Software Challenge

Genome Assembly

Owais Sarwar

Fall 2020

1. The N75 statistic is the maximal contig length for which all contigs greater than or equal to that length comprise at least **75%** of the sum of the lengths of all the contigs.
2. **N50 = 100** (100 + 200 = 300 = 600 \* 50%); **N75 = 60** (60 + 60 + 80 + 100 + 200 = 500 > 450 = 600 \* 75%)
3. **NG50 = 60** (60 + 60 + 80 + 100 + 200 = 500 = 1000 \* 50%)
4. **NGA50 = 50** (New Contigs = [20, 20, 30, 30, 50, 50, 60, 60, 80, 200]; 200 + 80 + 60 + 60 + 50 + 50 = 500 = 1000 \* 50%)
5. First, we could use the total estimated length of the genome to determine the approximate length of the gaps between the contigs in the scaffold (as well as the placement of the contigs, if we have a reference). We would also use the sequence and length of the contigs as key pieces of information to figure out where to place them in the scaffold (e.g. by comparing it to the reference), since the contigs’ sequence is the key informative characteristic.

|  |  |  |  |
| --- | --- | --- | --- |
| K | N50 | #LC | TLoLC |
| 25 | 59595 | 110 | 2802857 |
| 55 | 159616 | 38 | 2821839 |
| 85 | 188896 | 37 | 2825752 |

2. The k=85 sample performed the best because it had the highest N50, the lowest number of long contigs, and the greatest total length of long contigs. This likely performed the best because long kmers achieve longer contigs (which correlate favorably with all of the statistics mentioned) since each kmer contains more information that would avoid the uncertainties that shorten contigs.
3. Assuming the rest of the questions are asking for the k=85 case only, there were 29 misassembles.
4. The number of misassemblies has a significant effect on the resulting assembly. We can see (as shown in the results for the next question), NGA50 is reduced significantly from the NG50.
5. NG50 = 202267 ; NGA50 = 87161
6. The NGA50 is understandably smaller than the N50 because the NGA50 statistic has broken the contigs at their misassembly points, resulting in more smaller contigs which bring the NG50 statistic down. The NG50 statistic is actually higher than the N50 statistic, which must be because the assembly length is greater than the reference genome length