INVESTIGATING MULTIDRUG RESISTANCE IN *ESCHERICHIA COLI* WITH PHYLOGENETICS AND MACHINE LEARNING

by

David C. Brown

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics

Charlotte

2023

Approved by:

Dr. Daniel Janies

Dr. Jun-tao Guo

Dr. Alex Dornburg

Dr. Adam Reitzel

ABSTRACT

DAVID C. BROWN. Investigating Multidrug Resistance in *Escherichia coli* with Phylogenetics and Machine Learning. (Under the direction of DR. DANIEL JANIES)

The next pandemic is already underway in the proliferation of antimicrobial resistance (AMR) genes that reduce the effectiveness of therapeutics. The lessened potency of these current drugs results in increased economic costs, higher public health burdens, and greater loss of life when attempting to manage these increasingly treatment-resistant bacterial infections. Evolutionary principles guide this "silent pandemic", ultimately resulting in the development of superbugs, more formally described as multidrug resistant (MDR) bacteria. Such MDR phenotype bacteria resist three or more classes of antimicrobial compounds, representing significant obstacles to infection clearance and patient recoveries. To understand the forces driving the AMR pandemic, it is necessary to identify commonalities among bacterial genotypes with MDR phenotypes.

This dissertation aims to uncover the genetic determinants of MDR bacteria through a study of *Escherichia coli*. One hypothesis for the development of MDR phenotype bacteria theorizes that resistance results from an increased number of mutations. Researchers refer to these highly mutable strains as possessing hypermutator phenotypes, often attributed to poorly functioning mismatch repair systems. One specific example of flawed mismatch repair in hypermutator bacteria was identified by LeClerc et al., 1996, defined as *E. coli* strains with a deficient Mutator S gene (gene *mutS* encodes the protein MutS).

First, I analyzed the *mutS* genes from 817 high-quality E. coli isolates using phylogenetic comparative analyses. Although I observed 271 MDR isolates in this data set, I found no evidence for the existence of a deficient *mutS* gene as defined by LeClerc. Additionally, when modeling the coevolution of an MDR phenotype against all variant amino acid positions in the *mutS* gene, the evidence supports a largely independent evolution between MDR and the *mutS* mismatch repair gene.

Next, to better understand this confounding result, I undertook a statistical analysis of the whole genome sequences (WGS) for all *E. coli* isolates. I used supervised machine learning to identify the genetic annotations that best predict the development of resistance to several classes of antimicrobial compounds: aminoglycosides, folate pathway antagonists, macrolides, tetracyclines, and other. The five trained random forest estimators achieved a mean ROC AUC of 0.87 ± 0.04 on 66 features, engineered from 5,511 annotated genes in the calculated *E. coli* pangenome. I determined that the top performing features did not include the *mutS* mismatch repair gene, further confirming the results of my first investigation. Instead, I found evidence that genes associated with horizontal gene transfer (HGT) best predict MDR phenotypes, an idea called into question by LeClerc et al., 1996.

Finally, I examined the component annotations which comprised the most important engineered features. Interestingly, I found that resistance to a given class of antimicrobial treatments results from a unique and specific pattern of annotated genes that does not include commonly understood genetic determinants of resistance. MDR is best predicted, not by AMR genes themselves, but by accessory genes often involved in the horizontal and lateral transfer of genetic information. My investigation supports the current domain understanding, that horizontal gene inheritance mechanisms drive the proliferation of antimicrobial resistance, by indicating gene sets which predict resistance to specific categories of antimicrobial compounds.

This work demonstrates the importance of combining phylogenetic methods and statistical modeling tools like machine learning to arrive at working hypotheses for polygenic traits. Additionally, this investigation portrays the research value of survey data. By identifying unique mechanisms for the continued proliferation of different resistance classes, this dissertation addresses the root evolutionary drivers of the AMR

pandemic. In future, further investigation of the evolutionary history of the specific genes responsible may lead to new therapeutic targets or improved prescription strategies.

DEDICATION

To my wife, Rhiannon, thank you for your unending support. As it has been said, "It's the possibility of having a dream come true that makes life interesting."

ACKNOWLEDGEMENTS

My gratitude to Daniel Janies for aiming me towards interesting work and supporting me while I researched in his lab. I recognize the members of my committee: Dr. Juntao Guo, Dr. Alex Dornburg, and Dr. Adam Reitzel; without your expertise and direction, I could not have completed this process. Many thanks to Colby Ford, for encouraging me to develop the temperament of a true scientist. Much appreciation to Denis Jacob Machado for persisting with me when approaching difficult problems together. I am thankful for Way Sung's passion for research, and that he chose to share that drive with me early in my graduate career. Without Dr. Russel-Pinson, my writing process would not exist, thank you for all that you do. I offer credit to Lauren Slane for helping me navigate the forms and steps towards graduation; I greatly appreciate it. I am indebted to Jon Halter and the rest of the Office of OneIT URC staff for their expertise with the HPC and patience with my questions. Further gratitude to the sources of funding provided to me at various times by GAANN, GASP, and the U.S. ONR. I am deeply grateful for the time and energy invested in me by all my instructors throughout my long education. To all those learning, searching, and contributing to our collective knowledge online, I stand here on your shoulders.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3GCR	third-generation cephalosporin-resistance
E. coli	*Escherichia coli*
mutS	gene that encodes the MMR protein Mutator S
*tra***	gene that encodes transfer protein **, part of a transfer (tra) operon
AA	amino acid
ABR	antibiotic resistance
AIC	Akaike information criterion
AMR	antimicrobial resistance
AP	average precision
AR Threats	Antibiotic Resistance Threats, report published by the CDC
AUC	area under the curve
BLAST	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
CDC	Centers for Disease Control and Prevention
CLSI	Clinical and Laboratory Standards Institute
CR	carbapenem-resistance
CRE	carbapenem-resistant Enterobacteriaceae
CSV	comma-separated values file format, suffix .csv
DNA	deoxyribonucleic acid

DTN	data transfer node
F factor	fertility factor plasmid
FASTA	fast A file format
FDA	U.S. Food & Drug Administration
GenBank	National Institutes of Health (NIH) genetic sequence database
HGT	horizontal gene transfer
HPC	high-performance computing
Inc**	plasmid incompatibility group **
LGT	lateral gene transfer, synonymous with HGT
MDR	multidrug resistance
MGE	mobile genetic element
MMR	DNA methyl-directed mismatch repair system
multiFASTA	FASTA file including multiple sequences
MutHLS	biological MMR pathway formed by the protein complex of Mutators H, L, and S
MutS	MMR protein Mutator S
NAAT	nucleic acid amplification test
NARMS	National Antimicrobial Resistance Monitoring System for Enteric Bacteria
NCBI	National Center for Biotechnology Information

NCBI PD	NCBI Pathogen Detection system
NCBI PDIB	NCBI PD Isolates Browser, user interface for NCBI PD data
PCR	polymerase chain reaction
PGAP	Prokaryotic Genome (Automatic) Annotation Pipeline, formerly PGAAP
PR	precision recall
RefSeq	NCBI Reference Sequence Database
ROC AUC	area under the receiver operating characteristic curve
SQL	Structured Query Language
TA	toxin/antitoxin
TBLASTN	protein query against translated nucleotide BLAST search
Tra**	transfer protein **
TSV	tab-separated value file format, suffix .tsv
UN	United Nations
WGS	whole genome sequence
WHO	World Health Organization

CHAPTER 1: INTRODUCTION

1.1 The Silent Pandemic of Antimicrobial Resistance

In 2019, the Centers for Disease Control and Prevention (CDC) estimated there were 2.8 million antibiotic-resistant infections in the United States, resulting in 35,000 domestic deaths [1]. Antibiotic resistant infections are not limited only to the United States but instead represent a global healthcare concern. A recent report to the Secretary-General of the United Nations (UN) estimates that at least 700,000 individuals die worldwide from drug-resistant diseases each year, predicting that total could climb to 10 million deaths by 2050 [2]. The continuously increasing prevalence of antibiotic resistance phenotypes among pathogenic, or disease-causing, bacterial species drives these rising death tolls.

While antibiotic resistance is the common term for bacteria with drug-resistant phenotypes, the general trend in both academic and medical literature broadens the scope of the conversation to antimicrobial resistance (AMR). While some disagree with conflating the two terms [3], AMR leads as the currently preferred term [3, 4, 5, 6, 2, 7]. Therefore, in this dissertation, I use the term AMR when referring to microbial phenotypes that impart drug resistance. In bacteria, the presence of a specific AMR gene imparts the associated resistance phenotype [8]. Given this context, use of the term AMR throughout this manuscript relies on the above understanding, that simply the presence of an AMR gene within a bacterial genome indicates an AMR phenotype.

Research has long associated AMR with pathogenic bacteria, but not exclusively so. The existence of AMR genes among non-pathogenic bacteria is well-documented [9, 10, 11, 12]. The problem of AMR is pervasive, not limited to health care or geographic environments [10, 13, 14]. Many bacteria carry AMR genes, whether they remain benign or cause disease. Recently termed the “silent pandemic” [15], AMR genes continue to occur with increasing abundance among global bacterial strains. Simultaneously, AMR gene assortment also increases, as genetic variations continue to develop in order to resist various classes of antimicrobial drugs [16, 17, 18, 19, 20].

In the past, terms like “classes of antimicrobial drugs” and “multidrug-resistance” were variably defined [21]. For the purposes of this dissertation, I chose current definitions for both terms from the CDC and National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS). CDC NARMS categorizes antibiotics based on the Clinical and Laboratory Standards Institute’s twelve (12) antibiotic drug class definitions [22]. The CDC defines multidrug resistant (MDR) bacterial infections as infections that resist three or more of the CLSI antibiotic classes [23]. For the purposes of this dissertation, AMR refers to the quality of imparting resistance to antimicrobial substances, while MDR refers to a specific threshold (three or greater classes of resistance) within that broader definition of AMR.

1.2 Antimicrobial Resistant to Multidrug Resistant Phenotypes

MDR bacterial lineages preferentially gain AMR genes over time [5], a phenomenon representing the logical end stage of the silent AMR pandemic. MDR bacterial infections, commonly referred to as superbugs [24], involve serious challenges for human health in hospital settings [25]. Healthcare-associated infections, also known as nosocomial infections [26] often involve MDR pathogenic phenotypes. Nosocomial infections link to elevated healthcare costs, morbidity, and mortality [27, 28], so surveillance of the bacteria which cause nosocomial infections remains a top priority [26]. Two main propositions exist for the proliferation of MDR in healthcare conditions: mutation accumulation and horizontal gene transfer (HGT). Either MDR developed through rapid mutation within the highly competitive, antimicrobial-saturated environment of a hospital [29, 30, 31], or the arrival of the patient allowed the acquisition of groups of AMR genes which developed under external conditions [32, 33, 34, 35].

1.2.1 Hypermutator Phenotypes and Deficient Mismatch Repair

While the presence of an AMR genotype indicates the presence of an AMR phenotype [8], other mutations have been documented that increase virulence [36] or improve the mutational ability of a lineage [31]. A heightened rate of mutation to adapt to environmental stressors is called hypermutation [29], a property associated with a hypermutable phenotype [37]. The manuscript published in 1996 by LeClerc et al. theorized that hypermutator phenotypes result in MDR phenotypes. LeClerc suggested a mechanism for hypermutator phenotype *E. coli*: a defective Mutator S (*mutS*) gene. The *mutS* gene forms part of the DNA methyl-directed mismatch repair (MMR) pathway, a system which proofreads mistakes and corrects errors in the nucleic code following replication. LeClerc indicated that a deletion of 221 bp from the 3' end of the *mutS* gene results in a defective *mutS* gene, and asserted that such a defect leads to the hypermutator phenotype observable in *E. coli* [37]. I examine LeClerc's specific definition of hypermutator in this dissertation.

1.2.2 Horizontal Gene Transfer and the Mobilome

Previous understanding of bacterial evolution assumed that, like most species, lineages inherited genes vertically [38, 39]; however, researchers now acknowledge that horizontal (or lateral) transmission of genetic information affects enteric bacterial genomes, in addition to the vertical inheritance of genes [40, 41, 42]. These HGT events allow species of enteric bacteria to pass genetic information among and between lineages through mobile genetic elements (MGEs). The term mobilome refers to MGEs and their associated genes [43]. The existence of the mobilome indicates that enteric bacteria readily acquire and transmit AMR genes through HGT, often on MGEs such as plasmids [44, 45, 35].

