BINF7001 ASSESSMENT 3 (2024) | De novo genome assembly and phylogenomics

1. Genome data and assembly

1.1 Sequencing data

My dataset is ERR9872452. It was sequenced using the ILLUMINA NextSeq 500 sequencing platform. It produces 2x 150bp reads, with a total read count of 2197746. The G+C content is approximately 38%.

1.2 Genome assembly

I used the velvet program to assembly the genome. Velvet requires a kmer parameter to be selected. I used the jellyfish program to count various kmers and graph them. Values tested were k=11, 17, 21, 27, 31, 51.

By visual inspection, either k=21, k=27 or k=31 would be good candidates. The sharp peak and long tail of k=51 could capture too many long, infrequent occurring kmers, leading to not enough overlaps. Too short would lead to ambiguity, with too many overlaps that may be incorrect or not meaningful.

I then used velveth to generate the metadata for the alignment, and velvetg to do the assembly.

I ran this pipeline for the proposed kmer values:

kmer	N50 scaffold length	Maximum scaffold length	Total scaffold length	Total number of scaffolds	Config mean length	% assembled
k21	82/9.227 KB	60.132 KB	2,514,955	1,082	99.98%	2060637 / 2197746 (93.76%)
k31	36/22.561 KB	105.043 KB	2,532,370	521	99.99%	2188030 / 2197746 (99.56%)
k41	391/2.038 KB	10.011 KB	2,663,113	3,166	100.00%	2160007 / 2197746 (98.28%)

kmer	N50 scaffold length	Maximum scaffold length	Total scaffold length	Total number of scaffolds	Config mean length	% assembled
k51	44/18.254 KB	54.926 KB	2,506,012	464	99.99%	2165246 / 2197746 (98.52%)
k61	55/14.643 KB	45.492 KB	2,455,163	401	99.85%	2134797 / 2197746 (97.14%)
k71	38/20.195 KB	65.611 KB	2,439,224	254	99.83%	2106757 / 2197746 (95.86%)

I settled on k=31. Here's why:

- Has a high level of completeness (99.56%).
- While the number of scaffolds that contribute to 50% of the length is low (36), the 50th largest is 22.563KB. This means we have a smaller number of longer scaffolds - this indicates larger, more complete scaffolds.
- The L50 value is very large (105KB). Although only one scaffold, this metric supports the 36/22.561KB figure, and is further evidence for k=31, the final assembly is not fragmented.

2. Ab initio gene prediction

(a) Brief description of your approach

I used the mat file from the practical. It is for Staphylococcus Aureus; my sample is Staphylococcus Pseudintermedius, which is known to be very genetically similar and have many common proteins. I would like to build my own matrix file and see how that compares, if time permits. https://github.com/kuleshov/nanoscope My command:

```
genemark -opn -m /opt/BINF7001/2024/Prac8_2024/Staph_aureus_JKD6008.mat ./k31/contigs.fa
```

According to GeneMark, the GC content is 37.5% - this aligns with what we expected.

(b) Total number of predicted genes

The total number of predicted genes is 4560, found using the following command:

```
cat k31/contigs.fa.orf | grep "^>" | wc -l
```

(c) Average gene length

The average length of the predicted genes is 302 amino acids. This was derived using python from the `protein.fa.orf`, which the proteins as extracted from the GeneMark `configs.fa.orf`.

(d) Length and function of the longest gene

Longest gene is 1571 amino acids. I ran a BLAST on the protein on NCBI and the best matches suggest this protein is either

- LPXTG-anchored putative endo-alpha-N-acetylgalactosaminidase SpsG
- YSIRK-type signal peptide-containing protein

Both have been suggested to help secure surface proteins to the cell wall in gram-positive bacteria (Bae, 2003).

3. Phylogenomic analysis (maximum two pages; 8)

(a) Brief description of your approach (name(s) of program(s), key parameters used)

(b) Total number of homologous protein groups, and number of single-copy groups

(c) 16S rRNA gene tree (8 taxa, rooted using outgroup)

(d) protein tree of the chosen housekeeping gene (8 taxa, rooted using outgroup)

(e) one key difference/similarity between the two trees

(f) one plausible explanation as to why such a difference occurs

References

Bae, T., & Schneewind, O. (2003). The YSIRK-G/S motif of staphylococcal protein A and its role in efficiency of signal peptide processing. Journal of bacteriology, 185(9), 2910–2919. https://doi.org/10.1128/JB.185.9.2910-2919.2003