UC San Diego

Assignment No.2 – Arsham Mikaeili Namini

N50 statistic: N50 is a statistic that is used to measure the quality of an assembly. N50 is defined as the maximal contig length for which all contigs greater than or equal to that length comprise at least half of the sum of the lengths of all the contigs. For example, consider the five toy contigs with the following lengths: [10, 20, 30, 60**,** 70]. Here, the total length of contigs is 190, and contigs of length 60 and 70 account for at least 50% of the total length of contigs (60 + 70 = 130), but the contig of length 70 does not account for 50% of the total length of contigs. Thus, N50 is equal to 60.

NG50 statistic: The NG50 length is a modified version of N50 that is defined when the length of the genome is known (or can be estimated). It is defined as the maximal contig length for which all contigs of at least that length comprise at least half of the length of the genome. NG50 allows for meaningful comparisons between different assemblies for the same genome. For example, consider the five toy contigs we considered previously: [10, 20, 30, 60, 70]. These contigs only add to 190 nucleotides, but say that we know that the genome from which they have been generated has length 300. In this example, the contigs of length 30, 60, and 70 account for at least 50% of the genome length (30 + 60 + 70 = 160); but the contigs of length 60 and 70 no longer account for at least 50% of the genome length (60 + 70 = 130). Thus, NG50 is equal to 30.

1. **Based on the above definition of N50, define N75.**

   N75 is defined as the maximal contig length for which all contigs greater than or equal to that length comprise at least 3/4 of the sum of the lengths of all the contigs. For example, consider the five toy contigs with the following lengths: [10, 20, 30, 60, 70]. Here, the total length of contigs is 190, and contigs of length30, 60 and 70 account for at least 75% of the total length of contigs (30 + 60 + 70 = 160), but the contig of length 60 + 70 does not account for 75% (142.5) of the total length of contigs. Thus, N75 is equal to 30.

2. Compute N50 and N75 for the nine contigs with the following lengths:

   [20, 20, 30, 30, 60, 60, 80, 100, 200].

50% of Sequence = 300 N50: 100

75% of Sequence = 450 N75: 60

3. Say that we know that the genome length is 1000. What is NG50?

   NG50 = 60

4. If the contig in our dataset of length 100 had a misassembly breakpoint in the middle of it, what would be the value of NGA50?

   NGA50 = 50

5. Based on the definition of scaffolds, what information could we use to construct scaffolds from contigs? Justify your answer.

   - Additional long reads could be generated in an attempt to find reads that bridge the gaps in contigs. In other words, if we find a long read that begins at the end of contig A, and ends at the beginning of contig B, then we can conclude that the read extends across the gap between the contigs.
   - Contigs could be compared against a "reference genome", i.e., a complete genome sequenced from the same species (often at greater cost). The order of the contigs in the reference genome would indicate the order of the contigs in the desired scaffold.
   - Information from read-pairs could be used. In particular, if the first read in a read-pair maps to contig A, and the second read in a read-pair maps to contig B, and we know the distance between the paired reads, then we can infer the distance between contigs A and B. By gathering this information for different pairs of contigs, we may be able to infer distances between contigs and therefore their ordering with respect to each other.

6. fill in the 9 missing values in the following 3 x 3 table:

k	**N50**	**#long contigs**	**total length of long contigs**
25	59,595	110	2,802,857
55	159,616	38	2,821,839
85	188,896	37	2,825,752

7. Which assembly performed the best in terms of each of these statistics? Justify your answer. Why do you think that the value you chose performed the best?

The total length of long contigs is about the same for all three values of k. Accordingly, we conclude that the assembly using k = 85 performed the best because it has a larger value of N50 and fewer contigs than k = 25, while having the same number of contains as k = 55.

k = 85 performs the best because if the reads are too short (k = 25 or 55), then the reads contain too little information, and repeats may make it difficult to identify where a read came from.

8. When you increase the length of k-mers, the de Bruijn graph _____.

Becomes less tangled.
We saw in the class text (and lecture) that increasing the value of k used to generate k-mer reads led to a less tangled de Bruijn graph because the larger the value of k, the greater the amount of information contained in our reads, and the lesser the effects of repeats.

9. Aanswer the following questions using the QUAST reports.

1. How many misassemblies were there?

2. How significant is the effect of misassemblies on the resulting assembly?

3. What are NG50 and NGA50 for the QUAST run?

4. How do they compare with the value of N50 that you previously calculated? Why?

1. 29 in k=55,85 and 24 in k=25

2. in low k-mers the misassemblies starts in high coverage of genome but in higher k-mers, it starts in low coverage of genome reads, and obviously the number of misassemblies increases

3. NG50: (k=25,75624). (k=55,165671). (k=85,202267) NGA50: (k=25,35369). (k=55,92194). (k=85,87161) they are at least 50% of genome length that are made with greater contigs

4. NG50 is larger than the N50 value previously obtained. (Note: it roughly corresponds to the value of N50 that was previously obtained, because the contigs generated cover the entire genome.) However, NGA50 is about half as large as N50 because of the effects of misassemblies.