With time and research, our domain understanding of HGT, MGEs, and plasmids continues to improve, but the fundamental concepts are not new. LeClerc et al. 1996, specifically proposed their theory of hypermutation in opposition to the understanding of HGT from that era. This dissertation aims to examine LeClerc's definition of hypermutator phenotype as *E. coli* with defective *mutS* genes. Testing LeClerc's hypermutator hypothesis leads to one of two competing outcomes. First, falsifying LeClerc's definition calls into question the role of hypermutation in MDR, and weakens hypermutation in favor of HGT for proliferating AMR genes into MDR phenotypes. Conversely, an inability to falsify LeClerc's definition emphasizes the importance of monitoring hypermutable bacteria as a means to control MDR threats.

1.3 The Complex Problem of Multidrug Resistance

In Enterobacteriaceae research literature, both HGT and specific MGEs are strongly associated with the proliferation of AMR genes and consequently the acquisition of MDR phenotypes [46, 47]. Several additional mechanisms compound the uncertainty resulting from these two competing theories.

1.3.1 Reticulate Evolution and AMR

Reticulate evolution, the term for the non-independent nature of the complex evolutionary interactions between two distinct lineages, includes processes like HGT, hybridization, and recombination [48]. The current threat of MDR superbugs exacerbates the difficulty in describing these reticulate evolutionary patterns, especially among Enterobacteriaceae where acquired genes, originating from different species, circulate contemporaneously [44, 49, 50].

The phenomenon of reticulate evolution is an acknowledged concern for *Salmonella* genetics [8], and while a similar species, *E. coli* has not been investigated as thoroughly. Conservative estimates suggest that about a quarter of the *Salmonella* genome was derived by reticulate evolutionary forces [41]. Estimates for *E. coli* indicate a minimum of 12.8% of current protein-coding DNA is foreign, while not accounting for any prior evolutionary history of the species' genome [40]. According to recent research, 64.5% of *E. coli* are third-generation cephalosporin-resistant (3GCR) and 5.8% are carbapenem-resistant (CR); projections indicate that these values increase to 77% for 3CGR and 11.8% for CR *E. coli* by 2030 [51]. Due to reticulate evolutionary forces, these and other AMR genes are not exclusively available to *E. coli* but to all Enterobacteriaceae.

1.3.2 AMR Genes, Xenogenetic Elements, and MDR

Xenogenetic elements are invasive, mobilized AMR genes with replication capabilities [17]. Current evolutionary forces at work in *E. coli* accumulate AMR, accessory, and virulence genes into broadly drug resistant phenotypes [52]. As such, MDR phenotypes also fit the criteria of xenogenetic elements, regardless of their horizontal or vertical method of inheritance. Generally, MDR phenotypes resist multiple antimicrobial drug classes through several different AMR genes or mechanisms. Hypothetically, certain MDR phenotypes could associate with specific, limited sets of accessory and virulence genes, which I will refer to as gene sets. These gene sets would indicate specific genes or groups of genes that associate with MDR inform an understanding of the stepwise progression towards a specific, multiply-resistant, bacterial phenotype. I believe that this concept of gene sets indicates unique hereditary pathways for MDR phenotypes. The increasingly commonplace occurrence of MDR bacterial lineages threatens global food safety and public health [53]. A greater understanding of MDR *E. coli* within this global context is needed to better protect human lives.

1.4 One Health Initiative

While both proliferation concepts aim to describe the process of AMR gene accumulation, neither the clinical setting or external environment should be examined in isolation. The Manhattan Principles developed by the Wildlife Conservation Society encourage scientists to “take the complex interconnections among species into full account” [54]. These interconnections allow researchers to “[r]ecognize the essential link between human, domestic animal and wildlife health and the threat disease poses to people, their food supplies and economies…" [54].

The Manhattan Principles led to the subsequent development of One Health [54, 55]. One Health seeks to increase collaboration between various disciplines of academia, human healthcare, veterinary medicine, and public health to "promote, improve, and defend the health and well-being of all species" [55]. A main aim of One Health seeks to establish "[j]oint cross-species disease surveillance and control efforts in public health…" [55]. This specific aim of One Health has proven useful for both pandemics and microbial pathogen surveillance involving food safety [56].

1.4.1 GenomeTrakr

The U.S. Food and Drug Administration (FDA) established the GenomeTrakr laboratory network in collaboration with public health groups and academic institutions to further this One Health aim [57, 58]. Inspectors gather surveillance swabs and send samples to network labs for sequencing. These WGS data are then made publicly accessible in the National Center for Biotechnology Information's (NCBI) online databases. The web-based NCBI Pathogen Detection Isolates Browser (NCBI PDIB) serves as the main interaction point for international collaborators to access the data. These WGS data represent a valuable resource for food safety surveillance.

1.5 Food Safety and *Escherichia coli*

Many researchers and health organizations point to the misuse of antimicrobials in agricultural practices as a fundamental driver for AMR proliferation [59, 53]. AMR genes often impart a fitness advantage, for example lineages with AMR genes outcompete other lineages in the presence of antimicrobial compounds [60, 61, 62]. Monitoring livestock species for pathogenic bacteria is an important component of food safety surveillance. Food-related illnesses are not uncommon globally, so current systems of health surveillance sample food and identify pathogenic bacteria by WGS techniques.

The WGS data gathered by these surveillance systems provide insights into the evolutionary history of pathogens and the functionality of MDR traits. One group of pathogens associated with both food and MDR is Enterobacteriaceae, a family of Gram-negative bacilli [63].

1.5.1 Enterobacteriaceae

Specific pathogens within Enterobacteriaceae include Salmonella, Klebsiella, Shigella, and Escherichia coli (*E. coli*). The organisms within the family Enterobacteriaceae, often referred to as enteric bacteria, colonize the gastrointestinal tract of humans and other mammals. Enteric bacteria are a special cause for concern regarding MDR phenotypes due to their genetic flexibility, which allows them to survive in a broad range of environments.

Enteric bacteria colonize many segments of the farm-to-fork continuum, where they encounter substantially varying environmental conditions [64, 65]. Environmental variation drives the selection of specific heritable mutations which impart a fitness advantage. Gene transmission mechanisms improve the host range of enteric bacteria [66]. Host shifting from livestock species to humans and back necessitates different sets of genetic adaptations for survival.

Researchers often focus on Salmonella when addressing concerns of both food safety and MDR [67, 68, 69, 70, 71]. However, despite the recognized importance of monitoring *E. coli* in food safety and healthcare, little to no understanding exists of the broader, global MDR patterns for *E. coli* [72]. I chose to utilize *E. coli* data in order to address this gap in current knowledge.

1.6 Problem Summary - Resistance and Food Safety

Antimicrobial resistance genes diversify and proliferate as a silent pandemic [7, 15]. As bacterial lineages amass multiple antimicrobial resistance genes, in some environments gain a fitness advantage and become multidrug resistant [73]. While antimicrobial resistance may be examined under the separate lenses of laboratory, clinical, or food safety contexts, a lack of research examines the broader, interconnected threat MDR represents towards global public health. A One Health perspective focused on pandemic control cannot ignore this interconnectedness.

The data gathered by One Health-focused initiatives, like GenomeTrakr, monitors food for potential microbial pathogens [58]. Specifically within the area of food safety, effort focuses on Salmonella, with *E. coli* less explored. The silent pandemic of AMR, the usefulness of One Health concepts for pandemic control, and the relatively underexplored niche of *E. coli* when taken together represent an opportunity for novel research.

1.7 Purpose

The purpose of this research program aims to examine the evolutionary history and proliferation of the multidrug-resistance pandemic in Escherichia coli from both phylogenetic and correlational perspectives (i.e. machine learning). In Chapter 2, I employ phylogenetic comparative methods to examine the role of hypermutation in the evolutionary development of MDR through an analysis of the methyl-associated mismatch repair (MMR) protein Mutator S described by LeClerc et al. 1996. In Chapter 3, I use supervised machine learning to reveal the genetic annotations that best predict an MDR phenotype, comparing those predictors against the two prevailing theories of MDR development: hypermutation or HGT. With Chapter 4, I describe the biological significance of the genetic annotations that best predict an MDR phenotype.

CHAPTER 2: PHYLOGENETIC COMPARATIVE ANALYSIS OF HYPERMUTATION AND MULTIDRUG RESISTANCE

2.1 Introduction

The phenomenon of increasing bacterial antimicrobial resistance (AMR) genes is often described as a pandemic [3, 4]; this proliferation of AMR genes among bacterial lineages logically results in multidrug resistance (MDR) phenotypes, with resistance to three or more classes of antimicrobial drugs [74]. Understanding the epidemic and pandemic trajectories of pathogens and the gain, loss, and spread of disease causing genotypes and phenotypes requires such evolutionary histories displayed on phylogenies [5]. Phylogenies investigate and express hypotheses about the evolutionary history of organisms [75, 76]. I chose to investigate the evolutionary history of MDR through the lens of *Escherichia coli* food safety data due to 1) the amount of well curated whole genome sequence (WGS) and phenotypic data, 2) the ubiquity of *E. coli*, 3) the global nature of the survey data set, and 4) the common association of pathogens with MDR phenotypes.

2.1.1 Hypermutation

MDR phenotypes have long been linked with the terms “hypermutable”, “hypermutation", and “hypermutator” [77, 37, 78]. Hypermutable bacterial phenotypes possess deficient methyl-directed mismatch repair (MMR) systems that generate significantly more mutations than wild-type bacteria [79, 80, 37, 78, 81, 82, 83]. The widely accepted concept became that highly mutable, MMR deficient bacteria more easily adapt when challenged with antibiotics, allowing mutator strains to develop resistance to more compounds [84, 77, 37, 78], eventually resulting in MDR. Some disagree about the role of hypermutation with MDR, pointing to research in uropathogenic *E. coli* that could not link mutators with MDR [85], an investigation in enterococci that MMR mutations do not associate with MDR [86], and other studies [87] that did not provide experimental evidence linking the hypermutable, MMR-deficiency phenomenon with MDR. The goal of my dissertation is to advance this conversation about the cause or causes of MDR phenotypes. To begin, I chose to investigate an early, well-defined example of hypermutable, MMR-deficiency noted by LeClerc et al. 1996 involving Mutator S (gene *mutS* encodes the protein MutS).

MutS is one protein component of the *E. coli* MMR system, initially associated with hypermutation by LeClerc et al. 1996, who indicated a deletion of 212 bp in the 3' end of the MutS protein as responsible for causing a mismatch repair deficient, hypermutable phenotype [37]. As noted above [84, 77, 78], the MutS deficient concept introduced by LeClerc drove a substantial volume of research linking elevated mutation rates with hypermutable mismatch repair deficient phenotypes and AMR genes. The consequence, fully intended by LeClerc et al. 1996 [37], being that hypermutation is viewed in opposition to other mechanisms for MDR phenotype proliferation, like another prominent hypothesis involving the horizontal gene transfer (HGT) of AMR genes.

2.1.2 AMR, MDR, and Food Safety Data

Within the broader healthcare concerns about AMR progression towards MDR bacteria, attention focuses on maintaining the safety of the global food supply [88, 89]. Spearheaded by the FDA and assisted by other domestic and international organizations, the GenomeTrakr program seeks to fulfill the mission of keeping food safe [58]. GenomeTrakr laboratories monitor supply lines for the presence of bacterial pathogens through routine sampling and genome sequencing. GenomeTrakr partners collect and submit WGS data to the NCBI Pathogen Detection (NCBI PD) database [90]. The NCBI PD separates sequences by species, organizing isolates in phylogenetic trees to track and monitor foodborne bacteria for the presence of AMR and virulence genes [91, 92]. Several species of Enterobacteriaceae are among the many food associated bacteria monitored by the NCBI PD database.

