**Analyzing Intragenomic and Intergenomic Variation of 16S rRNA of E. coli Bacterial Strains using Long Read Technology** (Working title)

**Abstract**

With the increasing availability of long-read (LR) data, it is now possible to analyze a genome with increased resolution with long-read sequencing than the commonly used short-read (SR) sequencing. LR tech benefits from its increased length to cover highly conserved genomes like the 16S region of a bacterial genome, allowing for allele-to-allele comparison and identifying intragenomic variation. While SR sequencing provides a great representation of genomes on a genus and species level, it lacks the ability to differentiate distinct alleles from another due to insufficient overlap. Overall, LR sequencing seems to show better resolution than SR sequencing when analyzing species diversity, frequency, etc. (Jeong et al, 2021). To test this, we produced a python wrapper script that analyzes the inter and intragenomic differences in the 16S region of closely related bacterial strains of *Escherichia coli*. By deriving the edit distance matrix of *E. coli* genomes and plotting their distribution, we (were/were not) able to determine the validity of increased resolution of LR sequencing.

**Introduction**

Short-read sequencing of the 16S region still remains as an impressive method due to its availability, however its resolution is only ideal for differentiating on the genus and species level. This is due the fact that shorter reads do not span the entire length of the 16S region, making it difficult to differentiate alleles. Long-read sequencing can remedy this and has the ability to look at intragenomic variation, compare alleles from one genome to one another and overall provide a better resolution than short-read sequencing. This further allows us to find similarities in genomic DNA between homologous sequences. Briefly, our script downloads all *E. coli* genomes from NCBI’s assembly page, filters for long read sequencing technologies, and extracts all 16S rRNA copies from each of these genomes. Next, an edit distance matrix is created comparing all of the genome copies to one another. Lastly, the between and within genome edit distances are parsed out from the matrix for further bio-statistical analysis.

A link to further understand the code is here: <https://github.com/amyrold/LongReads.git>

**Implementation**

1. DATA DOWNLOAD

Development begins with downloading the metadata for all genomes for *E. coli* from NCBI’s assembly page. Genomes in the meta data are parsed out by a python script and outputs genomes that were sequenced by Nanopore/PacBio, these outputs a list of accession ids. With Unix commands, these whole genomes are downloaded into a large multi-fasta file called wgs.fasta.

1. BLAST

Genomes in the multi-fasta file are then BLASTed against a 16S rRNA database. First, with Unix command line tools, the 16S refseq is downloaded and the local BLAST database is created. Using BLASTn, the multi-fasta file is blasted against the 16S rRNA database, including the parameters accession, percent identity, query start, query end, length, and e-value, and outputted to the myresults.csv file. To distinguish the 16S copies in the same genome from each other, the accessions in this file are renamed with accession\_copy# using the “split\_acc” and “rename\_acc” functions, and outputted to the trim.csv file. Lastly, using the trim.csv file as input to the “store16S” and “trim\_fa” functions, each genome copy is extracted from its corresponding genome in the wgs.fasta file and written to the trim.fasta file, later to be used in subsequent edit distance matrix functions.

1. EDIT DISTANCE

First, an NxN edit distance matrix will be initialized, where N = the number of records in the trim.fasta file, and the column and row names will be the record names. After generating pairwise combinations of headers, the matrix will be sequentially populated with corresponding edit distances using the “ed” function and outputted as matrix.csv. Next, in order to determine statistics of edit distances based on 16S rRNA copies *within* one genome and *between* different genomes, the “within\_between”, “get\_ed\_within”, and “get\_ed\_between” functions will be used to generate separate lists of within and between edit distances (between.csv and within.csv).

1. VISUALIZING VARIATION

Using ggplot2 within RStudio, the between.csv and within.csv files will be used to create boxplots of between and within variation of 16S rRNA copies.

**Results and Discussion**

Our wrapper script shows success in downloading long reads and BLASTing *E. coli* against the 16S region… While LR sequencing is not as inexpensive as SR sequencing, the possibility of increased resolution and allele-to-allele differentiation is valuable information in microbial diversity. This wrapper script allowed us to test the validity of LR sequencing resolution and we found that… In future developments, we wish to research different means of visualizing edit distance plot distributions to better visualize intra/inter genomic variation…

Figure 1 - Edit Distance Matrix

Application

Description automatically generated with medium confidence

Figure 2 - Between Edit Distance Distribution - Box Plot

Chart, histogram

Description automatically generated

 Figure 3- Within Edit Distance Distribution - Box Plot

Graphical user interface, chart, histogram

Description automatically generated

**References**

1.     Jeong J, Yun K, Mun S, Chung WH, Choi SY, Nam YD, Lim MY, Hong CP, Park C, Ahn YJ, Han K. The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. Sci Rep. 2021 Jan 18;11(1):1727. doi: 10.1038/s41598-020-80826-9. Erratum in: Sci Rep. 2021 May 19;11(1):10861. PMID: 33462291; PMCID: PMC7814050.