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

The Genetic Basis of Antibiotic Resistance in E. Coli

Introduction

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].

Data

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 as shown in Figure 2, 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.

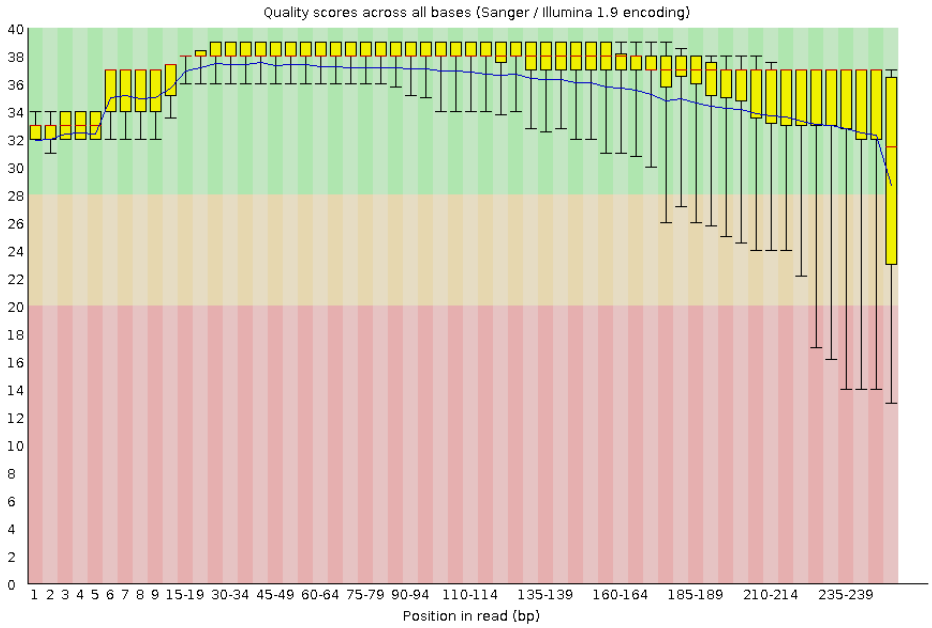

Figure 1: The distribution of quality scores for a single sample sorted by read length

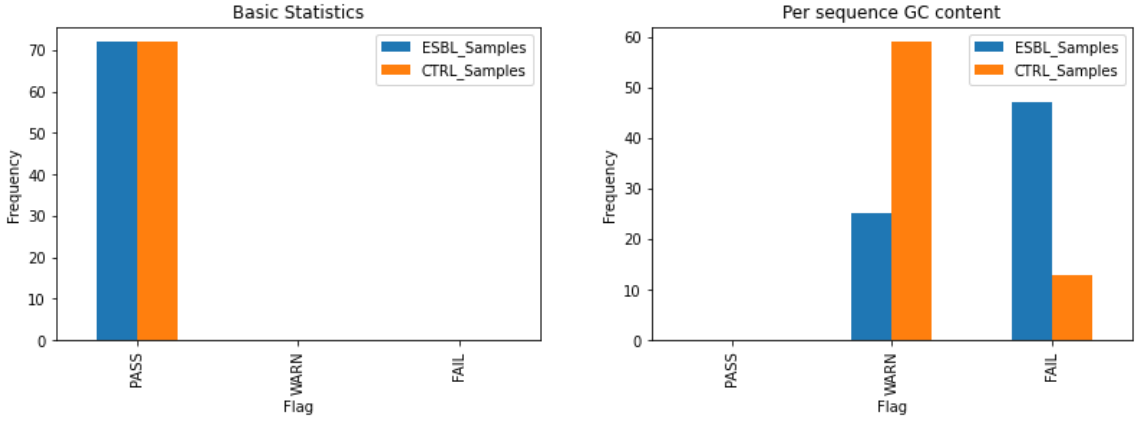

Figure 2: Aggregated count of quality Pass/Fail flags between ESBL and control samples.

Methodology

Figure 3 below provides an overview of our project pipeline, starting with downloading the raw reads and ending with our analysis from the genome wide association study and variant annotation using SnpEff. Once the sequencing data is downloaded and assessed for quality, it is processed using multiple bioinformatics software tools, starting with Cutadapt for adapter trimming.