In 2019, the CDC published an Antibiotic Resistance Threats (AR Threats) report, listing both typhoidal and nontyphoidal subtypes of *Salmonella* as "serious threats" [1] among the enteric bacteria. The data submitted by GenomeTrakr labs to the NCBI PD is collected to understand the evolution and spread of AMR pathogenic *Salmonella* [8]. Recent publications by the FDA indicate both the role of HGT and reticulate evolution in the development of the *Salmonella* genome, while also pointing to the broader usefulness of the NCBI PD data for monitoring and tracking pathogenic AMR genes [88, 58, 8].

The AR Threats report separates microbes into three categories of decreasing severity: urgent, serious, and concerning. Serious threats, like *Salmonella*, "are public health threats that require prompt and sustained action..." [1]. Of higher concern to human health, urgent threats "are public health threats that require urgent and aggressive action..." [1]. Urgent threats include the Carbapenem-resistant Enterobacteriaceae (CRE) group, which includes *E. coli.* Some lineages of enteric bacteria produce a carbapenemase that imparts resistance to carbapenems, cephalosporins, and penicillins, a multiply resistant phenotype termed the "nightmare bacteria" [93]. As such, the study of MDR *E. coli* is of utmost importance to public health.

2.1.3 *Escherichia coli Data Source*

The utility of *E. coli* as a well-understood laboratory model species for AMR development [61, 12] continues into the present day via computational techniques [94]. Research on *E. coli* has investigated AMR genes within the context of food safety [95, 89]. Also of evolutionary and regulatory interest is the finding that lineages of *E. coli* gain more AMR genes than they lose over time, a finding termed genetic capitalism [96, 97, 98]. Our publication, Ford et al., 2020 [5], demonstrated that many

E. coli lineages evolve under genetic capitalism instead of stabilizing selection, where isolates both gain and lose resistance over time. We extended the concept of genetic capitalism to show that variation in the gain phenomenon depends on the functional class of antimicrobial resistance. Our results indicated that positive selection for resistance overcomes the potential fitness costs bacteria incur when carrying AMR genotypes, allowing lineages under genetic capitalism to continue gaining AMR genes [96, 98, 5]. *E. coli* is therefore a valuable and well-understood model species to study the context dependent accumulation or regression of AMR phenotypes.

This study aims to understand the relationship between hypermutation and the AMR pandemic, by investigating phylogenetic correlations between genotype and phenotype through *mutS* and MDR phenotypes. To examine the relationship between *mutS* and MDR, I test whether hypermutable, repair deficient MutS variants (as defined by LeClerc et al., 1996 [37]) correlate with MDR phenotypes (as defined by the CDC [1]), my null hypothesis stating no relationship between deficient *mutS* and MDR.

2.2 Materials and Methods

2.2.1 Materials

2.2.1.1 Raw Data

For the Ford et al., 2020 study, we performed a search inclusive for both *E. coli* and Shigella on October 08, 2018 in the NCBI PDIB [5]. The search returned 29,255 isolates, viewable as a table from the NCBI PDIB, which I downloaded as a tab-separated value (.tsv) file. This file (hereafter, the metadata file) contains the identifying information for each isolate, as well as many types of non-genomic information associated with the uploaded sequences (region, source, sequencing platform, etc.). At time of publication, an explanation of the file contents is found at https://www.ncbi.nlm.nih.gov/pathogens/pathogens_help/#data-fields [Internet]. The file contains a column of sequence accession numbers, which match GenBank and RefSeq records.

In this manuscript, I use the sole term *E. coli* to identify these isolates, replicating the organizational scheme employed by NCBI Pathogen Detection. To reduce the total number of isolates and increase confidence in the conclusions of this dissertation, I filtered the accession numbers from the .tsv metadata file using custom scripts written in the Python language [99] distributed by Anaconda version 4.11.0 [100] to identify *E. coli* isolates (NCBI:txid562 found in column *species_taxid*) assembled to the Complete Genome or Chromosome designation levels defined by GenBank, thus separating out contigs and scaffolds that represent incomplete or partial genomes. This filtering resulted in 911 RefSeq accession numbers of *E. coli* with assembled WGS (schema GCF_#########). I used the accession numbers to download the WGS files in FASTA format from the NCBI FTP site via the UNC Charlotte Data Transfer Node (DTN). While I indicated all 911 accessions for download, several server requests failed for a variety of reasons, leaving 875 assembled, RefSeq-hosted, *E. coli* WGS FASTA files. I stored these data files on the UNC Charlotte High-Performance Computing (HPC) cluster. Further information on hardware specifications, Anaconda environments, and detailed program versions may be found in the Appendices. Code is available upon request.

2.2.1.2 Processed Data Set

In summary, I identified a subset of data from our earlier study, Ford et al., 2020 [5], involving about 29,255 isolates of *Escherichia coli* and *Shigella* accessed from the NCBI PDIB on October 08, 2018. I subset that data for my investigation according to the following selection criteria:

- Sequences of *Escherichia coli* marked by NCBI:txid562.
- Sequences assigned to the two highest levels of assembly (Complete Genome

and Chromosome) recognized by the NCBI.

- Sequences have a RefSeq identifier (schema: GCF_#########).

The application of these criteria resulted in 911 candidate RefSeq accession identifiers within the Ford et al., 2020 [5] data set. I downloaded the available whole genome sequences for the isolates to the UNC Charlotte high performance computing (HPC) cluster via the NCBI FTP site using the RefSeq accession identifiers, resulting in 875 successfully completed downloads. I took the antimicrobial resistance properties of these successfully downloaded isolates from the table we gathered in Ford et al., 2020 [5] from the NCBI PDIB.

2.2.2 Methods

2.2.2.1 Core Genome Calculation and Phylogenetic Tree

Comparison of WGS information requires annotated, quality-controlled, and aligned data. I quality-controlled the 875 successfully downloaded sequences with the program CheckM version 1.2.1 [101, 102, 103, 104] to reduce contamination of the data set with non-*E. coli* genetic sequences. I used the program CheckM version 1.2.1 [101, 102, 81, 104] to quality control the downloaded whole genome sequences for both completeness and contamination. I discarded sequences with less than 99.0% (exclusive) completeness OR (logical) greater than 1.0% (exclusive) contamination, as unlikely to represent high-quality *E. coli* whole genome sequences. Implementation of CheckM led to a final data set of 817 *E. coli* whole genome sequences, which I passed to the program Prokka version 1.12 [105] for *de novo* gene annotation.

After Prokka annotation, I employed the program Roary version 3.12.0 [106] to identify the pangenome of the *E. coli* isolates. The Roary program calculates a core genome, a set of genes where each gene is found in greater than or equal to 99% of the sequences passed as input [106]. Roary creates an alignment file for the core genome, based on a given set of annotations from Prokka [106, 105]. I passed the core genome alignment file created by Roary to the RAxML version 8.2.12 [107] program to generate a maximum likelihood phylogenetic tree. Maximum likelihood methods produce unrooted phylogenies [107], so I used the program Gotree version 0.4.3 [108] to root the RAxML output. I rooted the RAxML tree on the RefSeq sequence GCF_003287245 (NCBI PD accession number PDT000343334.1, also GenBank accession GCA_003287245.1) chosen by Ford et al., 2020 [5] as representative of an ancestral *E. coli.*

2.2.2.2 Identifying MDR from AMR

The investigation proposed here requires the binning or characterization of AMR phenotypes into a limited number of classes for both phylogenetic and machine learning analyses. I explain these analyses in more detail in their corresponding chapters. The NCBI Pathogen Detection database records the AMR genes associated with an isolate under the *AMR_phenotype* column in the .tsv metadata file. While many of the isolates are not clinically tested to determine levels of resistance, the presence of an AMR gene within Enterobacteriaceae genomes equates to the existence of a resistance phenotype [8]. I used custom scripts in the Anaconda version 4.11.0 [100] distribution of Python 3, versions 3.9 and 3.10, [99] to parse the metadata file and identify families of AMR genes associated with each isolate, often employing code from Biopython [109]. I grouped individual AMR genes by the three-letter prefix of each gene name. I referenced the Comprehensive Antibiotic Resistance Database (CARD) [110] to identify the antimicrobial compounds that a three-letter family group resists.

Upon further analysis, I noted the presence of the *bla* or beta-lactamase AMR family genes in nearly every isolate. To maintain conservative estimates of both AMR and MDR, I therefore excluded the *bla* family of AMR genes from considerations of resistance. I then grouped these families into representative AMR phenotypes commonly tested for by CDC NARMS under CLSI guidelines [23]. I assigned each isolate binary labels for MDR (1 for resistant to three or more drug classes exclusive

Table 2.1: Examples of the resistance categories and annotated resistance genes present in the *E. coli* data set.

Resistance Category	**3 Letter Resistance Gene Prefix**
aminoglycosides	*aac, aad, ant, aph, arm, rmt*
beta-lactam combination agents	*bla*
cephems	*abc*
folate pathway antagonists	*dfr, sul*
macrolides	*ere, erm, mef, mph, msr*
nucleosides	*sat*
penicillins	*amp*
phenicols	*cat, flo*
quinolones	*qep*
tetracyclines	*tet*
others	*arr, ble, cml, fos, lnu, mcr, oqx, qac*

of beta-lactam antibiotics, 0 for not resistant to three or more drug classes exclusive of beta-lactam antibiotics) and AMR. In this data set of 817 *E. coli* , I found 546 isolates (66.8%) without MDR (trait label 0) and 271 isolates (33.2%) with MDR (trait label 1). I used these MDR phenotype characters as input to map the progression of MDR on a phylogenetic tree, here, in Chapter 2, and modified them into the predicted labels for supervised machine learning analysis in Chapter 3.

2.2.2.3 Trait Characterization - MDR

I created a binary character matrix to represent presence or absence of the MDR phenotype by employing the CDC definition of MDR as microbial resistance to three or more antibiotic classes [74]. I identified families of antimicrobial resistance genes by the 3-letter prefix for all gene names found in the *AMR_phenotype* column of the NCBI PDIB table (see Table 2.1). I assigned gene families to a single, exclusive antimicrobial class by manually cross-referencing the list of CLSI compounds screened by NARMS [23] with information from the Comprehensive Antibiotic Resistance Database (CARD) to arrive at these assignments. I assigned an isolate with less than three (3) classes of antimicrobial compounds a 0 (absence) in the MDR character matrix, while in the positive case (presence of MDR) I assigned a 1. The final MDR character matrix has dimensions 817 x 1, representing the binary state of the MDR character for each isolate.

2.2.2.4 Trait Characterization - MutS

The question of *mutS* trait characterization is more complex. Roary can perform context-dependent paralogous splits, but for this inquiry, I did not enable that functionality [106]. I kept paralogous genes together in order to simplify the analysis: focusing solely on the MutS sequences, instead of potential controller regions up- and downstream of the annotated gene. I used helper scripts provided by the Microbial Genomics Lab at the Center for Bioinformatics and Integrative Biology [111] to organize an SQLite database of annotated chromosome references and generate a MutS multiFASTA file.

For clarity, I counted all three of the *mutS*, *mutS_ 1*, and *mutS_ 2* gene annotations from Prokka and Roary as potential variants of the *mutS* gene. This dissertation refers to that collective grouping as *mutS*. To identify the type of data for the *mutS* trait, I accessed the SQL database to extract the sequences of all annotated *mutS* genes into a single aligned multiFASTA file. I observed a *mutS* gene in 816 of the 817 isolates, so a binary character indicating deficient or functional *mutS* by LeClerc's definition was not be useful. To confirm if the observed incidence of hypermutators in my data set matched the expected hypermutator prevalence of 1% [79, 37], I performed a Chi-Square Goodness of Fit Test [112] at 0.05 p-value level of statistical significance.

To prepare for the possibility that the observed incidence of hypermutator *mutS* did not match the theoretical distribution, I translated the *mutS* nucleic sequences into protein space and aligned them again using MAFFT version 7.273 [113], resulting in an aligned protein multiFASTA file. I identified variant positions in this MutS protein multiFASTA file using the program `snp-sites` version 2.5.1 [114]. Out of 853 total amino acid (AA) positions for *E. coli* MutS, I located 141 variant sites. I simplified this categorical, protein space data to a binary state by developing a binary character matrix to represent an isolate's agreement with the consensus residue at a given sequence position. A 1 represents an exact match between the residue of an isolate and the consensus AA residue, while a 0 represents the presence of any other residue besides that found in the consensus sequence at the specified position. I allowed the consensus sequence to include missing residues, to account for potential cases of extended or truncated protein sequences in the variant positions. The final character matrix has dimensions 816 x 141, representing agreement (1) or disagreement (0) with the consensus sequence for each isolate at 141 positions with varying residues in the MutS protein. Depending on the results of the Chi-Square Goodness of Fit Test, I could remove a row from the MDR trait, resulting in a 816 x 1 matrix to match the MutS trait.

