Khan Inan
Assignment 14

1)

	contigs	avg contig length
100	100	0.333
1000	10	0.033
10000	1	0.003

2) What is an appropriate N (or coverage) for this project in order to assemble the genome, in your opinion ?

I believe the lower the N the better it is, just due to contig values being higher. It does seem to be inversely proportional in the sense that a lower N value results in a larger contig and avg contig length for these specific values

3) If L=200 bp, how does it affect the numbers just calculated in 1 ?

If L is 200 then the average contig length increases. the values go from 0.333, 0.033, 0.003 to 0.5, 0.055, 0.005.

4) If G=1,000,000 bp long, how would that affect your answers ?

If the genome increases to 1,000,000 then I would see my contig column increasing by a factor of 100. similarly if the G was halved then the contig values would also half. This means that the Genome is directly proportional to contig .

5) Why do you get numbers that seem to be nonsense (less than 1 contig, and average contig lengths growing beyond length of genome ?

This is mostly likely due to the fact that it is very rare to calculate for N values that are close to or equal to the value of G. And so in this case since we have an N value of 10,000, that makes some of the values no make any sense. Similarly, there is usually a cutoff for N values where anything less than a certain value makes the contig avg greater than 1. In this case, since the L value was sufficiently large it wasn't the case.

6) If the genome has many repeats (say 20% of genome). How will your sequencing strategy change, in terms of L and N if a) repeats are 20 bp long b) repeats are 120 bp long ?

I would say the sequencing strategy would change with equivalent ratios. The final values would be ⅛ of what they are now if it were 20, and then 6 times that if they were 120

Seq1:
TGAACGCGCCCGATCTCGTCTGATCTCGGAAGCTAAGCAGGGTCGGGCCTGGTTAGTACTT
GGATGGGAGACCGCCTGGGAAT
Seq2:
GAGATTTCCCAAGGCTGACTTTACAGAGATTTCCAAGATAGTGACAGATCTTGCAAAAGTCC
ACAAGGAATGCTGCCATGGTGA
Use the ucsc browser http://genome.ucsc.edu to map the two reads to the human genome and

find the 1) underlying gene,
The underlying gene for these two sequences is human hg3

and 2) the chromosome and position of the map location
The chromosome for these two sequences is chr1_KZ208906v1_fix
and the position for the map location 183106 - 183188 for the first and
73412017 - 73412100 for the second

Take two columns from the mRNAseq dataset (Mnemiopsis_count_data.csv)
Plot the cumulative distribution (x axis is value, y axis is number of genes)
Hint : use binning
Is there anything you can guess about the samples from the distribution ?
From plotting the distribution we can see that the samples are closely concentrated towards the x-axis of the graph

Perform scaling normalization (totals in each column are same) and plot the cumulative distribution,
Is this informative about the groups in the data ?
The cumulative distribution is not that useful for this particular set of data because there are no obvious clusters ti be found
Perform quantile normalization and plot the cumulative distribution (the name quantlie come from the fact that on a q-q plot (try this), they lie on a diagonal, all distributions have the same quantile distributions

For this graph although there are outliers on the Q-Q plot, most of the values do align with the diagonal of the plot.