Eddie Cho
BIOMI 609
Spring 2022

# Midterm Part 1 Summary

The three different genome assemblers, Velvet, SPAdes and ABySS were used to benchmark the performance. Instead of Unicycler, ABySS was used since the Unicycler caused the Python package along with gcc C++ compatibility issues on XSDE Virtual Machine on Jeststream cloud hosted by Indiana University. The reference genome fasta file was obtained from NCBI (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta). The NCBI indicates that it is a complete genome of Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1. The sequencing data implemented on the three genome assemblers was SRR17854410 (https://www.ncbi.nlm.nih.gov/sra/SRX14015288[accn]). The NCBI indicates that it is a sequencing of SARS-CoV-2 infections in Montana in 2020 and 2021. Then, the sequencing data (SRR17854410) was further split into two fastq files using the SRA toolkit. To streamline the benchmark, SRR17854410_1 was implemented on the three genome assemblers. For Velvet, the two step process took most of the time. Once installed, the *velveth* was performed resulting in three files: Log, Roadmap, Sequence. Then, *velvet* was performed yielding a final product or assembled file. Lastly, the QUAST (Quality Assessment Tool for Genome Assemblies) was utilized to compare the assembled file of SRR17854410_1 and the reference genome. For setting parameters, contigs of size were adjusted to 20bp. The QUAST assessment produced a report as an interactive html file containing details of two compared sequencing data along with a graph. According to the report, the numbers of fully unaligned contigs are 4716 contigs with the fully unaligned length that is 323110. The second genome assembler, SPAdes (St. Petersburg Genome Assembler) was easiest to implement. Specifically, this one step process

was easy to install the package and run the assembly. The package comes with a self checking test where a user can test if the package has been correctly installed and assembled. Similar to Velvet, the assembled sequencing data of SRR17854410_1 was adjusted to 20bp for contigs of size. Then it was compared with the reference fasta on QUAST. According to the report, the numbers of fully unaligned contigs are 10 contigs with the fully unaligned length that is 1540. Implementing the ABySS assembly took the most time because of the installation. In the end, I found out the shell had run on the env setting all the time. Once assembled, the assembled sequencing data of SRR17854410_1 was adjusted to 20bp for contigs of size just like Velvet & SPAdes. According to the report, the numbers of fully unaligned contigs are 19 contigs with the fully unaligned length that is 2495. To benchmark the performance of three genome assemblers, the table shown below would be helpful to see clear differences all based on the QUAST reports.

| | **# of Fully Unaligned Contigs** | **Fully Unaligned Length** | **Mismatches per 100 kbp** |
|---|---|---|---|
| **Velvet** | **4716** | **323110** | **4395.6** |
| **SPAdes** | **10** | **1540** | **61.82** |
| **ABySS** | 19 | 2495 | 73.06 |

Since the two compared sequencing data is SARS-CoV-2, I decided to focus on unalignment & mismatches per 100 kbp. In other words, these two were the parameters that I observed the performance. My hypothesis was the better the *de novo* assembler is, the less unalignment & mismatches per 100 kbp it produces. The table clearly indicates the Velvet has the highest numbers of unaligned contigs, highest fully unaligned length and highest mismatches per 100 kbp as noted in red color. On the other hand, SPAdes has the least numbers of unaligned contigs, least fully unaligned length and least mismatches per 100 kbp as noted in blue color. The ABySS

fits just in the middle in terms of numbers of unaligned contigs, fully unaligned length and mismatches per 100 kbp. Based on the table, the conclusion would be SPAdes is the best assembler while Velvet is the worst. The AbySS sits between Velvet and SPAdes.

As I worked on this, I made interesting observations and led me to thoughts why the Velvet is the worst assembler. The Velvet was first released in 2008. Based on the manual link, the last release was 2011. Considering the first iPhone was introduced in 2007, it might have been the newest assembly technology at that time. However, better technology emerged along with more complex problems. On the Github page, the AbySS was last released and/or modified in 2019. The SPAdes was first released in 2012 and continued to be developed as the latest release is 2022. The SPAdes is developed by the Center for Algorithmic Biotechnology in St.Petersburg State University located in Russia. Therefore, it is possible that the dedicated researchers with the support of the whole university continue work on the assembler bringing the best results. To sum up, there is no one-size-fits-all solution when it comes to gene assembly. It is upto a user to find best solutions.