Alan Chen, A13989230
Myra Haider, A14480471
Jiayi Wu, A15124058
Prof. Shannon Ellis
DSC180B Project Report Check-in

One of the most difficult problems in drug development today is the growing number of bacterial strains that have developed resistance to antibiotics. Due to their immunity to known medicine, their exposure to humans can lead to infections that are impossible to cure [1]. According to the International Federation of Pharmaceutical Manufacturers & Associations (IFPMA), by 2050, antimicrobial resistance could kill up to ten million people per year [2]. The biological mechanism behind antibiotic resistance involves changes in bacteria at the genetic level, through either random mutations in their own DNA or the acquisition of genetic material from the environment [3]. Although antibiotic resistant bacteria have been studied for decades, their whole genome sequencing has started relatively recently [4]. This study aims to investigate the genetic factors associated with antibiotic resistance by performing a genome-wide association study on antibiotic resistant E. Coli bacteria and comparing the results against non antibiotic resistant E. Coli bacteria.

Enterobacteriaceae, specifically E. Coli, commonly causes infections both in healthcare settings and communities [5]. Certain strains however have developed an especially dangerous resistance mechanism, the ability to produce an enzyme known as extended-spectrum beta-lactamase, or ESBL [5]. ESBL is capable of breaking down multiple types of antibiotics such as penicillin, rendering them ineffective [5].

The study is conducted using two groups of samples, one with an ESBL-enzyme producing strain of E.Coli which is antibiotic resistant, and a non-antibiotic resistant control group [6][7]. Sample sizes were 36 whole genomes for each group. Whole genome sequencing data was obtained via an Illumina MiSeq sequencer and formatted as two pair-end fastq files per sample [6][7]. The FastQ files were first piped through FastQC, a quality control program, with default settings for an initial summary of read quality [8]. The generated report outputs Pass/Warn/Fail flags for various categories (e.g. Basic Statistics, Per Sequence GC Content, Per Base Sequence, etc.) [9]. We chose to keep samples that passed the Basic Statistics check and compared overall quality between both groups. All of our files passed this check, although there were some differences amongst groups that may affect variant calling accuracy further down the

line. The most significant difference between the groups was the %GC content, with most of the ESBL producing samples having higher amounts on average compared to the control group. This may either be a physiological phenomenon or a confounding factor that may need to be addressed.

During the sequencing process, temporary adapter sequences of nucleotides are attached to the fragments of DNA in order to facilitate sequencing [8]. Sometimes these adapters are accidentally sequenced as part of the sample genome, so they need to be algorithmically removed using tools such as Cutadapt [10]. Cutadapt was run on all 72 samples using default parameters, in trimmed pair-end reads mode to "remove adapter sequences… from high-throughput sequencing reads" [10]. The trimmed fastq files were piped through FastQC again to ensure that the data still passes our quality check post-adapter trimming, which every file did [8].

Once all files have been cleaned, the reads from each sample can be aligned into a whole genome sequence using a program called Bowtie2 [11]. Bowtie2 takes in a pair of fastq files from a single sample and a known reference genome to assemble the reads into a contiguous sequence [11]. After performing alignment on each sample, 36 .sam files are produced for each group. These .sam files are converted into a readable .bam format using samtools, a package for manipulating .sam files. Once this is completed, the 72 .bam files are ready for variant calling and analysis using GATK [12].

GATK is a suite of tools used for the analysis of genomic data [12]. For this study, the HaplotypeCaller tool was used to identify Single Nucleotide Polymorphisms (SNP's) between each group and compare them [13]. The resulting output after GATK is an aggregated VCF file containing the list of variants for all 72 E.Coli genomes. The VCF file is then piped into Snpeff, a tool that "annotates and predicts the effects of genetic variants on genes and proteins" [14]. This enables us to observe the specific markers that may tell us why ESBL is able to nullify current antibiotics.