2.2.2.5 Phylogenetic Comparative Methods

I imported the phylogenetic tree to Mesquite, version 3.70 (build 940), [115] and mapped binary character matrices representing the MDR and *mutS* traits for each isolate, resulting in a broad but shallow phylogeny that was difficult to adequately visualize. To arrive at a more quantitative result, I also analyzed the independent evolution between the binary MDR character and all 141 MutS characters using Pagel's 1994 algorithm [116]. I chose Revell's implementation of the algorithm included with the package phytools version 1.2 [117] written in R version 4.2.2 [118]. This implementation enabled me to compare four models at each of the 141 positions: complete independence, MutS dependent on MDR, MDR dependent on MutS, and interdependence. I assessed the relative strength of each of the four models at each position using the Akaike information criterion (AIC).

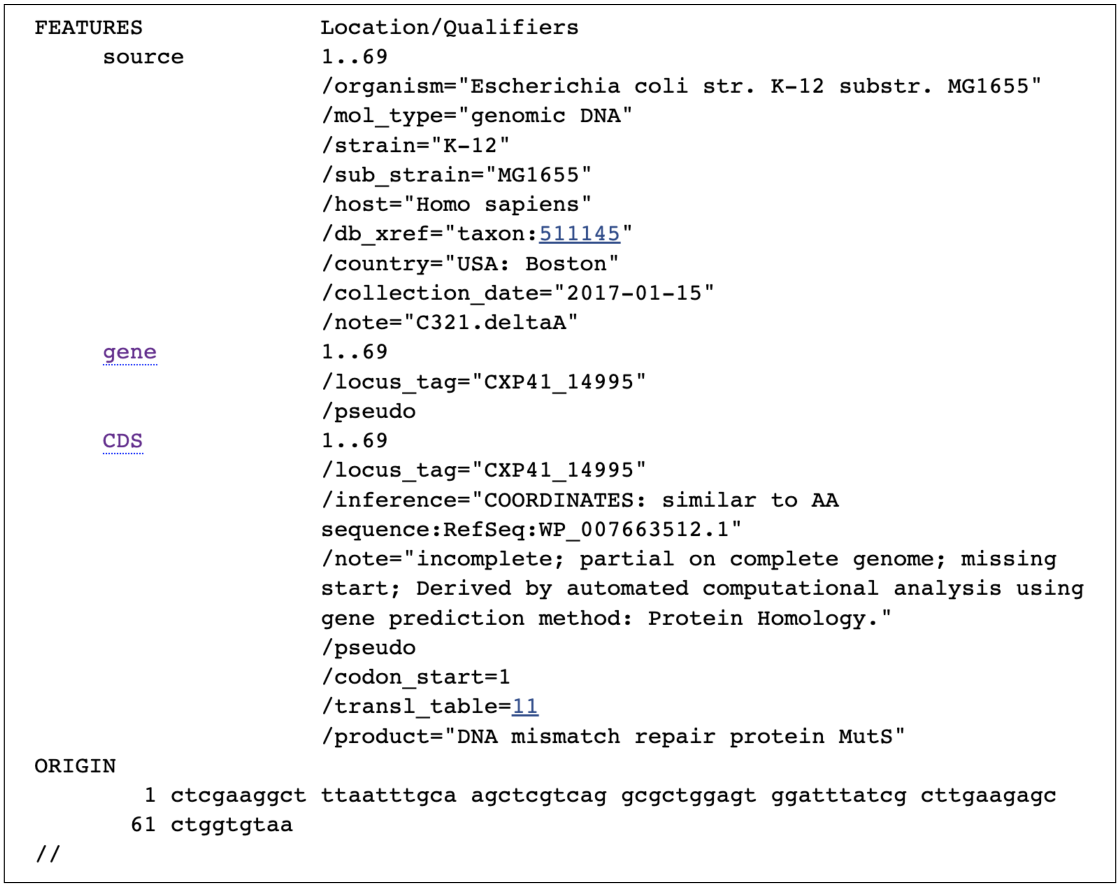
2.3 Results

2.3.1 No Evidence of *mutS* Sequences Described by LeClerc et al., 1996 as Hypermutators

I found a *mutS* sequence in 816 of the 817 isolates in this data set, with the exception of isolate GCF_002843685. Sequences uploaded to the NCBI are automatically annotated using the Prokaryotic Genome Annotation Pipeline (PGAP) [119, 120, 121], so I compared the online NCBI record for GCF_002846385 to the Prokka and Roary results. Upon investigation of GCF_002843685 on the GenBank website, the annotation of the MutS protein derives from the relative location of that fragment of DNA [122]. The nucleic sequence data is incomplete (see Figure 2.1, resulting in a DNA fragment too short for recognition by the program parameters I passed to Prokka and Roary. The fragment was also much too short to match with the criteria noted by LeClerc (significant large deletion of 212 nucleotide base pairs at the 3' end of the protein) [37].

The possibility existed for the GCF_001485455 isolate to represent a substantial deletion of *mutS*, potentially similar to LeClerc's definition of a defective MutS protein. I performed a Chi-Square Goodness of Fit Test to determine if the observed incidence of hypermutators in my data set matched the expected hypermutator prevalence of 1% noted in literature by LeClerc et al., 1996 [37] and others [79]. The data matches the assumptions (counts, exclusivity, independence) of the Chi-Square Test [112], including the expected count of 8 hypermutator *mutS* (1% * 817 = 8.17). I calculated the results of the test as a $\chi^2 = 6.19$, df = 1, p = 0.013. The result is significant at $p < .05$ with 1 degree of freedom, leading me to reject the null hypothesis of similar distributions. Instead, my observations of potential *mutS* hypermutator phenotypes are not drawn from the expected distribution of 1% hypermutator incidence.

The missing *mutS* annotation for GCF_002846385 could result from a combination



```
FEATURES             Location/Qualifiers
     source          1..69
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /host="Homo sapiens"
                     /db_xref="taxon:511145"
                     /country="USA: Boston"
                     /collection_date="2017-01-15"
                     /note="C321.deltaA"
     gene            1..69
                     /locus_tag="CXP41_14995"
                     /pseudo
     CDS             1..69
                     /locus_tag="CXP41_14995"
                     /inference="COORDINATES: similar to AA
                     sequence:RefSeq:WP_007663512.1"
                     /note="incomplete; partial on complete genome; missing
                     start; Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /pseudo
                     /codon_start=1
                     /transl_table=11
                     /product="DNA mismatch repair protein MutS"
ORIGIN
        1 ctcgaaggct ttaatttgca agctcgtcag gcgctggagt ggatttatcg cttgaagagc
       61 ctggtgtaa
//
```

Figure 2.1: Screen capture (Escherichia coli str. K-12 substr. MG...) of NCBI Assembly record annotations for MutS protein in RefSeq isolate GCF_002846385 (GenBank: CP025268.1; Assembly: GCA_002846385; BioProject: PRJNA421841).

of many upstream potential error sources: collection methodology, sample contamination, preparation techniques, or sequencing issues. All other isolates contained a *mutS* sequence of roughly the correct length (8̃53 AA). Therefore to maintain confidence in my analysis, I pruned the GCF_002846385 isolate from the phylogenetic tree and excluded it from the binary MDR character matrix when mapping.

Another isolate, GCF_001485455, showed a large number of deletions, but I observed that all the deletions occurred only in the previously noted 141 variant AA sites. As such, I observed no evidence of the specific 221 bp deletion from the 3' end of *mutS* described by LeClerc et al., 1996 [37]. I identified no long contiguous deletions, as the remaining 816 isolates with annotated *mutS* sequences contained

Table 2.2: Chi-square test for goodness of fit expressing the observed (Obs.), expected (Exp.), Difference (Diff.), square of difference (Diff. Sq.), and square of difference divided by the expected fraction (Diff. Sq. / Exp. Fr.) values rounded to two decimal places for hypermutators and not hypermutators in this population of *E. coli.*

	Obs.	**Exp.**	**Diff.**	**Diff. Sq.**	**Diff. Sq. / Exp. Fr.**
Hyper.	1	8	-7.00	49.00	6.12
Not Hyper.	816	809	7.00	49.00	0.06
Sum	817	817			6.19

relatively few gaps when compared to overall length. I concluded that none of the *mutS* genes in this data set are hypermutators as defined by LeClerc et al., 1996 [37], hypermutator phenotype *E. coli* due to defective *mutS* genes caused by large deletions.

2.3.2 Low Variation in Mutator S Component Amino Acids

As the hypermutator trait could not be represented by a deficient *mutS* binary state, I made a choice to reconsider the hypermutation phenotype in light of MutS protein sequence variants as opposed to the complete deficiency definition used by LeClerc et al., 1996 [37]. The MutS AA sequences only varied at 141 of the 853 possible locations, with the most common sites of variance found at positions seen in Figure 2.2. Figure 2.3 shows the number of amino acid polymorphisms per MutS variant. Figure 2.4 and Figure 2.5 deomonstrate that MutS variants occur with low prevalence, matching the expectation of a highly conserved sequence given the important role of MutS in DNA mismatch repair. I found no deficient MutS sequences matching LeClerc's definition, nor did I identify any missing *mutS* annotations that I could not attribute to potential sequencing error. All annotated MutS proteins displayed little variation in terms of the protein product size, the location of variant residues, or the prevalence of variant residues. Therefore, I chose to test each of the 141 variant AA sites of the MutS protein for correlation with the MDR binary trait.

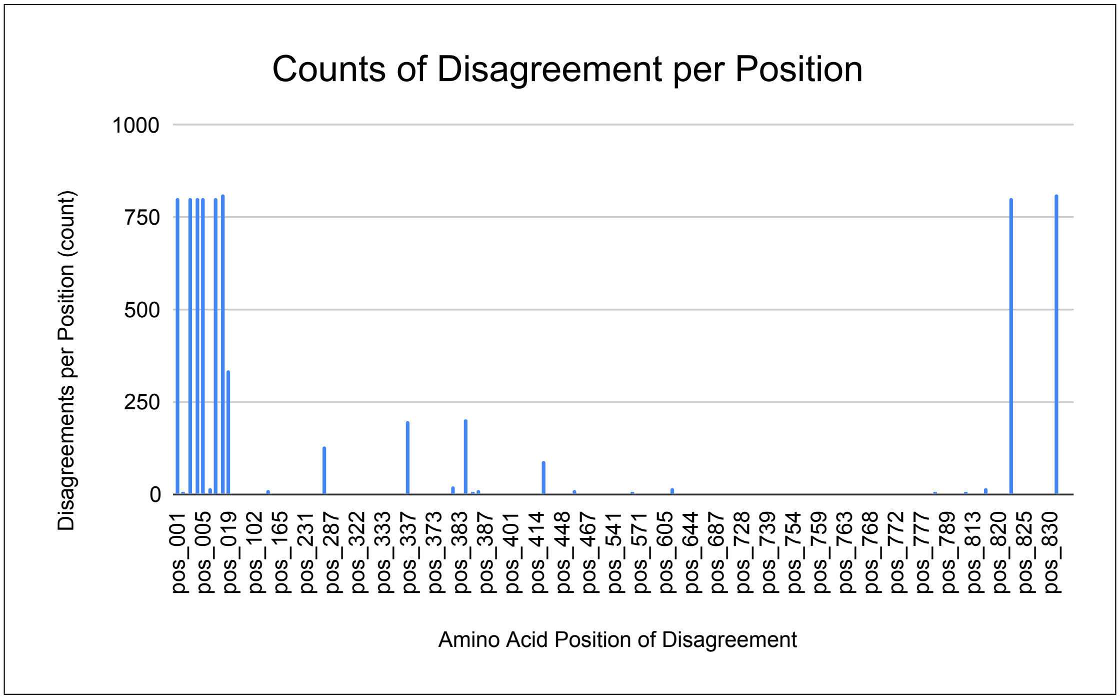


Figure 2.2: Displays the counts of disagreement from the consensus sequence residue at each amino acid position (pos_###) in the MutS protein.