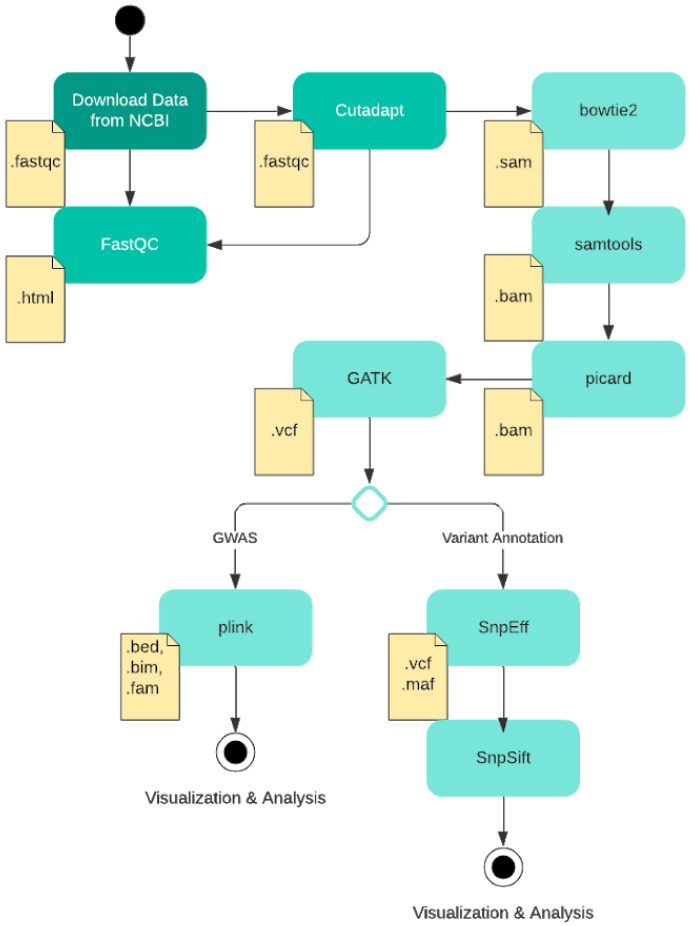

Figure 3: Project Pipeline Overview

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 which we then used for a genome wide association study using the software Plink, as well as variant annotation using SnpEff to identify variants of significance and their functional effects.

Before proceeding with the genome wide association study (GWAS), a basic principal component analysis was conducted using Plink to determine if there are potentially meaningful differences between the two groups. The first two principal components were calculated for each sample and plotted in the scatterplot shown in Figure 4. It can be seen that the antibiotic-resistant samples clustered together on the right of the graph, indicated by the points in blue. The orange points represent samples from the control group, which appear to have much more variance. It is unknown why the three outliers in the top left vary so significantly from the rest of the control group, but they were retained in the study due to the sample populations being so small.

The genome wide association study was performed using a logistic regression model to identify the SNP's that are statistically significant between the two groups. The model computes a p-value for each SNP indicating this level of significance. The model was tuned using the first two principal components as covariates and by filtering out all SNPs with a genotyping rate of

less than 60%. The genotyping rate is used as a measure of missing data, with 60% being chosen as it was the best threshold to minimize the skew of the distribution of p-values.

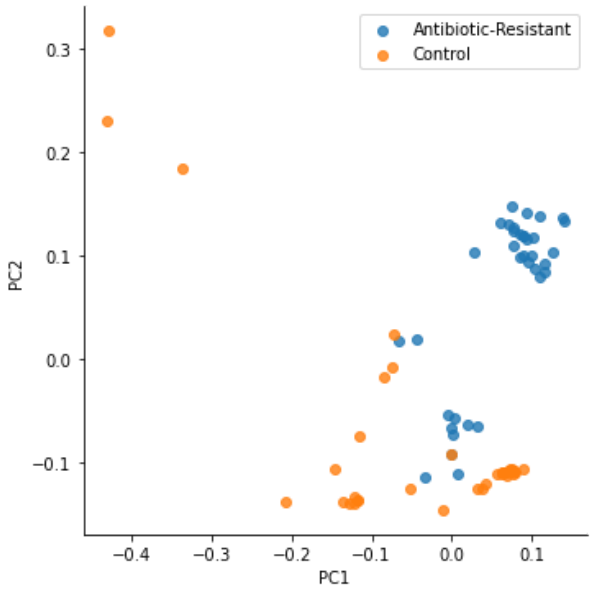

Figure 4: Principal Component Analysis Results

The variant annotation and filtration was performed using snpEff and snpSift. We used the built-in E-coli database that snpEff provided as the reference genome during annotation. This step gave us an annotated vcf file, a succinct summary report, and a gene_name table. We could then run snpSift on the VCF file to filter out the common variants which are present in both groups and leave only the variants that differ between groups. Finally, We used the extractFields command in snpSift to extract useful information from these remaining variants for further analysis.

