Reproducing "Completing Bacterial Genome Assemblies with Multiplex MinION Sequencing" by Wick et al. 2017

Alex Baryiames, April 30, 2019

Abstract

In the study of infectious bacteria, it is important to understand the structure of their genome in order to design therapeutic treatments for patients. Antibiotic resistance genes are often flanked by repetitive sequences that make it difficult to identify their location within the bacterial genome. This issue arises because Illumina are physically short, which makes it hard to arrange the fragments of reparative sequences correctly in whole-genome sequencing. Researchers Wick et al. sought to determine if nanopore data is suitable to accurately construct a genome by comparing three different assembly types: Nanopore, Illumina, and Hybrid assembly. Nanopore data was assembled using Canu, and the hybrid and Illumina data was assembled using Unicycler. These methods produced a .gfa file containing the structural assembly data that made it possible to construct the assembly of the *K. pneumonia* genome in Bandage. I was able to reproduce the findings of the paper, concluding that a hybrid assembly containing nanopore long reads with short Illumina reads was able to assemble the structure of the genome with 99.3% accuracy.

Introduction

The objective of "Completing Bacterial Genome Assemblies with Multiplex MinION Sequencing" was to assess three different methods for whole-genome construction. The researchers investigated whether the assembly of a genome using Illumina only, Nanopore only, or a hybrid assembly of both would be the most successful at determining the genomic structure of *K. pneumonia* (Wick et al. 2017). Success was measured by percent identity to a reference genome, and physical continuity to a known genome structure. The organism studied in this paper is *K. pneumoniae,* which is known for causing pneumonia in humans. Sequencing the DNA of human pathogens has become a routine procedure for researchers due to the pharmaceutical implications and the development of personalized medicine. Understanding how pathogens operate is vital to designing new drugs and treating symptoms (Wong et al. 2015). For example, if doctors knew that a patient's bacteria were resistant to a certain antibiotic

prior to administering it, then treatment options could be tailored to what would be the most effective in that particular situation. However, it is difficult to discern where these sequences are because they are typically flanked by repetitive insertion sequences. Short reads are incapable of providing the data required to located them in a genome assembly because of their similarity to other sequences (Gurevich et al. 2013). This makes it impossible to know whether the resistance genes are located within the main genome, or a smaller plasmid (He et al. 2017). However, Oxford Nanopore Technology (ONT) is able to achieve ultra-long reads capable of placing these genes but has low accuracy when compared to Illumina sequencing. Therefore, if Illumina short reads and ONT long reads were combined, it would be possible to have the accuracy of sequencing by synthesis and the resolution of ONT. In my analysis, I aimed to recreate the results of a paper analyzing the generation of genome assemblies using various techniques. The results of the paper comparing assemblies using hybrid, nanopore, and Illumina sequencing will be tested for reproducibility by running their raw data through their proposed pipeline.

Methods

The data selected for this study was base-called nanopore data and trimmed Illumina reads published in the supplementary because albacore was not available to trim the adapter sequences. In the original study the nanopore sequences were barcoded, so Porechop v0.2.1 (https://github.com/rrwick/Porechop) was used to remove adapter sequences. To complete the nanopore assembly, Canu v1.5 (https://github.com/marbl/canu/tree/f356c2c3f2eb37b53c4e7bf11e927e3fdff4d747) was used to assemble the chopped sequences. Instead of using nanopolish, Minimap https://anaconda.org/bioconda/minimap) was used to find the mapping positions of long sequences and Racon (https://github.com/isovic/racon) was used in conjunction with reference data as an error correction tool to polish the assembly.

Next, trimmed Illumina reads Barcode01_1 and Barcode01_2 were assembled with Unicycler v0.4.0 (https://github.com/rrwick/Unicycler). Polishing was not required because Unicycler automatically preforms a SPAdes read correction. The hybrid assembly was also generated using Unicycler, except the chopped nanopore reads were used as a template for the short Illumina reads. Accuracy was assessed using dnadiff in the MUMmer3 package (https://github.com/marbl/MUMmer3) as well as quast (https://github.com/ablab/quast). For visualization of the assembly graph Bandage was used (https://github.com/rrwick/Bandage).