2.3.3 MDR Correlates with Only a Few Variant Amino Acid Positions

I implemented Pagel's 1994 [116] algorithm using the `fitPagel` function available from the package phytools version 1.2 [117] written in the R language version 4.2.2 [118]. With `fitPagel`, I calculated the phylogenetic correlation between each of the 141 variant MutS positions and the MDR trait according to four hypotheses: independence, X dependent on Y, Y dependent on X, and interdependence. I ran 141 iterations of the four hypothesis tests, assigning X to for the variant position and Y always reserved solely for MDR. From the weighted Akaike information criterion (AIC) values calculated for all 141 variant positions, the `fitPagel` results only support the Y (MDR) dependent on X (variant positions) model at 3 positions (see Table 2.3). At all other positions, the independent model of character evolution is supported (see Supplementary Data in the Appendices).

I observed the first location at position 7, found in domain I (2-115) [123]. MutS

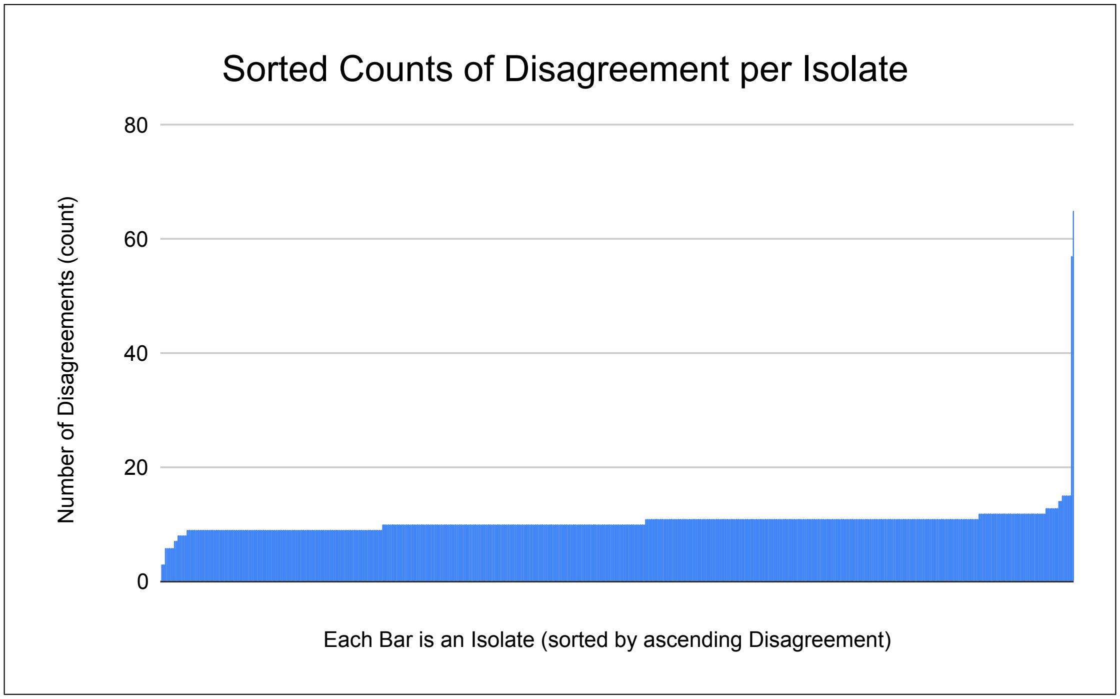


Figure 2.3: Sorted bar chart (from least to greatest) of counts of disagreement from the consensus MutS protein sequence for each *E. coli* isolate.

domain I, also referred to as the mismatch-binding domain, is globular in structure and is responsible for DNA binding interactions [123, 124]. The weight of the dependent model at position 7 is 0.69. Consensus from these isolates at position 7 indicates a phenylalanine residue, with variants of either leucine or missing data.

The second is found at position 140, with consensus at 140 for serine and variants composed of cysteine or missing data. Residue 140 is part of domain II (116-266), or the connector domain, an internal, mixed region comprised of alpha helices and beta sheets, linking the DNA binding domain I with the MutS core domain III (267-443 and 504-567) [123, 124]. As calculated by `fitPagel`, the weighted AIC of the dependent model at position 140 is 0.73.

Finally, the third location occurs at position 804, adjacent to the C-terminal side of the helix-turn-helix (HTH) motif (from 766-800) where the MutS dimers interact [123, 124]. The weight of the dependent model at position 804 is 0.46, with consensus for residues of alanine, varying with aspartic acid, threonine or incomplete data. The

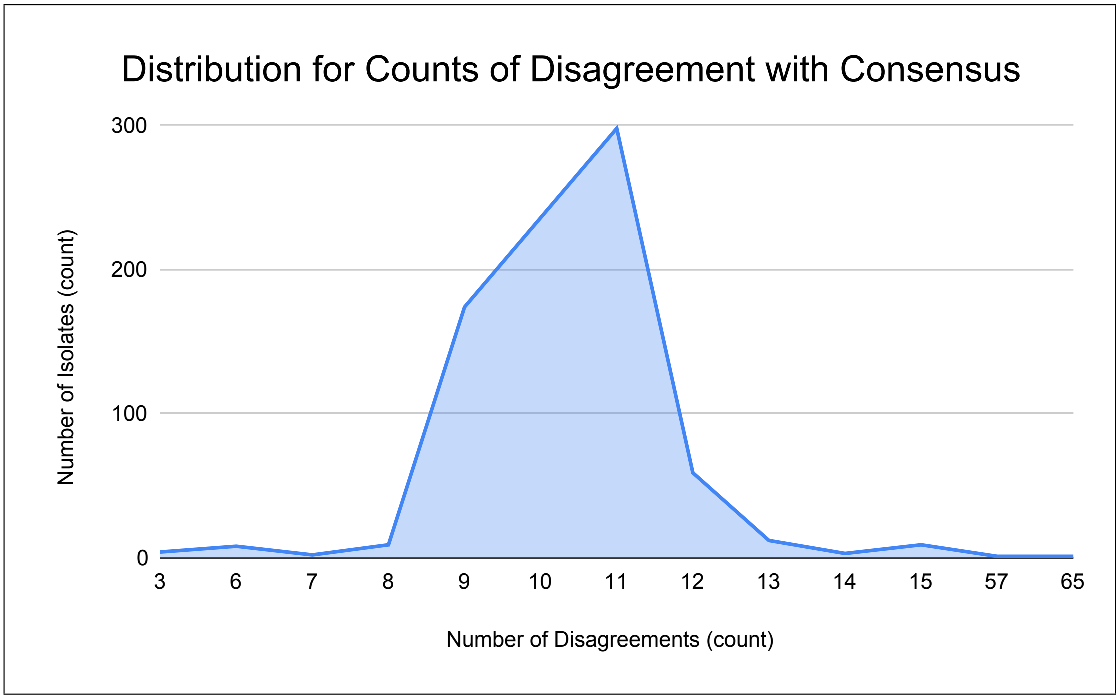


Figure 2.4: Distribution of counted disagreements from the MutS protein consensus sequence for each isolate.

support for the dependent model at the third location is noticeably weaker than at the previous two positions (see Table 2.3).

Interestingly, these locations only disagreed with the consensus sequence 16, 11, and 8 times respectively (compare against Figures 2.2, 2.3, 2.4, and 2.5). The median value for disagreements per position is 10 with a mean of 10.4 and a standard deviation of 2.8. Position 7 lies within two standard deviations of the mean, while positions 140 and 804 are within one standard deviation. Of the 3 positions correlated with binary MDR, none differed significantly from the mean for total disagreements.

Table 2.3: Selected residues (Res.) with Independent (Ind.) and Dependent (Dep.) model AIC values, both raw and weighted (Wt.), as calculated by `fitPagel`. Values represent greatest weighted AIC support for the "Dependent Y" model of evolution, where the MutS residue position depends on the MDR binary trait.

Res.	**Ind. AIC (raw)**	**Ind. AIC (%)**	**Dep. X AIC (raw)**	**Dep. X AIC (%)**	**Dep. Y AIC (raw)**	**Dep. Y AIC (%)**	**Dep. Both AIC (raw)**	**Dep. Both AIC (%)**	**Wt. AIC (sum %)**
7	983.221	0.153	998.104	0.000	980.193	0.695	983.233	0.152	1.000
140	1048.579	0.000	913.057	0.165	910.078	0.732	913.992	0.103	1.000
804	957.586	0.324	959.452	0.128	956.863	0.465	960.312	0.083	1.000

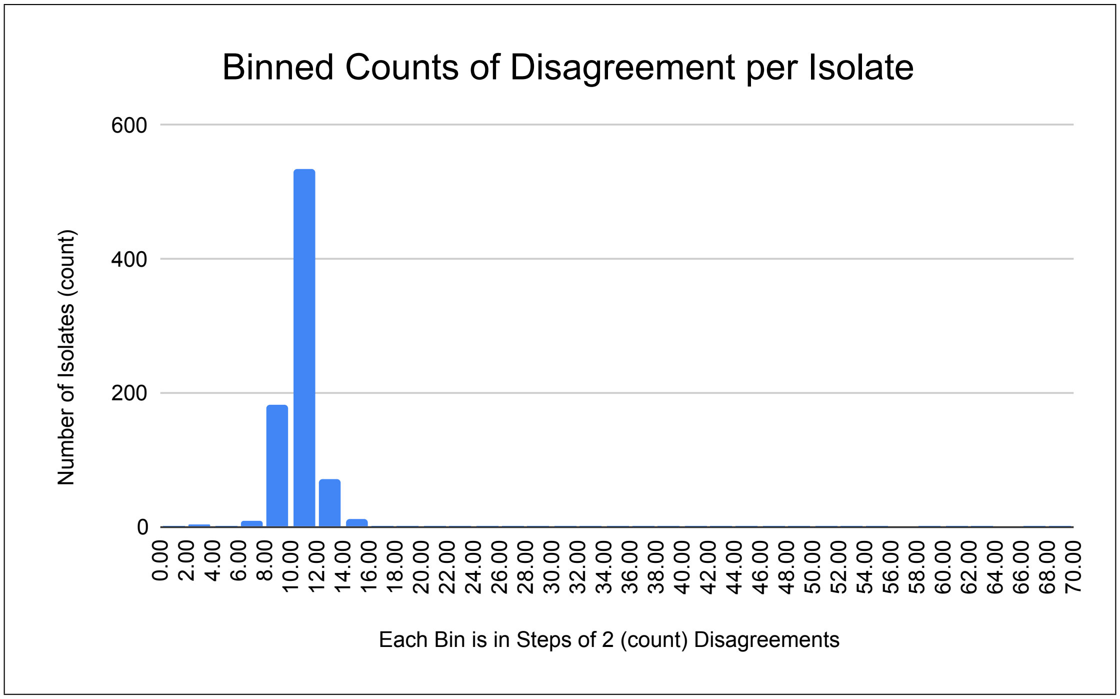


Figure 2.5: Histogram representing number of isolates in binned (2) counts of disagreement from the consensus sequence MutS residue.

2.4 Discussion

2.4.1 MDR Does not Require a LeClercian Hypermutator

I found no truncated *mutS* in this dataset, however I observed many examples of MDR (271 isolates). As I found no evidence for the existence of a hypermutator as defined by LeClerc et al., 1996 [37], therefore, the LeClercian hypermutator phenotype *mutS* could not be associated with MDR. This high quality, high confidence, high completeness *E. coli* data set consists of hundreds of isolates drawn from laboratory, health care, and environmental settings. These isolates represent *E. coli* exposed to a broader set of real world conditions than those bacteria kept in a laboratory.

