Genomic assembly proposal

Molli Coombs

Background

Whole genome sequencing gives the ability to classify bacteria in a clear-cut and effective way using average nucleotide identity (ANI) values (Ha et al. 2019). The outcome of this sequencing has helped differentiate species on a more certain level and is often used in medical settings with isolated unknown bacteria. This sequencing, specifically for bacteria, is widely achieved using the Illumina MiSeq system with identification based on genome data analysis (Chun et al. 2018).

In this experiment the genome of an unknown species of gram-negative bacillus will be sequenced to allow for classification. This is the first set of steps to classifying this bacterial species and beginning treatment in patients with this bacterium. The bacteria used in this study was isolated from an immunocompromised patient with a blood infection who was undergoing chemotherapy as treatment for pre-diagnosed acute myeloid leukemia. The bacteria found in the patient showed signs of antibiotic resistance and lead to consistent high fever, abdominal pain, and eventual sepsis.

The sequencing of the entire genome of the bacteria will allow for the pathogen to be identified on a species level and compared to known species to determine identity or significant genetic difference using ANI scores. The ability to quickly sequence isolated sample genomes in this fashion is important in medical settings to allow for quicker diagnosis and treatment. The next step in this study would be to compare the genome being sequenced to known genomes and comparing the ANI scores to determine speciation of the unknown bacteria. If this is a novel species of bacteria, meaning ANI scores no greater than 95% appear between current species, then to delineate species an RNA analysis and comparison of 16S rRNA would be used to determine speciation (Chun et al. 2018).

Methods

Raw genomic data from the Illumina MiSeq run will be downloaded as the FastQ file available and run through a pipeline utilizing Velvet assembly, which works will with Illumina data. Velvet manipulates de Bruijn graphs as an alignment technique through compression to create contigs (Miller et al. 2011). A simple pipeline can be written to include the downloaded genomic data and run through Velveth and Velvetg once the read lengths are categorized and the hash length is chosen (Zerbino 2010). VelvetOptimiser will be utilized to estimate the proper parameters to set to assemble the genome correctly. This package bundles with the total Velvet package and helps determine the ideal parameters to use for the best possible assembly. Velvet software works in two steps, with Velveth reading the sequence files and building a dictionary, and Velvetg reading the alignments and building the de Bruijn graphs and removing errors (Lantz et al. 2018). Velvet include quality trimming in the program which increases the quality of the reads being used and will be utilized for this project to help ensure high-quality assembly.

The code will be written in BBedit and will be run in terminal connected to Spruce to ensure enough storage space is available. Velvet is a simple package to install and includes Velveth, Velvetg, and Velvetoptimizer. I have some experience with coding already and am relatively familiar with running shell commands. My previous experience with bioinformatics will help with running lines of code and also with accessing and working with genomic data for this project. I plan on continuing to learn about genome sequencing and assembly to be used later on in my academic career, so this study will help introduce me to the coding that is required to achieve this type of analysis.

Creating a pipeline using Velvet is a reasonable approach to this genome assembly because the coding needed for the pipeline is a relatively simple process. The data from the isolated bacteria is publicly available through NCBI at (<https://www.ncbi.nlm.nih.gov/assembly/GCA_008632125.1>) and is easy to access and download in a FastQ file.

Expected Outcomes

From this study it is expected to assemble the genome of an unidentified bacteria isolated from a blood culture of an immunocompromised patient. This will be the starting point for further analysis of the bacteria to determine the species using comparison to already sequenced genomes of bacteria. Once the genome is compared to similar genomes to determine genus the next step would be identification at a species level. That step can be done using RNA analysis of highly variable 16S rRNA. The high variability of this RNA allows for a more accurate determination of species, which can be hard to achieve in bacteria that can have similar genetic makeup. This analysis will allow for better diagnosis and treatment of patients facing infections caused by this bacterium by giving a species level identification.

Possible challenges that could be faced in this study are problems in the coding pipeline, which will require re-writing code and troubleshooting. If the genome is large then space on the computer being used could be a problem, which is why spruce is being utilized for extra storage. Other possible challenges to be faced could be poor data quality from the isolated bacteria DNA that was extracted. If the DNA quality is low, it will be difficult to assemble the genome accurately and could result in large gaps in the genome. If this occurred, it would make errors in identification of the genome more likely and increase the chances of incorrect identification of genus.

References

Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, et al. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol. 68(1):461–466. doi:10.1099/ijsem.0.002516.

Ha SM, Kim CK, Roh J, Byun JH, Yang SJ, Choi S Bin, Chun J, Yong D. 2019. Application of the whole genome-based bacterial identification system, TRUEBAC ID, using clinical isolates that were not identified with three matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) systems. Ann Lab Med. 39(6):530–536. doi:10.3343/alm.2019.39.6.602.

Lantz H, Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, Amselem J, Bouri L, Bocs S, et al. 2018. Ten steps to get started in Genome Assembly and Annotation. F1000Research. 7. doi:10.12688/f1000research.13598.1.

Miller JR, Koren S, Sutton G. 2011. Genome Sequencing Algorithms. 95(6):315–327. doi:10.1016/j.ygeno.2010.03.001.Assembly.

Zerbino DR. 2010. Using Velvet *de novo* assembler for short-reading sequence technologies. Curr Protoc Bioinforma. 31(11):1–13. doi:10.1002/0471250953.bi1105s31.Using.