Apr 27th, 1:45 PM - 3:00 PM

New Paenibacillus Bacteria Genome Assembly

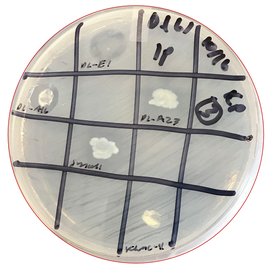
Makenna Kager
Ouachita Baptist University

New *Paenibacillus* DLA16 Bacteria Genome Assembly

Makenna Kager

Introduction

Bacterium DLA16 was isolated from an Austin, Texas, soil sample in 2017. **DLA16 was screened for antibiotic production in nutrient limitation swab patch assays**.

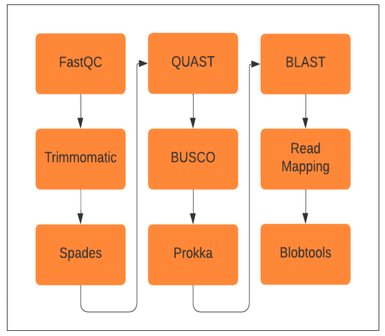


In these assays, DLA16 **produced antibiotics** inhibiting growth of the bacteria *Escherichia coli*, *Enterobacter aerogenes* NRRL B-407, and *Klebsiella pneumoniae* NRRL B-3521(RG2). Because it inhibited multiple pathogenic species, DLA16 was **selected for whole genome sequencing**.

Sequencing was performed at the Hubbard Center for Genome Studies at the University of New Hampshire on an Illumina HiSeq 2500 instrument with 250 bp paired end reads being sequenced.

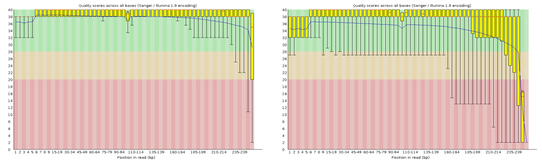
Goals

The goal of this project was to turn the DLA16 sequencing reads into an assembled genome using the Bash command line interface on the UNH server "Ron", and to do preliminary analysis on the assembled genome following the steps above.



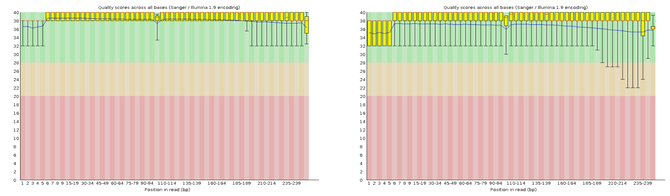
FastQC

This program shows the quality of the raw reads on the unassembled genome. This program was run before and after Trimmomatic to confirm that the trimming step worked. The images below show the quality of the reads before trimming. Note the read bars in the yellow and red sections.



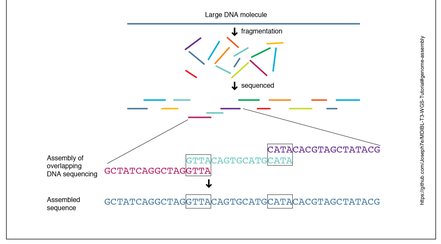
Trimmomatic

This program takes the raw reads and trims off low quality reads, resulting in a higher quality set of reads for assembly into a genome. The FastQC results after trimming are seen in the images above. Note that all of the read bars are in the green section.



SPAdes

This program takes the trimmed reads and overlays them in the correct overlapping positions to create larger contigs.

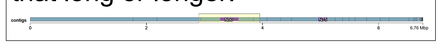


QUAST

This program gives several quality measurements for the contiguity of the assembled genome.

Largest contig	228 kilobase pairs
Total length	6.76 Megabase pairs
GC%	41.91%
N50	3.69 kilobase pairs

N50 represents the median contig length, with half of contigs being that long or longer.



BUSCO

The idea of this program is that the bacteria genome should contain one of each single-copy orthologs (BUSCOs) in its genome. A single-copy ortholog is a gene with one functionally identical copy in every organism. The percent of single-copy BUSCOs present will tell us about the coding completeness of the genome.

Complete BUSCOs	124	100%
Complete and single-copy	123	99.2%
Complete and duplicated	1	0.8%
Fragmented	0	
Missing	0	
Total BUSCO groups searched	124	

Prokka

Prokka runs multiple programs at once that can recognize different features of a genome such as CDs (coding sequences), rRNA, tRNA, tmRNA, rRNA, and much more. This program provides a whole genome annotation by identifying features of interest in the DNA sequences and labeling them.

Contigs	250
Bases	6.83 Mbp
CDs	6066
rRNA	10
repeat regions	14
tRNA	82
tmRNA	1

BLAST

This program compares the DLA16 genome to other bacterial genomes. This can tell us specifically what genus our sequence is the closest match to. The database used in this research is the nucleotide (nt) BLAST database. **The BLAST results indicate that DLA16 is in the genus *Paenibacillus***.

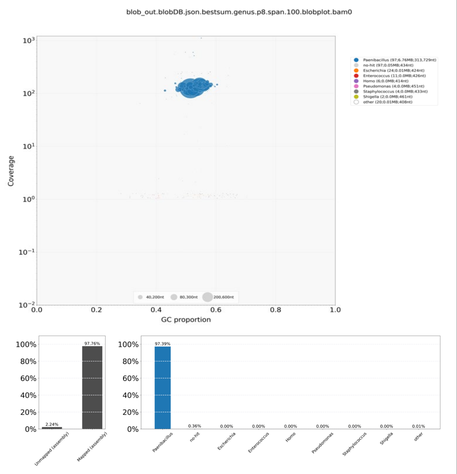
Read Mapping

This program maps the individual reads back to the assembled genome, showing where sequencing reads match to the assembly. The main output with this command is in a Sequence Alignment Map (SAM) format. This file provides the information about and information.

Blobtools

This tool allows us to visualize the genome assembly. The top plot shows the GC content on the x-axis; GC content is often genus specific. Blob size indicates the contig length and color shows the taxon with the corresponding GC content. Reads with different GC content would have blobs of a different color.

The bottom plot shows the percent of DLA16 reads that mapped to the assembled DLA16 genome during Read Mapping (black bar), and the percent of mapped reads that corresponded to various taxa.



Conclusion

Both BLAST and Blobtools identified DLA16 in the genus *Paenibacillus*, allowing us to confirm that our bacterium is this genus. *Paenibacillus* is a soil bacterium that aids plants in growth through things such as nitrogen fixation. Bacteria in this genus can also be used for protection against other bacteria so it in commonly used in medicine, agreeing with our observation of antibiotic production.

Comparing the DLA16 genome analysis results to previously published *Paenibacillus* research made it simple to verify which of our results were good. For example, the DLA16 genome is 41.91% GC and research into *Paenibacillus* shows that this genus has a GC content between 39 and 59 percent. This means these results have a high probability of being correct since they follow the trend within this genus.

Acknowledgments

Thank you to Dr. Ruth Plymale, Ouachita Baptist University, Department of Biology, Patterson School of Natural Sciences, and the Hubbard Center for Genome Studies at the University of New Hampshire.