Lab 2 – Genome Assembly with SPAdes and Assembly Evaluation with Quast

***Introduction***:

Assembling a genome can be compared to piecing together a jig-saw puzzle, except when you empty the box you realize the same puzzle is represented in the box multiple times. When the first human genome was assembled it required approximately 10,000 hours of computation. Since that time, computational speeds have increased and the algorithms improved so that the time required to assemble a genome is measured in minutes, not hours. Many approaches exist for assembling a genome, and often the approach is determined by the type of data that is collected. Genome size, error rate, and the type of technology used to sequence the genome are all factors that impact how to assemble the genome. One approach to assembling a genome is by using the software SPAdes. SPAdes is a genome assembly algorithm that can be used for single cell and multi-cellular bacterial sequencing. It is versatile due to its compatibility with many data types. It is compatible with sequence data sets created from Ion Torrent, PacBio, Oxford Nanopore, and Illumina. It can assembly genomes by single end, paired end, and mate pair reads and can take many other inputs to include reference genomes and long read sequences for comparison. In lab 2, we are provided a E. coli genome to be assembled, a reference genome and PacBio long read sequences to help with error correction. After the genome is assembled, it is important to evaluate the quality of the assembly. We use the program Quast to evaluate the assembly of the E. coli genome. Evaluation of assembly quality can be examined in two sets of results, correctness score and size statistics. Correctness score looks at the accuracy of the contigs compared to the reference genome by matching the alignment of the contigs to the reference. For size statistics, Quast evaluates various assembler metrics such as the percentage of the reference genome accounted for by the contigs assembled, the number of contigs created, the length of the N50 contig, and number of unaligned contigs.

***Method***:

SPAdes is a great tool to use to assemble the E. coli genome. It is a short, bacterial genome and when compared to the reference will have very few sequencing errors. The scripts below are used to assemble the genome and then evaluate the quality of the assembly with the program Quast.

The following sequence files were provided:

ERR008613.fastq – paired end Illumina sequence reads from 200 base pair E. coli fragments

ERR022075.fastq – paired end Illumina sequence reads from 600 base pair E. coli fragments

PacBio\_2kb\_CCS\_500bp.fastq - PacBio CCS E. coli reads

NC\_007898.fasta – chloroplast reference genome

NC\_007898.gff – chloroplast reference genome annotation file

The first two assemblies were completed using only the 200 base pair E. coli fragments (ERR008613.fastq). One assembly also utilized the PacBio 500 base pair CSS E. coli reads to help assemble the genome. I used Quast to evaluate the accuracy of the assembled genome. After completing the Quast analysis on these assemblies I wanted to see if I could improve on the assembly metrics by including the Illumina 600 base pair reads with the 200 base pair reads for a more accurate assembly. I completed two more assemblies using the 200 base pair fragments as well as the 600 base pair E. coli fragments (ERR022075.fastq). The first assembly from the second attempt was completed using the script shown in figure 1.

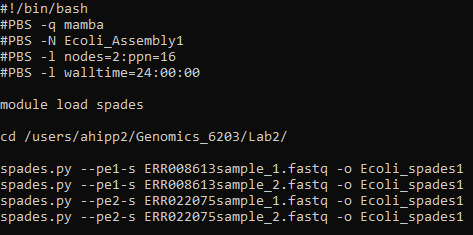


Figure : Assembly of E. coli with both sets of Paired End reads

Figure 2 shows the second assembly from the second attempt. This script uses both E. coli fragment files, as well as the PacBio CCS file to complete the assembly.

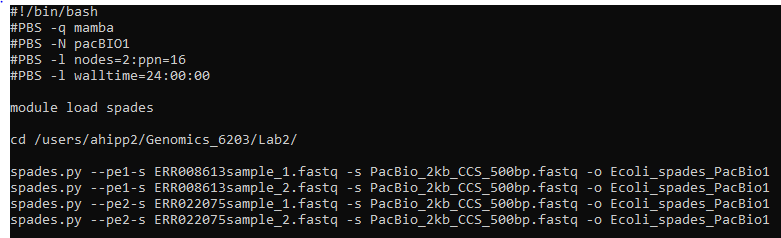


Figure : Assembly of E. coli with both sets of paired end reads and PacBio reads

Once the genomes were assembled, I once again evaluated the quality of assembly using Quast. Figure 3 shows the first analysis, where I used the reference genome provided for us (NC\_007898.fasta) and the annotation file (NC\_007898.gff) to compare the E. coli assembly against.

