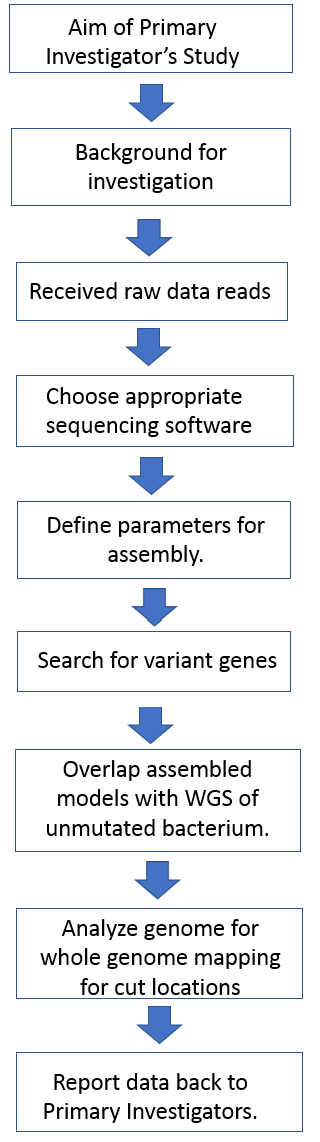
Heather Price

Genomic Data Analysis

Genome assembly

9.24.20

Genome Assembly of Pathogenic *E. coli* Species

**Introduction:**

*Escherichia coli* is a native gut bacterium that, at its base level, is a harmless coliform that is part of the normal flora of the human and animal intestinal tract. Even so, there are a multitude of strains of *E. coli* which have mutated to take on traits that can be harmful to the host that it happens to infect. The strain that is being examined here is a mutated form of *E. Coli* which has proven to be muti-drug resistant, producing a form of beta-lactamase (1,2) which inhibits the cell wall inhibition of penicillin and like drugs. These mutants can also bring about a variety of infections to the host which can range from simple diarrhea to sepsis (1).

One of the primary reasons it is so important to be able to type the new and emerging pathogenic species of bacteria, and *E. coli* in particular, is that in emerging areas of society where hygienic norms substandard and by any means less than desirable, these mutated pathogens can wreak havoc on an already struggling populace. These pathogenic bacteria can also migrate to different areas due to vector transport (2), which is the primary concern of the primary investigator who has requested the sequencing of the raw data collected. Due to vector transport such as common house flies, the pathogenesis of new strains of bacteria such as *E. coli* make for a larger issue than just those faced by a developing population.

**Methods**:

The data collected by the primary investigator and his/her/their group will be procced via protocols established by the investigator in accordance with all ethics committee standards and protocols set. It is expected that data will be properly collected and purified for a well purified high concentration of bacterium to be analyzed for whole genome sequencing (WGS) by Illumina.

Figure 1: Analysis Flowchart

Raw data will be handed over from the primary investigator’s team for bioanalysis. Coverage of data will be expected to be deep read coverage calculated by Illumina’s standard: C = (L (N/G). Where C is the calculated coverage, L is the read length, N is the number of reads and G is the genome length (3).

Raw data is expected to be delivered in the FastQ file format which is standard for Illumina (4). As such, Illumina originated software A5-miseq (5) would be preferable to use with Illumina generated raw data. However, if this is not an option due to funding, open source software such as BaseSpace Sequence Hub (6), which is also developed by Illumina with some limited sequencing capabilities as compared to A5-miseq. After software has been determined, appropriate parameters will be set for short read base pair assembly referencing the E. coli genome which can be found at: [https://www.ncbi.nlm.nih.gov/assembly/GCF\_005221985.1 strain 131](https://www.ncbi.nlm.nih.gov/assembly/GCF_005221985.1%20strain%20131) . Data quality would be checked with the FastQC tool to assessed to determine aspects such as quality of reads, read length, number of N’s which need to be trimmed. Once this is determined, data will be trimmed appropriately and assembled. Gap filler tools may be required, which can be procured for use through open source programs. Protocols followed for these steps can be found through Melbourne Bioinformatics (4).