Results

The snpEff output consists of a comprehensive summary report, an annotated vcf file and a gene_name table. The summary report provides us a brief overview of our variants. We have encountered 341666 warnings in the control group and 514589 warnings in the esbl group.Since the number of errors we encountered is relatively small compared to our datasize, we decided to focus more on the numerous warnings. These warnings mainly consist of ref_dose_not_match_genome error. We theorized that it may be a result of using two different reference genomes: the original reference genome and the one provided by snpEff. It was

unavoidable in this study as our initial reference genome was missing an essential .gff file. Our results could significantly improve if we built our own database. Another observation is that the variant rate of the esbl group is 1 variant every 16 bases, which is much higher than the control group with 1 variant every 24 bases. We hypothesize that it is due to the possibility that the ESBL group is more prone to mutations.

Besides the runtime information, the statistical data generated by snpEff illustrates that the control and ESBL group are incredibly similar with a few variance. For both groups, the Missense/Silent ratio is about 3 which is quite high. As this ratio increases, the more mutations happening in our samples directly affect the amino acid it produces. This could be the reason for E. Coli bacteria gaining resistance to antibiotics. The rest of the statistics are similar and do not reveal how E. Coli bacterias become antibiotic resistant. Next, we applied snpSift filtration to reduce the scope of the dataset to observe any noticeable changes.

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	84,527	72.196%
NONSENSE	4,684	4.001%
SILENT	27,869	23.803%

Missense / Silent ratio: 3.033

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	128,049	71.635%
NONSENSE	7,528	4.211%
SILENT	43,174	24.153%

Missense / Silent ratio: 2.9659

Figure 5: Missense/Silent ratio

In this step, we passed the annotated VCF file we got from snpff to snpSift. We then filtered them out and kept only high/moderate impact variants present in all of our ESBL samples and absent in at least one of our control samples. Originally we had over 60,000 variants after annotation, but with filtering we ended up with 50. We extracted their information and counts for the ten most frequently presented Mutated Genes, each one denoted by a specific gene ID. Their gene IDs are listed below.

[HUS2011_1899 , HUS2011_1904, HUS2011_1897, HUS2011_1898, HUS2011_1900, HUS2011_1901, HUS2011_1902, HUS2011_1903, HUS2011_1896, HUS2011_1905]

We can observe that in the remaining variants the most frequent mutations are from A -> G (see Base changes). In addition, the variants_impact_MODIFIER is quite high indicating that there

are more variances, with greater impact, in this gene. There are also many variants identified to have a missense variant effect, which usually resulting in a different amino acid sequence.