As investigative techniques improved over the past 25 years, it is possible that LeClerc was constrained by the tools of the time. LeClerc's identification of the 221 bp deletion could be due to experimental methodology. They identified the deletion with long polymerase chain reaction (PCR) analysis and sequencing in only a single one of the four (4) *E. coli* mutator strains [37]. The long PCR analysis failed to amplify any *mutS* products in the other mutators [37]. To investigate the three (3) remaining mutator strains, the investigators opted to employ low-stringency oligonucleotide probes. “Because probe results that were positive with nonmutator O157:H7 and O55:H7 strains were negative with these [the remaining three] *E. coli* mutators, we conclude that they carry extensive deletions, affecting not only *mutS* but also surrounding genes.” [37]. This assumption and slight change in methodology might have introduced errors into the analysis. Failed amplification of PCR products in three of four attempts could result from any number of scenarios, including sequencing error, primer failure, temperature controls, buffer inconsistencies, etc.

Tool choice and methodological technique influence experimental results. For my investigation, I specifically set high thresholds for annotation confidence, sequence completeness, sample contamination, and data quality. I used these parameters during my annotation process, while verifying discrepancies against the NCBI PGAP process. These thresholds identified *mutS* in all isolates with the single exception of GCF_002843685 (see Figure 2.1). I found no evidence for significant deletions to *mutS* that could not result from sequencing error. While *E. coli* hypermutation phenotypes are well-documented experimentally and in literature [60, 125, 37, 126, 127], I could not observe LeClerc’s definition of a deficient *mutS* gene. As I could not verify the presence of deficient *mutS* in my data, I have to rule out LeClerc’s definition of hypermutation when determining the causes of the MDR phenotypes that do exist in these isolates. Without any observed significant deletions in *mutS* matching LeClerc’s definitions and given the significant, high-confidence results of the chi-square goodness of fit test, I was unable to observe the phenomena described by LeClerc and therefore cannot reject my null hypothesis: there is no correlation between significant *mutS* deletions and MDR phenotypes.

2.4.2 Limited Variation in MutS Protein Does not Support LeClercian Hypermutation

I compensated for the absence of deleterious *mutS* sequences by investigating all observed MutS protein variants. The 816 MutS annotated in my dataset were highly conserved. I found only 141 variant amino acid sites in the 853 residue protein; 712 positions or 83.5% of the total residues in the observed MutS protein are invariant (see Figure 2.2). The variation in disagreement from the consensus sequence is normally distributed (see Figure 2.4). Most isolates also differed from the consensus sequence by at most roughly 11 residues from the possible 141 variant positions (see Figures 2.3, and 2.5), meaning that across the entire length of the 853 sites, the observed MutS differed from each other in 1.3% of total residues. This lack of variation is further seen by the outcome of the models calculated by `fitPagel` [117], where evolution of MDR depends on three positions in MutS.

For those three identified positions, 7, 140, and 804, I found no literature suggesting that those residues are known to be associated with hypermutation. A literature search noted substitution mutations at Val352Ile and Met628Leu [86] in *E. coli* MutS or Val246Ala and Val421Phe [128] in *Salmonella* MutS (corresponding to *E. coli* MutS Val244 and Val419) have been linked with hypermutation, although the authors of those studies employed much broader definitions of hypermutation than the strict sense used in this dissertation. Of those substitution mutations, I only observed Val244Ala in my data, but the `fitPagel` results did not support a dependent evolutionary relationship between position 244 and MDR. The AA at positions 7, 140, and 804 in *E. coli* MutS are novel findings that link MutS and MDR, but not under LeClerc's definition of hypermutation as explored in this investigation. Instead, these residues could indicate locations within protein MutS responsible for weak mutator activity [129, 85, 87, 130].

In combination, the limited variation in MutS and the largely independent evolution between the binary characters of MDR and MutS variants suggest that the structurally derived hypermutation put forward by LeClerc does not exist in these data. By redefining the MutS character from the annotated gene space to the protein space, I found little evidence that supports rejecting the null hypothesis of no correlation. Only three specific variant MutS residues correlate with the MDR phenotype, so in only those cases, could I reject the null hypothesis. Instead, the models support a largely independent evolutionary relationship between the characters for deficient mismatch repair and MDR.

2.4.3 A New Question - What Predicts MDR Phenotype Bacteria?

In light of my research, I find it unlikely that the specific hypermutable phenotype *E. coli* defined by LeClerc (as isolates whose MutS protein contains large deletions) exists in the global survey data set I collected from GenomeTrakr. I did not observe a repair deficient hypermutable phenotype bacterium (defined as large deletions within *mutS* by LeClerc et al., 1996 [37]) in my data, so any elevated mutation rates in these samples cannot be the result of deficient MutS. Some confusion in research and literature results from combining these two separate definitions for hypermutation: 1) LeClerc's deficient MutS, and 2) an observed increase in mutation rates. For example, some researchers demonstrate that increased mutation rates result from inducible, reversible phenotypes [79, 84, 131, 132, 133, 134], often referring to the phenomenon as hypermutation. While a redefinition of the term, this inducible hypermutator phenotype with increased rates of mutation is likely due in part to the increased fitness burden carried by repair deficient mutator strains [80, 135, 136, 137, 138], as an induced phenotype only experiences a temporarily reduced fitness.

Hypermutation in the broader, increased mutation sense drives the development of novel mutations that impart AMR phenotypes, especially given the increasing evidence for induced, transient hypermutable states [79, 84, 131, 132, 133, 134]. This inducible, transient phenotype was also discussed by LeClerc, "... the ultimate pathogen would possess an elevated mutation rate that is transient (or conditional), providing genetic variation during the first few hours when the pathogen must survive, invade, and colonize its host." [37]. While this conjecture is noteworthy, LeClerc's published definition of hypermutation was structural, explicitly focused on a significant deletion resulting in deficient MutS, which I did not observe in my data.

In addition to the lack of an observed, truncated *mutS* to confer a persistent hypermutable phenotype, my models supported independent evolution between amino acid variant positions in MutS and the MDR phenotype. This evidence suggests that there is little to no correlation between the *mutS* gene and the MDR phenotype. Only three positions (7, 140, 804) indicated support for a dependent model between those residues and MDR. These three locations may indicate a MutS amino acid phenotype that more readily enters a transient hypermutable state in the broader sense, or those variant sites could suggest a genetic lineage that colonizes environments often exposed to antimicrobial compounds. Future work could ascertain the evolutionary lineage of those three sites and quantify the functional impact of amino acid base substitutions at those locations.

The investigation of this chapter set out to verify the work done by LeClerc et al., 1996. In 816 *E. coli* isolates, I could not identify a *mutS* sequence matching the hypermutator defined by LeClerc. In addition, when I chose to more flexibly define the hypermutator phenotype from significant *mutS* deletion to variable MutS AA positions, I found strong support for independent models of binary character evolution. I was unable to verify LeClerc's claims, and therefore failed to reject the null hypothesis of independent evolution between *mutS* and MDR. The phenomenon of hypermutable phenotypes is best considered separate from the growing problem of MDR. The results of my inquiry lead to the open question: **what then are the genetic predictors of MDR phenotype *E. coli*?**, which I explore in a subsequent chapter.

CHAPTER 3: CORRELATIONAL ASSESSMENT OF *E. COLI* PANGENOME FOR PREDICTING MDR PHENOTYPES

3.1 Introduction

Within academia and industry, an ongoing push exists for the widespread adoption of machine learning and artificially intelligent systems [139, 140, 141, 142, 143]. Machine learning in this context is broadly defined as the ability of computers to learn information without explicit programming [144, 145]. The use cases for machine learning involve problems without explicit solutions, that require large rule sets, or that routinely encounter novel data [144]. Such problems are becoming commonplace for computational biologists.

3.1.1 Supervised Machine Learning

Supervised machine learning, a subtype of artificial intelligence, approximates solutions to intractable problems with an associated knowledge base [146]. For example a potential problem could be: what combination of genetic markers and health history data predispose an individual to developing cancer? Supervised machine learning trains algorithms to predict the outputs for a given set of inputs [147]. When those outputs represent exclusive label values, the outputs are referred to as classes and the involved algorithms are called classifiers [147]. These classifiers learn a defined set of inputs from a training data set to generate useful output labels [148, 146, 147]. The two inputs required for training supervised machine learning classifiers take the form of matrices: a feature matrix and a label matrix.

For the feature matrix, each row is a sample, and each column denotes a feature. Features represent specific measurable properties. The simplest label matrix contains the same number of rows (samples) as the feature matrix, but only a single binary column. This column notes the labeled class value for a given sample. Every sample in the full data set has both measured features and a labeled class. Sampling of both the feature matrix and the label matrix separates the initial data into training and testing data sets. During training, the classifier associates certain sets of features or specific feature values with known labels. Those features or values which most influence the model towards correct labeling are the best predictors.

Once trained, the researcher assesses the classifier using the testing data subset to determine accuracy, precision, and recall metrics [148, 149, 150, 146, 147]. The researcher supplies the trained classifier with the test feature matrix and then tasks the estimator to predict the correct labels for that data. The researcher compares the predicted labels against the ground truth of the test label matrix. This iterative process of training and testing, allows further tuning of the model. Following best practices, I tested and tuned the model in order to make reproducible predictions with statistical significance, based on several calculated metrics [149, 150, 151].

3.1.2 Supervised Machine Learning in Computational Biology

Supervised machine learning finds use within computational biology to predict the shape of proteins, locate protein domains, anticipate molecular interactions, and surveil food safety data [148, 152, 146, 147]. The label matrix in these examples is often drug interaction information, protein characteristics, associated traits, or labeled phenotypes. The feature matrices for biological data are generally genetic, often containing nucleic acid, protein, and gene presence or absence information. At their core, machine learning analyses are applied statistical measures of correlation.

3.1.3 Research Question and Hypothesis

Recent research shows promise for the application of these correlative machine learning techniques to predict food, zoonotic source, location, and certain aspects of AMR from WGS data [152, 153]. These papers are contemporary with the US FDA's call for more research into improving predictive models based on WGS data [57, 8]. Machine learning approaches are of special interest to the FDA for improving genotype-to-phenotype hypotheses, clarifying outbreak investigations, and identifying nonsynonymous changes that allow a given pathogenic lineage to contaminate food [8].

Similar concerns exist for the silent pandemic of AMR proliferation which too requires improved hypotheses, outbreak tracking, and identification of pathogenic mutations. With such similarities, the machine learning methods may also inform our understanding of the progression from individual AMR genes to MDR phenotypes. The previous inquiry (Chapter 2) resulted in the rejection of deletion of the MutS gene as correlated with an MDR phenotype, leading to the open question: **what are the genetic predictors of an MDR phenotype *E. coli*?**

To better understand this open question, I selected data from the NCBI PD repositories. I conceptualized the genetic sequence data and character matrices common to phylogenetic analyses as analogous to the feature and label matrices required for supervised machine learning. Given the uncertain evolutionary history of the bacterial MDR phenotype, its polygenic cause, and the highly variable *E. coli* pangenome exclusive of the core genes, this investigation aims to test an alternate hypothesis for the MDR phenotype, that genetic patterns other than *mutS* better explain the development of MDR. In keeping with the previous chapter's focus on hypermutation and the potential for paralogous variation in *mutS* to affect MDR, the null hypothesis of this current investigation asserts that a random forest classifier trained on annotated *E. coli* gene features will reveal *mutS* or its paralog as the best predictor of MDR.

3.2 Materials and Methods

3.2.1 Materials

I used the same raw *E. coli* food safety data set analyzed in Chapter 2 and described in Chapter 1 here. That data subset draws from our earlier study involving 29,255 isolates of Escherichia coli and Shigella accessed from the NCBI PDIB on October 08, 2018 by Ford et al., 2020 [5]. I again chose the following selection criteria:

- Sequences of Escherichia coli identified by NCBI:txid562.
- Sequences from the two highest levels of assembly (designated as Complete Genome and Chromosome) .
- Sequences have a RefSeq identifier (schema: GCF_#########).

I downloaded available (875) whole genome sequences for the 911 isolates identified by the criteria from the NCBI FTP site via their associated accession identifiers. I took the antimicrobial resistance properties of these isolates from the NCBI PDIB table we downloaded in Ford et al., 2020 [5]. I quality controlled the 875 successfully downloaded isolates with CheckM [101, 102, 103, 104], to arrive at 817 remaining *E. coli* WGS.

