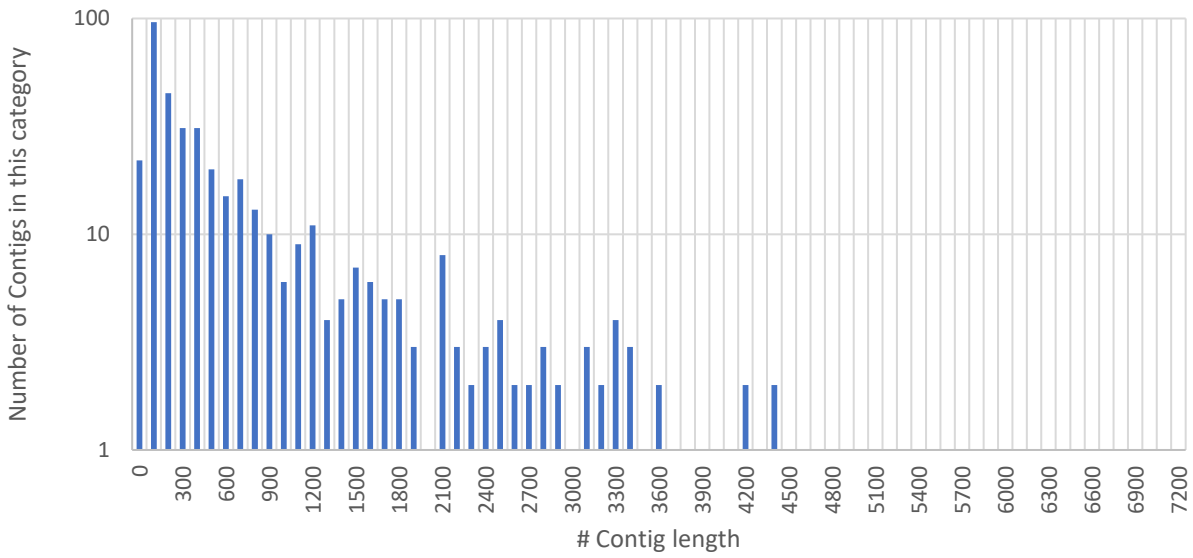


Figure 6. Contig Length Distribution: k-mer size 49, automatic coverage cut off

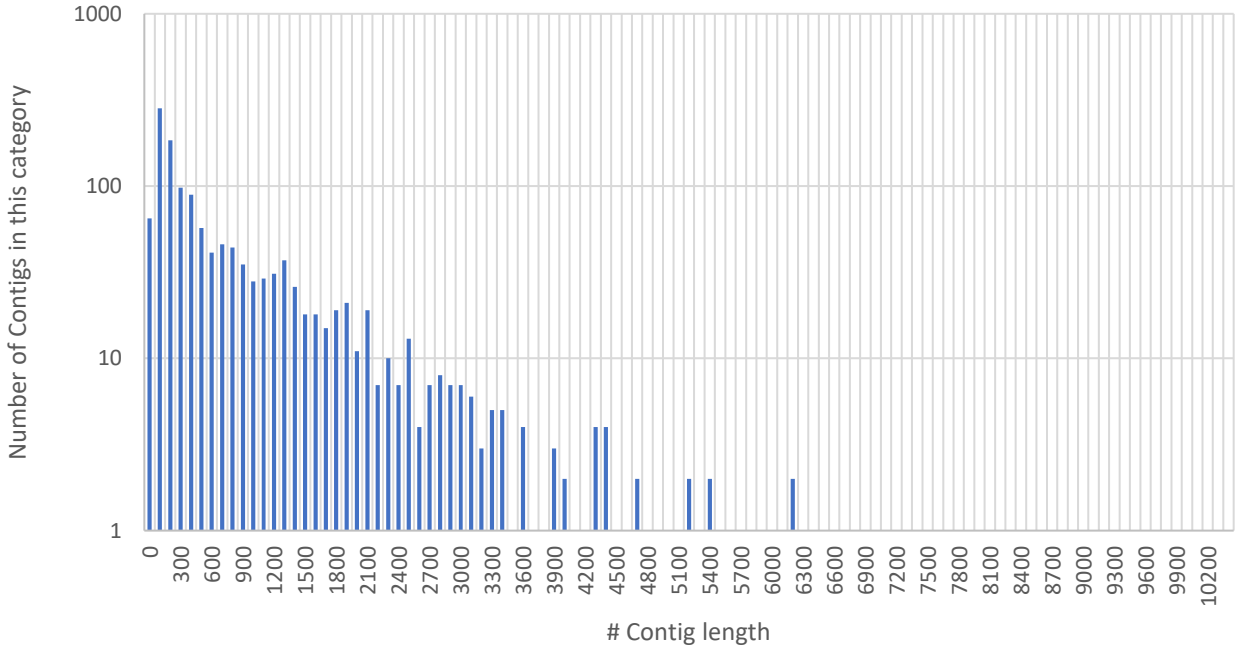


Figure 7. Contig Length Distribution: k-mer size 49, automatic coverage cut off, minimum contig length 500

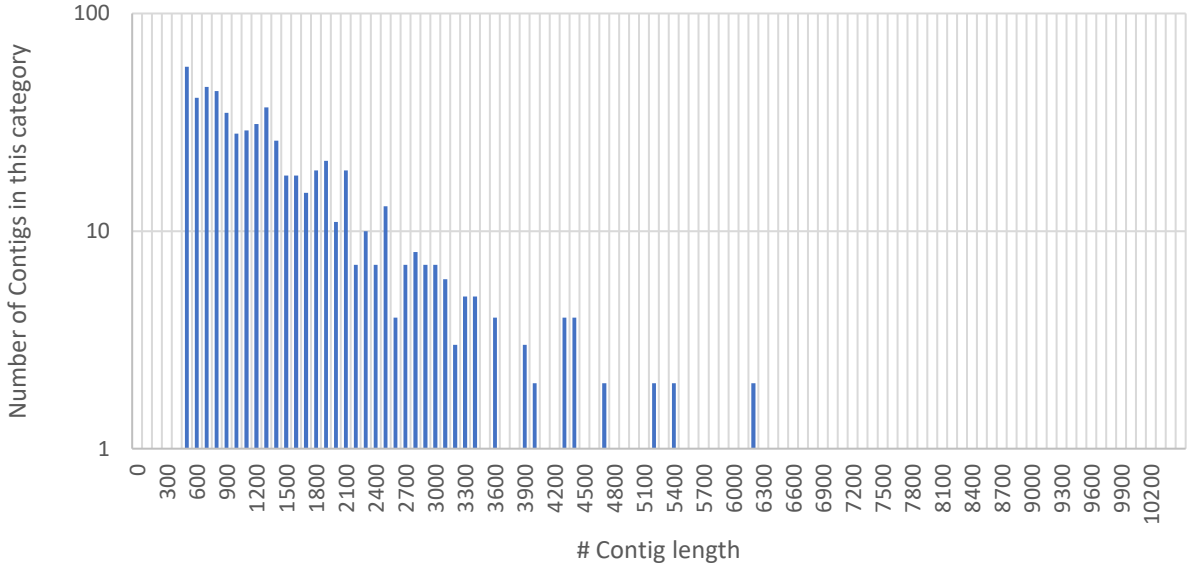


Figure 8. Contig Length Distribution of contigs.fa test file: k-mer size 49