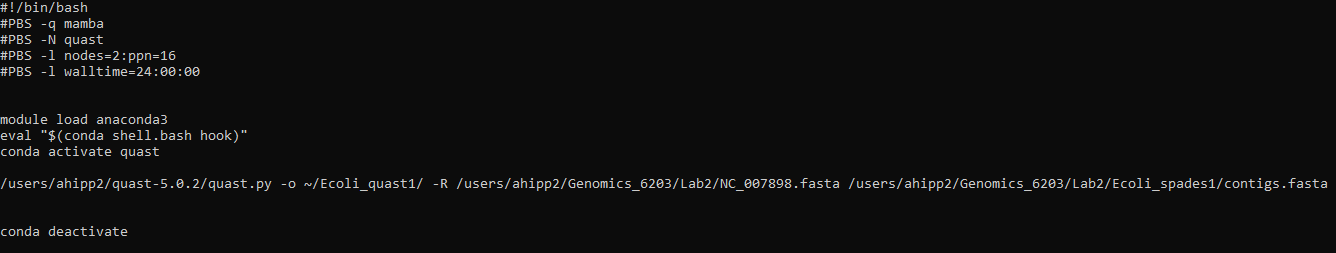


Figure : Quast evaluation comparing E. coli assembly against reference genome

The second assembly that was completed with the PacBio sequences was then analyzed in Quast using the reference genome (NC\_007898.fasta) and annotation file (NC\_007898.gff) for comparison.

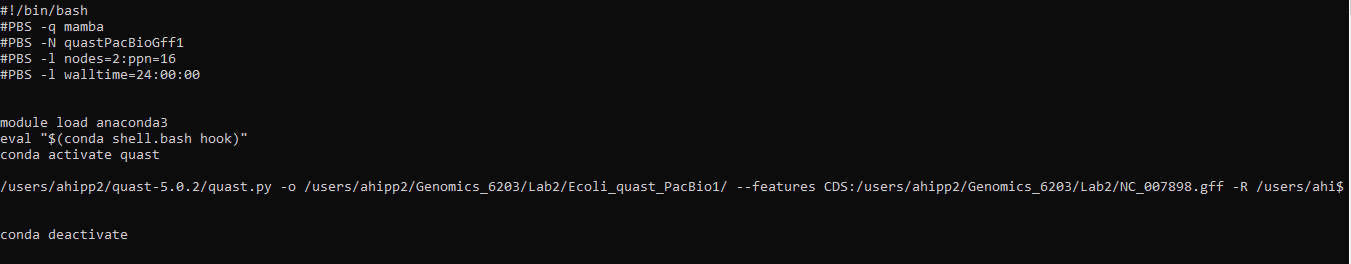


Figure : Quast evaluation of E. coli assembly with annotation file and PacBio sequence against reference genome.

***Results:***

From the first attempt at assembly and evaluation. Figure 5 shows all the graphs I will include from the first attempt at assembly. All remaining graphs and data are representative of the second attempt at assembly.



Figure : Initial Quast analysis report with E. coli short read sequence data. Both runs included reference .fasta and .gff file

Second attempt of assembly and evaluation:



Figure : Quast analysis generated with E. coli long and short read data. Both runs included reference .fasta and .gff file

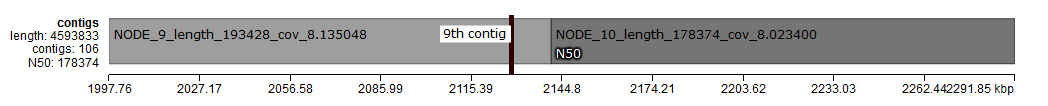


Figure : Representation of N50 contig from Icarus viewer.

Figure 7 shows the 10th contig which represents the N50 for our assembled genome after assembling with 200 and 600 base pair fragments and the PacBio reads.