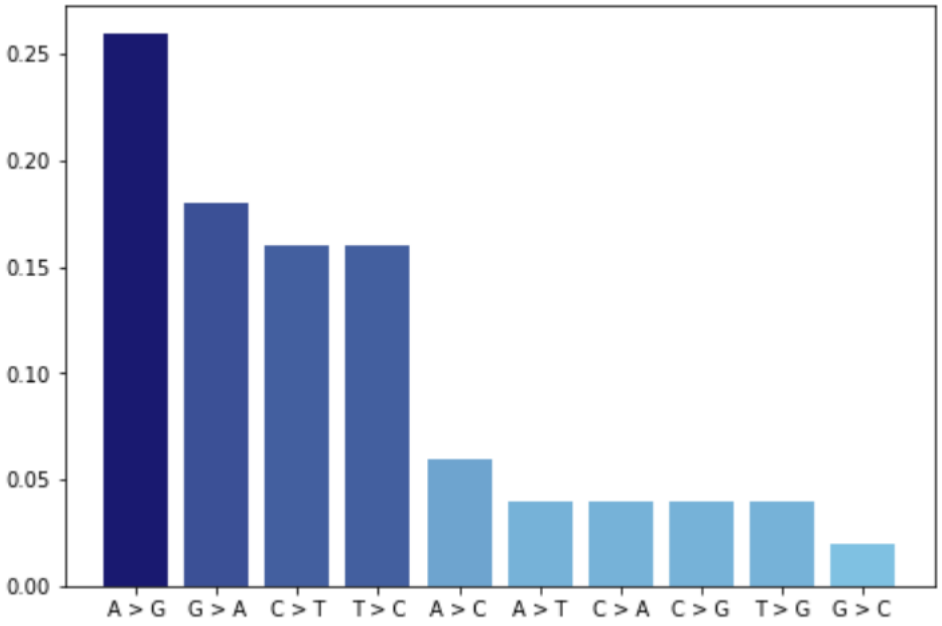

Figure 6: The count of base change in remaining variants

#GeneName	HUS2011_1896
GeneId	HUS2011_1896
TranscriptId	CCQ28774
BioType	protein_coding
variants_impact_HIGH	3
variants_impact_LOW	15
variants_impact_MODERATE	41
variants_impact_MODIFIER	814
variants_effect_downstream_gene_variant	422
variants_effect_missense_variant	41
variants_effect_stop_gained	3
variants_effect_synonymous_variant	15
variants_effect_upstream_gene_variant	392

Name: 954, dtype: object

Figure 7: The count of variants associated with one of our top 10 mutated genes

The genome wide association study output is a table of 32,414 SNPs, each row containing the specific base, genome location, and computed p-value. This table was filtered using Python's pandas package in a Jupyter Notebook to find the SNP's that had a p-value below the significance threshold. For the study, we initially set this threshold to 0.05. 1,598 SNPs were identified with p-values less than this threshold. However, because these comparisons are being tested simultaneously, the likelihood of incorrectly identifying a SNP as statistically significant increases. In order to compensate for this, Bonferroni correction was applied, so the new threshold was calculated by dividing the desired threshold by the number of comparisons: 0.05 / 32,414 = 1.54e-6. After filtering using this corrected threshold, it was found that 0 SNP's had p-values that could indicate statistical significance. The ten SNPs with the lowest p-values in the study are shown in Figure 8.

BP	A1	TEST	NMISS	OR	STAT	P
2463308	A	ADD	68	23.33000	3.685	0.000229
610604	C	ADD	70	15.67000	3.657	0.000255
971470	A	ADD	72	23.75000	3.630	0.000283
5165211	C	ADD	59	21.36000	3.597	0.000322
1711730	T	ADD	48	18.75000	3.594	0.000326
500989	C	ADD	66	14.14000	3.590	0.000330
1206051	T	ADD	71	23.47000	3.589	0.000332
5035860	T	ADD	46	16.00000	3.589	0.000333
610613	G	ADD	57	14.46000	3.563	0.000366
4438842	C	ADD	70	0.04061	-3.561	0.000370

Figure 8: The ten SNPs with the lowest computed p-values

The p-values from this table were then visualized using a QQ-plot to check if their distribution resembles data that is normally distributed. From the plot shown in Figure 9, there appears to be a skew of more extreme values, indicated by the longer tails at the ends. A Manhattan plot was also produced to visualize how these p-values vary across the genome. In Figure 10, each dot represents a single SNP, with the -log10 of the p-value on the y-axis and the genome location on

the x-axis. None of the SNPs hit the threshold for Bonferroni correction, indicated by the horizontal red line at the top of the plot.