Preferably several models of the genome will be constructed in the assembly process, at least three, to compare to the NCBI assembled genome for *E. Coli* to assess homology. This comparative process will give indicators as to variant genes within the assembly which could indicate or verify a mutation which could be pathogenic in nature (7). It is hoped that a high number of reads will be available to overlap a repeated suspected variant to give evidence for the correct mutation. Scaffold assembly will need to have parameters which are stringently set due to short reads in order to give rise to evidence that the suspected mutation is indeed a mutation and not a misread. The number of reads will be reported to support or refute the original hypothesis of the vector carried suspected pathogen strain of *E. coli*. Whole genome mapping to search for sites which may prove to have the ability to be cut by restriction enzymes would be preferable for strain typing due to the drug resistance that is posed by newly mutated bacterial pathogens (8).

While the team proposing this assembly has not directly been involved in whole genome sequencing procedures, RNAseq protocols have been observed in laboratory settings as well as RNA isolation procedures done by the team in question. Bioinformatics experience presented by this team is primarily in datamining with a basic background of programming knowledge attained in an undergraduate setting. It is believed that if given the chance, the team will be able to apply knowledge gained from previous experiences to genome assembly for this project.

**Expected Outcomes**:

The expected outcomes of this process would be, primarily, a fully assembled genome of the mutated *E. coli* pathogen species. Secondary expected outcomes to proper data processing and analysis would be addition to a growing library of sequenced genomes for known pathogens with quicker ability to treat underlying causes before they become a severe threat to an already struggling populace, as well as cumulated experience for the team in question in regards to whole genome sequencing processing and analysis from beginning with the raw data to ending with a completed genome with noted areas for disruption in the genomic data.

Challenges which could be faced during the process could include corruption to the original samples which damages the data pipeline from beginning of construction to end. In this case, subsequent data would need to be collected and processed through Illumina for additional raw data to provide to the bioinformatics team. Other challenges posed could be limitations of determined software for processing. Most open source software is more than serviceable when properly utilized, however, and would depend mostly on trouble shooting from the bioinformatics team as well as consultations with expert sources.

**References:**

1. Alizade H. (2018). Escherichia coli in Iran: An Overview of Antibiotic Resistance: A Review Article. Iranian journal of public health, 47(1), 1–12.

2. Heiden, S.E., Kurz, M.S., Bohnert, J., Bayingana, C., Ndoli, J.M., Sendegeya, A., Gahutu, J.B., Eger, E., Mockenhaupt, F.P., Schaufler, K. (2020) Flies from a tertiary hospital in Rwanda carry multidrug-resistant Gram-negative pathogens including extended-spectrum beta-lactamase-producing *E. coli* sequence type 131. *Antimicrob Resist Infect Control* **9,**34 (2020). https://doi.org/10.1186/s13756-020-0696-y

3. Illumina (2020) Read Length Recommendation: How to calculate the right read length for your sequencing run. [Accessed September 24, 2020] <https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html>

4. Melbourne Bioinformatics. De novo Genome Assembly for Illumina Data.[Accessed September 24, 2020] <https://www.melbournebioinformatics.org.au/tutorials/tutorials/assembly/assembly-protocol/>

5. David Coil, Guillaume Jospin, Aaron E. Darling, A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data, Bioinformatics, Volume 31, Issue 4, 15 February 2015, Pages 587–589, <https://doi.org/10.1093/bioinformatics/btu661>

6. Illumina. BaseSpace Sequence Hub. [Accessed September 24, 2020] <https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html>

7. McElroy, K., Thomas, T. & Luciani, F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb Informatics Exp* **4,**1 (2014). https://doi.org/10.1186/2042-5783-4-1

8. Miller, J.M. (2013). Whole-Genome Mapping: a New Paradigm in Strain-Typing Technology. Journal of Clinical Microbiology. Vol 51, p. 1066 –1070.