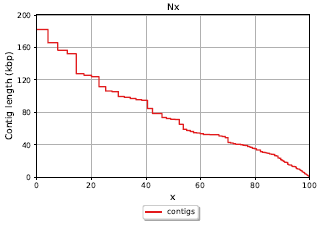
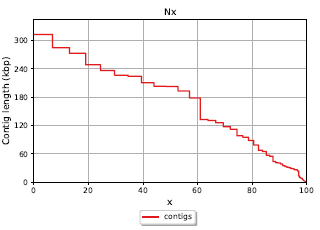
 

Figure : Nx graph – (Left) Assembly without PacBio reads. (Right) Assembly using PacBio reads.

The two graphs above show the Nx graph for the second attempt at assembling the E. coli genome. The graph on the left represents the contig composition for the assembly without using PacBio reads, while the graph on the right represents the E. coli genome assembled using the PacBio sequence reads.

***Discussion***:

My first attempt at assembling did not produce metrics that I expected, and I felt it necessary to assemble the E. coli genome again. The first assembly, which we can call phase 1, was completed by trying two different strategies that I could use to compare against each other. For the first assembly, I only used the Illumina paired end 200 base pair E. coli fragments in SPAdes and the chloroplast reference genome (NC\_007898.fasta). The second strategy I used the 200 base pair reads, the chloroplast reference genome, and the PacBio CCS reads as the input files into SPAdes. I hypothesized that the addition of the PacBio reads would help connect contigs that might have overlapping reads over two neighboring contigs. The goal was to compare the accuracy of assembly using the E. coli fragments with and without the input of the PacBio sequence reads. I ran these two assembled genomes in the Quast program to evaluate the quality of my assembly. My results from Figure 5 show that the PacBio reads significantly improved the assembly metrics. It shows the PacBio reads reduced the total number of contigs from 137 to 79 compared to the assembly without the PacBio reads. Also, the PacBio reads help increase the size of the N50 contig to 202,795 bases compared to only 71,567 without PacBio. Additionally, the number of unaligned contigs was reduced from 136 to 78. The assemble metrics generated by Quast confirm that the assembly with the addition of PacBio reads helped to provide a higher quality assembly.

I wanted to further improve the quality of assembly, so I completed phase 2, which consisted of retracing the assembly and evaluation steps from phase 1. The major difference in phase 2 was that I used the 600 base pair Illumina paired end E. coli fragment sequence reads as well as the 200 base pair reads. I also did two different assembly strategies in phase 2, running one assembly with PacBio reads and one assembly without PacBio reads. I hypothesized that my metrics would improve even more so than the second assembly in phase 1. With the 600 base pair fragments, I believed long repeating sequences could be better aligned with the longer fragments for comparison. The longer paired end reads could potentially recognize repeating sequences split over two contigs and join them together. The longer fragments could be used to join larger sequencing gaps by providing a better context for order and orientation. The first assembly, which was completed without PacBio reads, was then evaluated by Quast. The results did show an improvement from the first assembly in phase 1. I compared these to each other because they were assembled using the E. coli fragments and the chloroplast reference genome. Comparing these two assemblies did show improved number of contigs (137 to 122) and improvement on fully unaligned contigs (136 to 121). Furthermore, the largest contig from phase 2 was almost 20,000 base pairs longer than its corresponding assembly in phase 1. These numbers were improved, but I expected the phase 2 assembly with PacBio reads to be even more improved.

I was surprised at the results from Quast on the phase 2 assembly that included the 200 base pair and 600 base pair fragments, the PacBio reads, and the reference genome sequence. Although the largest contig was measured at 317,165 bases, the Quast report showed the number of contigs to be 106 and the number of fully unaligned contigs equal to 105. Additionally, the N50 contig was only 178,374 bases. These results all seemed to be inferior compared to the phase 1 assembly with PacBio reads. Comparing the phase 2 reads side by side did see expected improvements in assembly metrics. Figure 8 shows the Nx graph of the phase 2 assemblies, and the assembly with PacBio reads (right) is superior for virtually any x value.

Based on the results from this experiment, I could conclude that the PacBio reads provide a more significant improvement to the overall assembly quality than including the 600 base pair E. coli fragments. I am not convinced this finding is universally true, and I would want to run several more tests before making this claim. The 600 base pair fragments are important to resolving redundant paths and repeating regions that may not visualized by the 500 base pair PacBio reads.