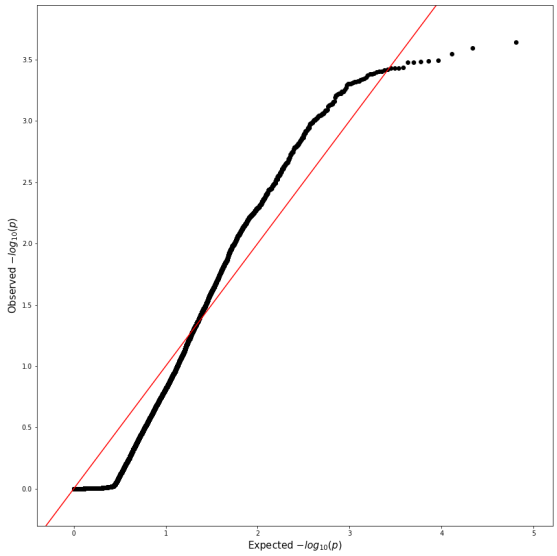

Figure 9: QQ-plot of computed p-values

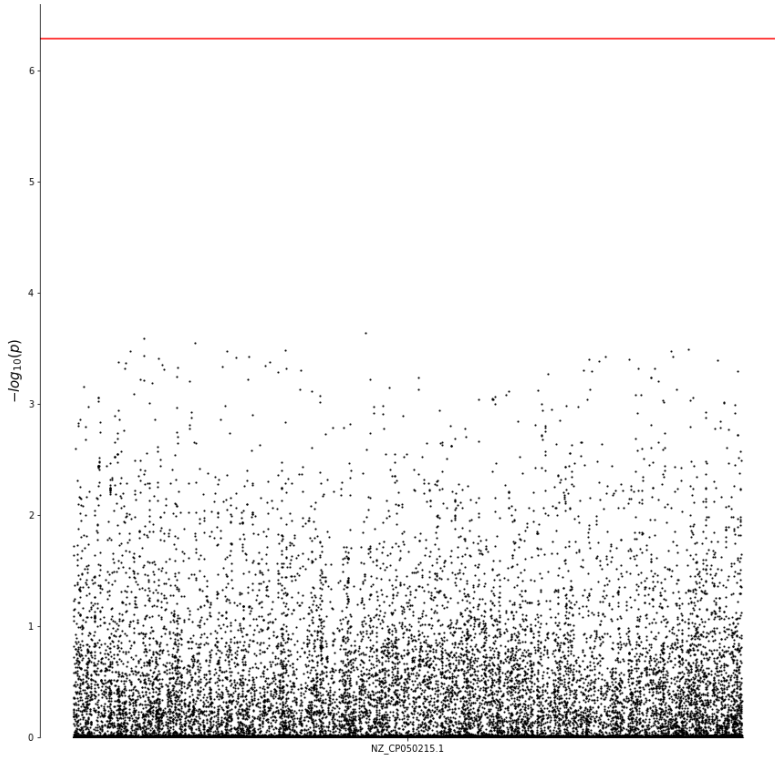

Figure 10: Manhattan-plot of SNPs across E.Coli genome

Conclusion

Our study aimed to explore and identify the genetic markers that characterize antibiotic resistance in ESBL-producing E. Coli. The project involved comprehensive processing of our genomic data, followed by a genome wide association study and analysis of found variants and their functional effects. Through our analysis, we noticed some observable differences through principal component analysis and variant annotation, but the location of the specific SNPs responsible for ESBL production is still unknown.

One of the primary limitations encountered was determining a definition for antibiotic resistance that could be feasibly tested using a GWAS. Although ESBL production has been proven to render E.Coli resistant to common antibiotics, there are many different physiological mechanisms that could also be used to classify other bacterial strains as antibiotic resistant. Another limitation was the limited sample size. This paired with the possible outliers from our principal component analysis could have given skewed results. If the study were to be repeated it would ideally involve hundreds, if not thousands of E.Coli samples. Time was another limiting factor, as a single sample takes approximately 60-70 minutes to be processed from raw reads to the Variant Call Format file needed for Plink and SnpEff. We also did not have access to a .gff file at the beginning of the study, which is an annotation file that is meant to be paired with the reference sequence used for alignment and variant calling. Because of this, we had to use two different reference genomes at different stages of the pipeline, which may have influenced the final GWAS and SnpEff outputs. In order to improve the logistic regression model, covariates such as %GC Content and more granular SNP filtering could be implemented.

This study is a small stepping stone toward addressing the much broader problem of antibiotic resistance. The project pipeline can be applied to any strain of bacteria with a case and control group which can not only help identify significant genomic variants, but also their effects on protein transcription. Understanding these markers for antibiotic resistance can better inform scientists of how such physiological changes occur and what can be done to prevent the emergence of new antibiotic resistant strains in the future.

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%20Modules/
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/.
15. "DNA Structure - The School Of Biomedical Sciences Wiki". *Teaching.Ncl.Ac.Uk*, 2021, https://teaching.ncl.ac.uk/bms/wiki/index.php/DNA_Structure?fbclid=IwAR00bHpJnec5eIT31PAJPRufMWzlob86-cJiY-rfdRGFC7Py3AT1ueOcZ3w.

Original Project Proposal

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