Results

Summary Statistics

Assembly with all three methods produced results comparable to the original study. ONT only assembly had the lowest percent identity to the reference genome, whereas the hybrid assembly had the highest. While this is true, the hybrid and ONT assembly only differed in accuracy by .9% which is most likely because the quality of the raw nanopore reads was high with over 200x coverage (Figure 1). This was expected because measuring the accuracy of these assembly methods was not the primary purpose of the study. However, it is important to understand that the data used for the study is accurate so that the results of the results from the assemblies are accurate and justified.

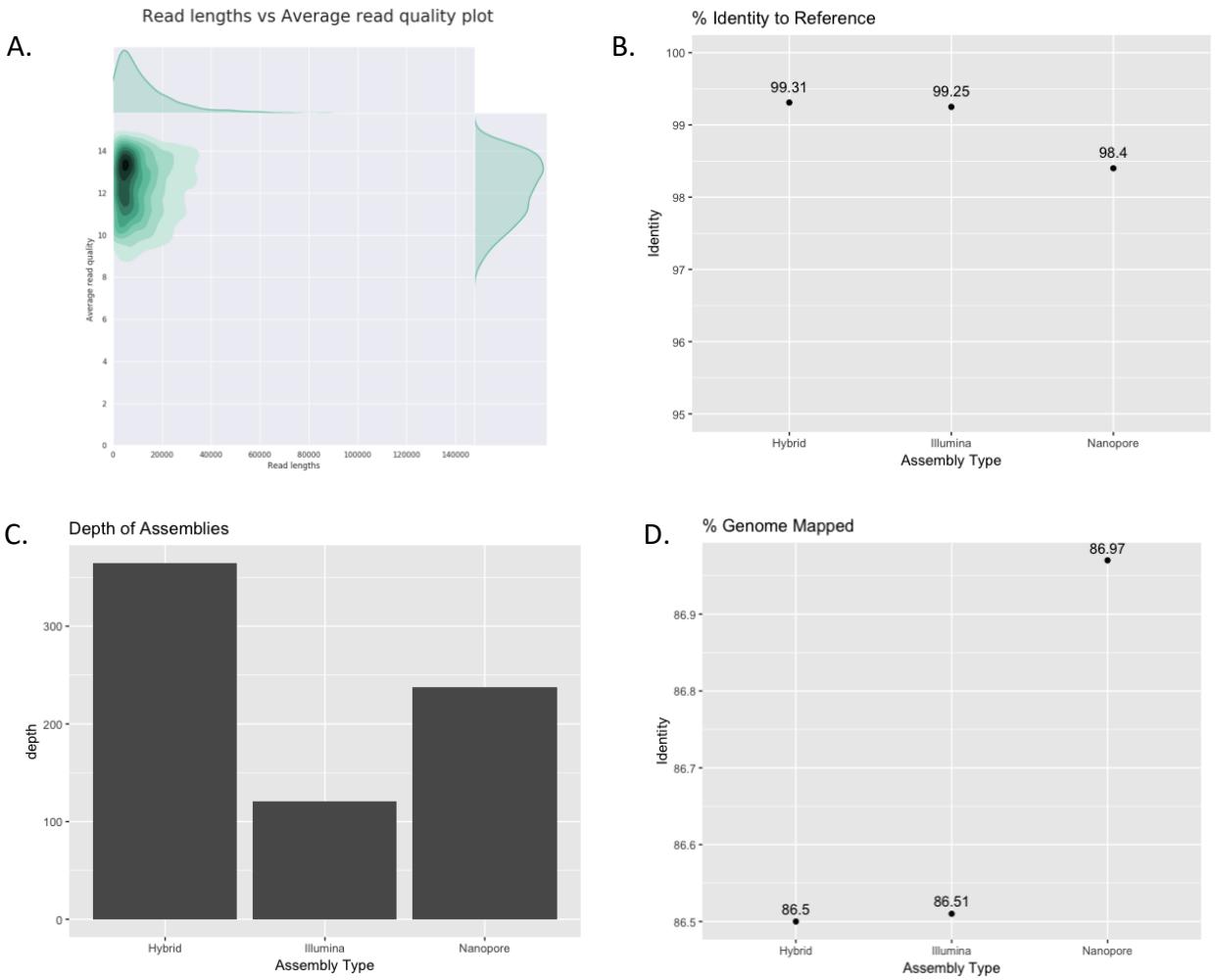

Figure 1.

A) The analysis results of Nanoplot on the raw ONT reads. The mean read length was 13,976.8 bp with a mean quality of 12.1. B) The identity to the reference genome generated by dnadiff. C) Sequence depth generated by quast. D) Percent of the genome mapped determined by the total number of aligned bases in the genome divided by the genome size.