As described in Chapter 2, I transformed the data to arrive at the input feature and label matrices, by engineering the annotated WGS of the isolates and their AMR phenotypes respectively. I defined the feature matrix with annotations assigned via Prokka version 1.12 [105]. I employed Roary version 3.12.0 [106] to develop the pangenome of the 817 sequences to prepare input features. Notably, for the methods in this third chapter I changed the parameter flags I input to Roary. In Chapter 2, I simplified the core genome prior to phylogenetic analysis by instructing Roary to keep paralogous genes together under a single annotation. The supervised machine learning analysis in this third chapter benefits from an increased number of features by splitting paralogous genes, in an attempt to allow some phylogenetically derived information about *mutS* paralogs to influence the random forest classifiers. Chapter 2 describes exact methods for arriving at the binary MDR phenotypes. I extended those methods, as noted below, to create the appropriate label matrices, representing the binary MDR phylogenetic character as predicted AMR labels for supervised machine learning. I manually curated the AMR categories for the label matrix, based on definitions provided by the CDC NARMS [1] and cross-referenced through CARD [110].

3.2.2 Methods

3.2.2.1 Random Forest Classification

To model the supervised learning, multi label classification problem posed by my hypothesis, I decided on the random forest classifier [149]. Zhang et al., 2019 [154] used random forests trained on *Salmonella* data to successfully predict host species using WGS data. Deng et al., 2021 [144] suggests the usefulness of random forests for food safety applications. A random forest is an ensemble model that builds multiple decision tree classifiers [155]. During the training stage, the decision pathways from each decision tree are averaged and grouped to estimate the best overall prediction for the input data [144, 155]. The resulting prediction is indicated by the largest number of individual decision trees [144, 155].

Decision trees are easily interpreted, as only a single decision pathway exists from the root node to the final leaf. Random forests represent complex meta estimators, spreading the decision pathway across multiple decision trees [144, 155]. Despite this added complexity in interpretation, the random forest has several advantages over the decision tree. Random forests improve upon the tendency of decision trees to overfit the data, allowing random forests to outperform decision trees [149, 155], while also scaling well and handling high-dimensionality data [144, 155]. Given the large size of my data set, the potential for variation in the number of data input features, and the importance of not overfitting, the random forest estimator seems to best fit these constraints.

To implement my random forest classifier, I used Python 3, versions 3.9 and 3.10, [99] from Anaconda version 4.11.0 [100] and the scikit-learn version 1.1.2 package [151, 156] for machine learning and data processing by calling the function `sklearn.ensemble.RandomForestClassifier` [157]. The code is available upon request. I tuned the number of decision trees, machine learning bootstraps, sampled data subsets, and other parameters according to the documentation [156]). Random forest classifiers require two input matrices, one representing the desired features (predictor values, derived in this case from annotated genetic sequence information) and the second representing the predicted labels (in this case phenotypes, such as resistance to aminoglycoside and tetracycline drugs). Detailed information on hardware specifications, Anaconda environments, and supporting programs may be found in the Appendices.

3.2.2.2 Feature Matrix

Other research indicated the usefulness of reference-free comparisons for machine learning [158], so I took a similar approach. For this second inquiry, I chose the *gene_presence_absence.csv* file generated by Roary and transformed it to a binary matrix representation of the annotated genes. Each row in the matrix represents an isolate, while each column notes the binary presence or absence of a given annotation for each isolate. I generated Roary visualizations with suggested scripts provided by Marco Galardini [159].

The Roary analysis process does not map sequences to a reference genome, but instead aligns genetic sequences of identical annotation [106]. This process effectively removes sequence context from the annotations, since I made the choice not to split paralogs in Chapter 2. I was not concerned with this potential loss of information for the phylogenetic investigation, and I made concessions for the sake of processing power and computational time when developing the RAxML tree of 817 highly similar *E. coli* isolates with 2916 shared, core genes. For the purposes of this machine learning inquiry however, I made the choice to allow Roary to split paralogs in order to recover context dependent information (such as control sequences, accessory genes, or *mutS* variation) that affects the development of MDR.

Additionally, I shifted focus from the core genome to genes with lower observed prevalence in the data. Roary subdivides the pangenome into four sections based on prevalence thresholds: core (99 - 100% prevalence inclusive both), soft-core (95 - 99% inclusive left), shell (15 - 95% inclusive left), and cloud (0 - 15% inclusive left) (Page et al., 2015), as shown in Figure 3.1. A binary input matrix derived from the core genome is nearly uniform, containing a large majority of cells with a value of 1. In a similar way, low prevalence data (see Figure 3.2 leads to a sparse matrix with mostly values of 0. Both of these qualities prove challenging for machine learning analyses, by causing the classifier to overfit the training data. To circumvent these challenges, I took the input feature matrix from the central portion of annotations, the majority of which encompasses the shell and a small portion of the region considered cloud. I excluded any annotations of greater than 95% or less than 10% prevalence from the machine learning input. These operations resulted in the creation of an 817 x 5700 binary matrix. Each row is an isolate, and each column represents an annotation assigned by Prokka and Roary.

Collinearity, commonly found when considering genetic data, poses another challenge for machine learning. Groups of annotations in my 817 x 5700 data matrix vary linearly with each other as seen in Figure 3.3. To address the challenge of collinearity, I calculated Spearman's rank order correlation coefficient for the matrix to perform hierarchical clustering [160] using Ward's method [161] as implemented in the Python 3 library SciPy version 1.9 [162]. Common practice selects a number of new features from the hierarchical clusters, where both the number of clusters and the number of members per cluster are sufficiently large [160]. Based on exploration of the data see Figure, I chose a value of 66 clusters shown in Figure 3.4 I chose the relative intersection of the two curves (at 66 clusters) to maximize both inter- and intra-cluster information demonstrated in Figure 3.5. This transformation resulted in new features, each representing a cluster of collinear annotations. The final input feature matrix is 817 x 66. Each row is an isolate, while each column is a feature representing a cluster of hierarchically grouped annotations. I performed all transformations in such a way as to maximize the explainability of any identified best predictors for each classifier.

3.2.2.3 Label Matrices

I created the label matrix from a modified version of the MDR binary character from Chapter 2, basing the MDR binary character on the *AMR_genotype* column downloaded in the earlier study, Ford et al., 2020 [5]. However, I made the decision to modify the MDR binary character matrix from Chapter 2 in preparation for supervised machine learning due to two main concerns. First, many varieties of MDR phenotypes occur, each composed of different combinations of AMR genes arising in different lineages in various ways over time [163]. Second, individual AMR genes behave under genetic capitalism, allowing lineages to arrive at stable AMR genotypes by gaining more AMR genes than they lose [5]. Both phenomena complicate the question posed in this chapter, especially due to the first concern noted above. This specific issue of uncertainty in AMR gene heredity causes significant disagreement between proponents of hypermutation and HGT as to how a specific gene arrived in a given lineage. To improve upon the ability of my models to demonstrate what are the best predictors of MDR, I built upon our work in Ford et al., 2020 [5] and the investigation of the previous chapter by creating labels more granular than binary MDR, but less granular than individual genes. (See Figure 3.6.)

To account for these realities, this inquiry treats each category of drug resistance separately. With this strategy, the initial multilabel classification problem (i.e., how to identify, rate, and explain predictors of MDR phenotypes consisting of multiple, non-exclusive groupings of AMR genes?) is best formalized as a combination of several binary classifiers. I selected five categories of drug resistance recognized by CDC NARMS [23] and treated each one as a separate binary classification problem. This approach improved the explainability of each individual classifier, as each model identifies only a single type of resistance. I compared these individual models to identify the genetic predictors of different types of MDR. Overlap between predictors indicates that an isolate contains the necessary genetic hardware to build towards multiple types of AMR, and eventually an MDR phenotype. I selected this approach in order to better understand if multiple genetic (and therefore evolutionary) paths towards MDR exist.

I designed the five separate binary classification problems to predict genes that confer resistance to **aminoglycosides**, **folate pathway antagonists**, **macrolides**, **tetracyclines**, and **others**. I reserved the category of **others** for genes not found to confer resistance to aminoglycoside, beta-lactam combination agent, cephem, folate pathway antagonist, macrolide, nucleoside, penicillin, phenicol, quinolone, or tetracycline-based compounds (see Table 2.1). I chose not to train classifiers for the following resistance categories: cephems, nucleosides, penicillins, phenicols, and quinolones. Due to their low attributed prevalence (each excluded category is found in less than 17 of the 817 isolates or $< 2.1\%$ of isolates), I excluded these categories to prevent issues of overfitting when training on small sample sizes for the positive class, a known problem in supervised machine learning [164, 165].

Notably absent from the list of five is the beta-lactam combination agent category. Data exploration revealed the presence of at least one gene imparting beta-lactamase drug resistance in nearly 100% of the isolates. As noted above, a uniform matrix of nearly all 1 values poses a challenge for machine learning algorithms, so I chose to exclude the beta-lactam category. Additionally, the near universal prevalence of beta-lactam resistance would modify the definition of MDR from >=3 classes of resistance to >=2 classes of resistance, exclusive of beta-lactam resistance. I avoided this modification of terms, by choosing only the above five categories of resistance. The input label matrix for each of the five classification problems is a 817 x 1 matrix, where each row is an isolate and the column represents the binary presence (1) or absence (0) of resistance to that category of antimicrobial drug.

3.2.2.4 Subsetting and Sampling Data Matrices

I split the feature and label matrices into training, testing, and validation data sets using a multilabel stratified shuffle split method [166]. The multilabel stratified shuffle split ensured proportionate representation (without over or under fitting) from all predictor variables while also accounting for the proportions of all five binary label values. I kept the overall data splits consistent using a set seed for the multilabel stratified splits during the tuning, training, validation, and testing processes.

For this inquiry, the training data set refers to 517 isolates (70% of the data) used for training each binary classifier. Validation refers to 123 isolates (15% of the data) used to check the trained classifiers. Testing refers to 123 isolates (15% of the data) held out from the hyperparameter tuning, model training, and model validation processes in order to reserve a blinded sample for final hypothesis testing.

I used the combined training and validation sets (640 isolates or 85% of total data) to tune the hyperparameters of each classifier. When calculating errors during the hyperparameter tuning stage, I only employed the validation set. During model training, I supplied only the training set to the classifiers. To validate the models produced by the training set, I compared the validation set without resampling, by calculating the permutation importance of each input feature as scored by the Area Under the Receiver Operating Characteristic Curve (ROC AUC). Finally, I used the testing data (without resampling) to analyze the trained classifiers as to their predictor effectiveness.

Generally, machine learning classification inputs are referred to as balanced or imbalanced data, with respect to the ratios between the positive and negative classes. In my data, there are more instances of non-MDR (negative) than MDR (positive) isolates, so my data is imbalanced. Base scikit-learn functions work best with balanced data [167, 156], so I looked at the sampling strategies provided by the imbalanced-learn library [167]. I selected a strategy of oversampling the minority class while training the random forest classifiers, by implementing the `RandomOverSampler` function within imbalanced-learn [167]. I chose this oversampling strategy from a head-to-head comparison of SMOTE [168], SMOTEENN [169], SMOTETomek [170], random over sampling [171], and random under sampling methods by the metric of highest ROC AUC score (see Figure 3.7).

3.2.2.5 Determining Feature Importance

I used permutation importance [149, 151] to identify the best predictor values for each trained classifier. To measure permutation importance, the algorithm randomly permuted every input feature to ascertain the difference in predictor score (ROC AUC) from varying that feature [149]. Larger values for permutation importance indicate decreased model score when disturbing a given feature [149]. The greater the permutation importance, the more necessary a feature to the classifier when arriving at a precise and accurate prediction score.

For comparison, I decided to compare the features on both correlation and importance. I calculated chi-square values for each feature to assess correlation, as well as the impurity-based Gini feature importances [150] to approximate feature importance in an alternate manner from the explicit, expensive calculation of permutation importance. In this way, I applied multiple metrics to distinguish between commonly occurring features and features useful to the estimators. As each feature represents a collection of hierarchically clustered, collinear, annotated genes, the resulting permutation importance values indicate groups of annotated genes whose presence or absence best predict resistance to the five classes of antimicrobial drugs.