Citations

1. Center for Disease Control. "About Antibiotic Resistance." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 13 Mar. 2020, www.cdc.gov/drugresistance/about.html#:~:text=Antibiotic%20resistance%20happens%20when%20germs,and%20sometimes%20impossible%2C%20to%20treat.
2. Cueni, Thomas B. "By 2050, Superbugs May Cost the Economy $100 Trillion." *IFPMA*, International Federation of Pharmaceutical Manufacturers & Associations , 13 Nov. 2018, www.ifpma.org/global-health-matters/by-2050-superbugs-may-cost-the-economy-100-trillion/#:~:text=Antimicrobial%20resistance%20(AMR)%20is%20on,%E2%80%9Csuperbugs%E2%80%9D.
3. Reygaert, Wanda C. "An overview of the antimicrobial resistance mechanisms of bacteria." *AIMS microbiology* vol. 4,3 482-501. 26 Jun. 2018, doi:10.3934/microbiol.2018.3.482
4. Ikegawa, Shiro. "A short history of the genome-wide association study: where we were and where we are going." *Genomics & informatics* vol. 10,4 (2012): 220-5. doi:10.5808/GI.2012.10.4.220
5. Center for Disease Control. "ESBL-Producing Enterobacteriaceae." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 22 Nov. 2019, www.cdc.gov/hai/organisms/ESBL.html.
6. Patel, IR. *National Center for Biotechnology Information*, U.S. National Library of Medicine, 22 May 2015, trace.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA230969&o=acc_s%3Aa.
7. Hokkaido University. *National Center for Biotechnology Information*, U.S. National Library of Medicine, 8 Dec. 2020, trace.ncbi.nlm.nih.gov/Traces/study/?acc=PRJDB10450&o=acc_s%3Aa.
8. Babraham Institute. *Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, 26 Apr. 2010, www.bioinformatics.babraham.ac.uk/projects/fastqc/.
9. Index of /projects/fastqc/Help/3 Analysis Modules. (3333). Unknown. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Mod

ules/

10. *Cutadapt — Cutadapt 3.1 documentation*.https://cutadapt.readthedocs.io/en/stable/
11. John Hopkins University. “Bowtie 2.” *Bowtie 2: Manual*, 5 Oct. 2020, bowtie-bio.sourceforge.net/bowtie2/manual.shtml.
12. Broad Institute. “GATK”. *GATK - How to Map and clean up short read sequence data efficiently*. Unknown. https://gatk.broadinstitute.org/hc/en-us/articles/360039568932--How-to-Map-and-clean-up-short-read-sequence-data-efficiently
13. “HaplotypeCaller.” *HaplotypeCaller - GATK*, Broad Institute, 7 June 2020, https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller.
14. Cingolani, Pablo. “SnpEff&SnpSift.” *Home - SnpEff & SnpSift Documentation*, Github, pcingola.github.io/SnpEff/.

Appendix

One of the most difficult problems in drug development today is the growing number of bacterial strains that have developed resistance to antibiotics. Because these bacteria cannot be killed using known medications, their exposure to humans can lead to infections that are virtually impossible to cure [1]. The biological mechanism behind antibiotic resistance involves changes in bacteria at the genetic level, through either random mutations in their own DNA or the acquisition of genetic material from the environment [2]. Although antibiotic resistant bacteria have been studied for decades, their whole genome sequencing has started relatively recently. This study aims to investigate the genetic factors associated with antibiotic resistance, and use these findings to develop a machine learning model to predict whether a new bacterial strain has the potential to be antibiotic resistant.

The data used in this study would include whole genome sequencing data from multiple strains of bacteria, particularly the strains classified as threats by the CDC as of 2019 [3]. These would be compared to the strains of their non-resistant counterparts, as well as each other for commonalities that might characterize antibiotic resistance. We plan on using FastQC to check the quality of the dataset and Cutadapt for adapter trimming. Afterwards, we will pipe the processed data into Bowtie2 for read alignment and finally GATK and Snpeff for gene annotation and analysis. With this data we will engineer features and test various machine learning models. The project output would include a report containing the results of the investigation and the predictive model used to classify a strain as antibiotic resistant.

The replication paper is similar to this study as high throughput sequencing data is being processed and analyzed, however there are multiple major differences. Bacteria is the organism being studied, and DNA is being studied rather than RNA. This is because we are more concerned with genetic variation through SNP's in this case, rather than gene expression levels. Alongside this, the results of the investigation will guide the production of a machine learning model, something that was not done in the replication paper.

We will be studying a family of bacteria known as Enterobacteriaceae, specifically E. Coli. This bacteria commonly causes infections both in healthcare settings and communities. Certain strains however have developed an especially dangerous resistance mechanism, the

ability to produce an enzyme known as extended-spectrum beta-lactamase, or ESBL. ESBL is capable of breaking down multiple types of antibiotics such as penicillin, rendering them ineffective. Our goal is to study approximately 90 E. Coli samples and identify the genes that are responsible for the production of this enzyme. The data was collected using whole genome sequencing. There have been previous studies that have identified genetic mutations in other species of antibiotic resistant bacteria, however we would like to utilize these results to produce a machine learning model that can be used for prediction of future strains.

Sources

1) https://www.cdc.gov/drugresistance/about.html#:~:text=Antibiotic%20resistance%20happens%20when%20germs,and%20sometimes%20impossible%2C%20to%20treat
2) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6604941/
3) https://www.cdc.gov/drugresistance/biggest-threats.html
4) https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-018-3581-5
5) https://ann-clinmicrob.biomedcentral.com/articles/10.1186/s12941-015-0098-9