Bandage Assembly

Bandage was run in order to determine which method produced the best alignment genome assembly. The results were consistent with the original study. Illumina only assembly produced a graph with many intersecting regions because of the inaccurate mapping of repeating or homologous regions (Figure 2A). The more repeating regions there are, the more tangled the graph becomes. This is because short reads don't have enough information to resolve where these repeating regions are located. Without a template, the assembly software aligns these regions to multiple locations in the genome. The Nanopore only assembly produced a much cleaner assembly. It was successful in finding two plasmids, however, there were two repeating regions in the genome that the assembler was unable to resolve (Figure 2B). There was also a short sequence in one of the plasmids that resulted in a non-circular assembly. Consistent with the findings of the paper, the hybrid assembly produced the best results. Unicycler was capable of producing the *pneumonia* genome without any mismatched sequences and successfully identified the two plasmids (Figure 2C). I hypothesize that the mistakes in oxford only assembly was caused by inaccurate sequencing because the mistakes were corrected once Illumina data was introduced. If the accuracy of nanopore sequencing was on par with Illumina, then the genome structure should have been assembled without issue.

A. B. C.

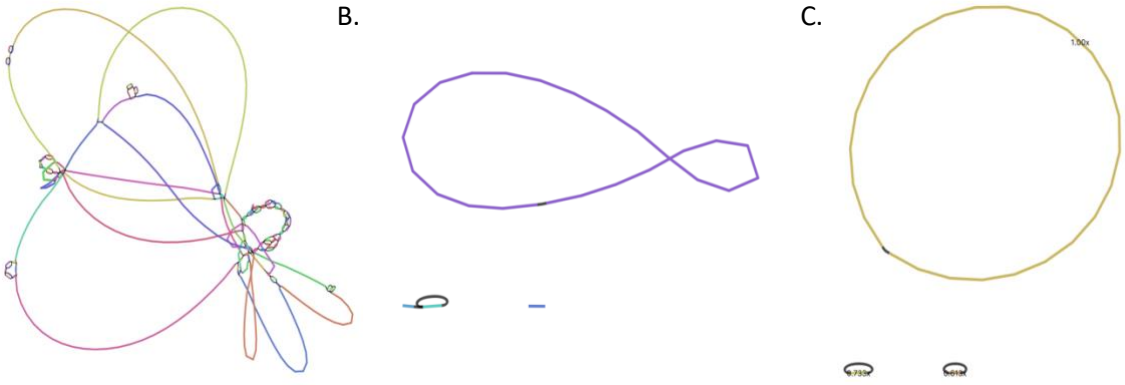

Figure 2.

From left to right: Illumina assembly graph, Nanopore assembly graph, Hybrid assembly graph. Colors are represented by unique contigs.

While the findings of my study were consistent with the original paper, the assembly graphs differed slightly. This could be due to the fact that I did not use their script designed to subsample the Illumina and ONT data to exclude low quality data. Ultimately, my assemblies had more contigs present so this may be responsible for the discrepancy in the Illumina assembly graph. Furthermore, I used minimap and racon to polish my Illumina data instead of Nanopolish. However, the difference in polishing steps was not significant because the nanopore data quality was so high to begin with. In conclusion, my findings were consistent with the original research paper. While the accuracy of Illumina data is high, the reads are too short to arrange into a single continuous genome. On the other hand, data produced by ONT is still too inaccurate to completely resolve the structure of a genome. Therefore, hybrid assemblies are required to produce assemblies accurate enough to identify resistance genes in bacteria.

References:

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E., 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial genomics, 3(10).

Inouye, M., Dashnow, H., Raven, L.A., Schultz, M.B., Pope, B.J., Tomita, T., Zobel, J. and Holt, K.E., 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome medicine, 6(11), p.90.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), pp.1072-1075.

He, S., Hickman, A.B., Varani, A.M., Siguier, P., Chandler, M., Dekker, J.P. and Dyda, F., 2015. Insertion sequence IS26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. MBio, 6(3), pp.e00762-15.

Wong, V.K., Baker, S., Pickard, D.J., Parkhill, J., Page, A.J., Feasey, N.A., Kingsley, R.A., Thomson, N.R., Keane, J.A., Weill, F.X. and Edwards, D.J., 2015. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter-and intracontinental transmission events. Nature genetics, 47(6), p.632.