3.3 Results

To name the clusters of annotated genes, I selected one constituent annotation from each cluster. ***Named clusters*** are shown with bold italics to remain consistent with the representation of **classifier models** (bold) and *gene annotations* (italics) in this manuscript. The component annotations within each cluster remain the same across all five tested AMR resistance classes. As seen in Table 3.1, ***cluster yfjQ_4*** is the best predictor for the **aminoglycosides**, **folate pathway antagonists**, **tetracyclines**, and **others** machine learning classifiers. The ***cluster ydaM*** best predicts resistance to **macrolides**.

Table 3.1: Demonstrates the top five ranked best predictors per classifier: **aminoglycosides** (Amino.), **folate pathway antagonists** (Fol. Path. Antagonists), **macrolides** (Macro.), **tetracyclines** (Tetra.), and **others** (Others). While a few of the classifiers share some best predictors, the rank order of those predictors differs. Only twelve sets of genes are needed to describe the top five ranks for all five resistance classifiers.

0-based Rank	**Amino.**	**Fol. Path. Antagonists**	**Macro.**	**Tetra.**	**Others**
0	yfjQ_4	yfjQ_4	ydaM	yfjQ_4	yfjQ_4
1	cbeA_1	yaiP	ybhC	mhpC	gltJ
2	hybE	group_6154	yfjQ_4	cbeA_1	ybhC
3	group_6154	cbeA_1	yihT	ligB	ydaM
4	yggT	hybE	cbeA_1	gltJ	cbeA_1

Permutation importance demonstrates the relative usefulness of each cluster when predicting each class of resistance. As seen in Figures 3.8, 3.9, 3.10, 3.11, and 3.12, the training (upper) and validation (lower) show relatively similar rankings by permutation importance for the feature clusters in each model. Both changes in rank and the increased variation represented by the whiskers demonstrate the effect of smaller sample sizes on the permutation calculations (517 samples in training and

123 samples in validation sets). Further discussion of results focuses on the training set, due to its larger size and lowered variation. While there are similarities in the top twenty best predictive clusters across the models, Table 3.1 shows how the top five best predictive clusters for each model differ from each another.

Figure 3.13 shows the prediction metrics. At left, observe the precision recall curves, each model colored the same as the respective Figures 3.8, 3.9, 3.10, 3.11, and 3.12. I report the average precisions, the highest of 0.80 for **others**, the lowest of 0.62 for **tetracyclines**. The precision recall curve displays the change in precision between precision (Y axis) and recall (X axis). At right, view the ROC AUC figure, where again, estimator colors are kept consistent across the graphics. The highest AUC is 0.90 for both **macrolides** and **others**, with the lowest as 0.79 for **tetracyclines**.

The bold blue line represents the mean AUC across all classifiers as 0.87 +/- 0.04. This value describes the ability of the five combined binary classifiers to predict a variety of MDR phenotypes with high accuracy. The shaded gray region represents +/- 1 standard deviation from the mean. As seen in Figure 3.13, most of the ROC curves remain within a single standard deviation of the mean, with the major exception of **tetracyclines**.

When comparing the PR and ROC AUC, the best performing model is for predicting the resistance category **others**, as indicated by the greatest AUC and AP values (AUC = 0.90 & AP = 0.80). The next best performing models are **aminoglycosides** (AUC = 0.87 & AP = 0.77), **folate pathway antagonists** (AUC = 0.87 & AP = 0.75), and **macrolides** (AUC = 0.90 & AP = 0.74). **Tetracyclines** stand out as the most poorly performing model (AUC = 0.79 & AP = 0.62). Further evidence is seen in the normalized confusion matrices of Figure 3.7.

Interestingly, when I compared features by their Gini importance and Chi-square correlation values, when sorted by descending mean importance, features with high mean importance did not exhibit a similar pattern in χ^2 values (see Figure 3.14).

The trained models do not rely solely on correlation to achieve accurate predictions. Table 3.2 displays the contents of the highly important feature ***cluster yfjQ_4***.

Table 3.2: Includes all component annotations from Roary found within ***cluster yfjQ_4***, which ranked as most important for the **aminoglycosides**, **folate pathway antagonists**, **tetracyclines**, and **others** models.

Gene (Roary)	**Annotation (Roary)**
group_16177	hypothetical protein
group_10103	hypothetical protein
group_10424	hypothetical protein
group_12760	hypothetical protein
group_7878	hypothetical protein
group_19502	hypothetical protein
group_4098	hypothetical protein
group_24638	hypothetical protein
group_9019	hypothetical protein
group_11161	hypothetical protein
group_5897	hypothetical protein
group_4701	hypothetical protein
group_9337	hypothetical protein
group_17382	hypothetical protein
group_2543	hypothetical protein
group_4818	hypothetical protein
umuC_2	SOS mutagenesis and repair
traM	Relaxosome protein TraM
traY	Relaxosome protein TraY
group_2526	hypothetical protein
ccdB	hypothetical protein
traA	Pilin
yfjQ_4	CP4-57 prophage; predicted protein
group_5903	hypothetical protein

3.4 Discussion

After the results of Chapter 2, I saw little to no correlation between the evolution of characters for *mutS* and the MDR phenotype. To further understand this result, I investigated the question of MDR phenotype development from a supervised machine learning standpoint. The results of this investigation falsify the null hypothesis that a random forest classifier trained solely on the presence or absence of annotated *E.*

coli gene features indicates *mutS* as the best predictor of MDR. Instead, I observed other clusters of annotated genes that best predict multilabel MDR phenotypes with accuracy.

In this analysis, I found no evidence that the presence of *mutS* paralogous genes improves the models' ability to predict MDR phenotypes. The most important feature, ***cluster yfjQ_4***, to the random forest estimator models did not contain *mutS*. I found no *mutS* genes in clusters with high feature importance, by the calculated metrics of chi square score, Gini impurity-based feature importance, or permutation importance. In contrast to the phylogenetic analysis in Chapter 2, this analysis specifically ignored the core genome, leading me to the conclusion that the core genome is less useful than other portions of the pangenome when investigating the predictors of the polygenic MDR phenotype.

Instead, as seen in Figure 3.14, the bars labeled ***ybaQ*** represent the two paralogous *mutS* genes identified by Roary. Both genes occupy the same cluster, with a near 0 chi-square score and less than 0.01 mean importance (Gini impurity). The ***cluster ybaQ*** is not found among the best predictors for any of the trained machine learning classifiers. These results further support the conclusions of Chapter 2: there is no evidence for a correlation between *mutS* and MDR.

In the 1996 paper, LeClerc suggested the theory of hypermutator phenotypes in opposition ("as a counter to the current paradigm") of HGT for the acquisition of resistance genes [37]. Research over the past 30 years has emphasized the role of HGT in the acquisition and proliferation of AMR genes [44, 172, 32, 173, 174, 175, 83], often when considering MDR phenotypes [45, 71, 176]. My results are in contrast with LeClerc, instead supporting the HGT hypothesis. The cluster labeled ***yfjQ_4*** includes many genes associated with HGT (see Table 3.2), most notably the Pilin, TraM, TraY, and prophage protein annotations. This evidence supports the HGT hypothesis for the proliferation of AMR genes into MDR phenotypes.

While both a literature search and this inquiry demonstrate support for the role of HGT in MDR, inducible hypermutation defined in the broader, increased mutation rate sense (occurring with or without *mutS* involvement) may play a role in MDR phenotype development. As noted in Chapter 2, a temporary, induced hypermutator phenotype [79, 84, 131, 132, 133, 134] could reduce the fitness burden of repair deficient hypermutators [80, 135, 136, 137, 138]. In contrast to the understanding of inducible hypermutation, my initial phylogenetic analysis, which relied on the core genome and LeClerc's definition, could not reject the null hypothesis that MDR evolves independently from MutS. In addition, the highly similar sequences of the *E. coli* core genome led to a broad, shallow phylogeny where the MDR character occurred sporadically, never resolving to any isolated clades. To overcome these challenges, I decided to pursue supervised machine learning to investigate genes outside of the core genome, allowing me to compare paralogous sequences for their ability to predict MDR. This investigation revealed that the clusters of annotated genes which best predict MDR phenotypes do not contain *mutS*, but instead HGT. However, while I achieved high precision and accuracy, further work is necessary to examine if the predictor features favored by the models indicate a biological context that improves understanding of MDR. While this investigation and the previous chapter both provide evidence against the role of deficient *mutS* in MDR, the open question still remains: **what are the genetic predictors of an MDR phenotype *E. coli*?**

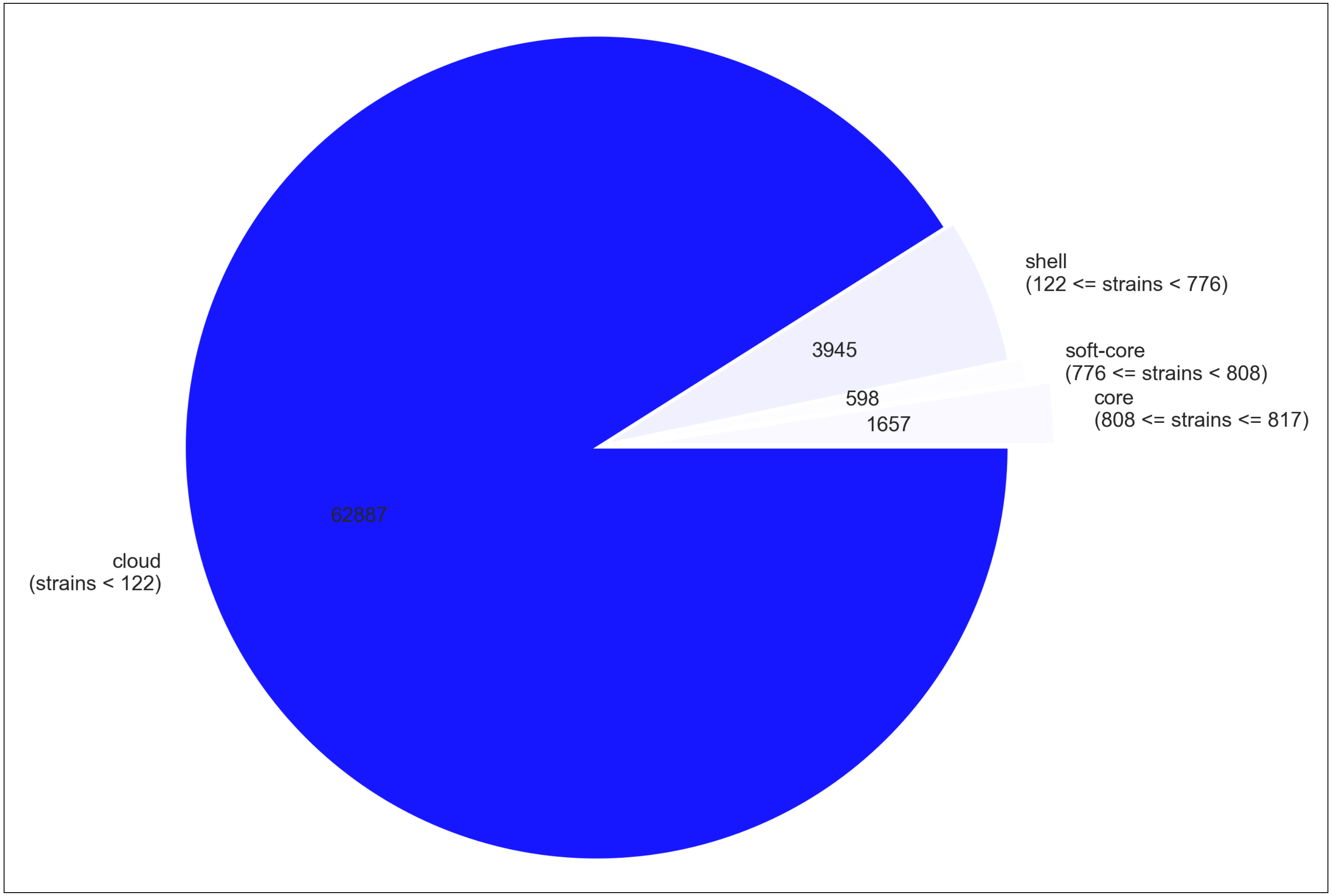


Figure 3.1: Proportions of annotation prevalence in the pangenome calculated by Roary when splitting paralogous genes.

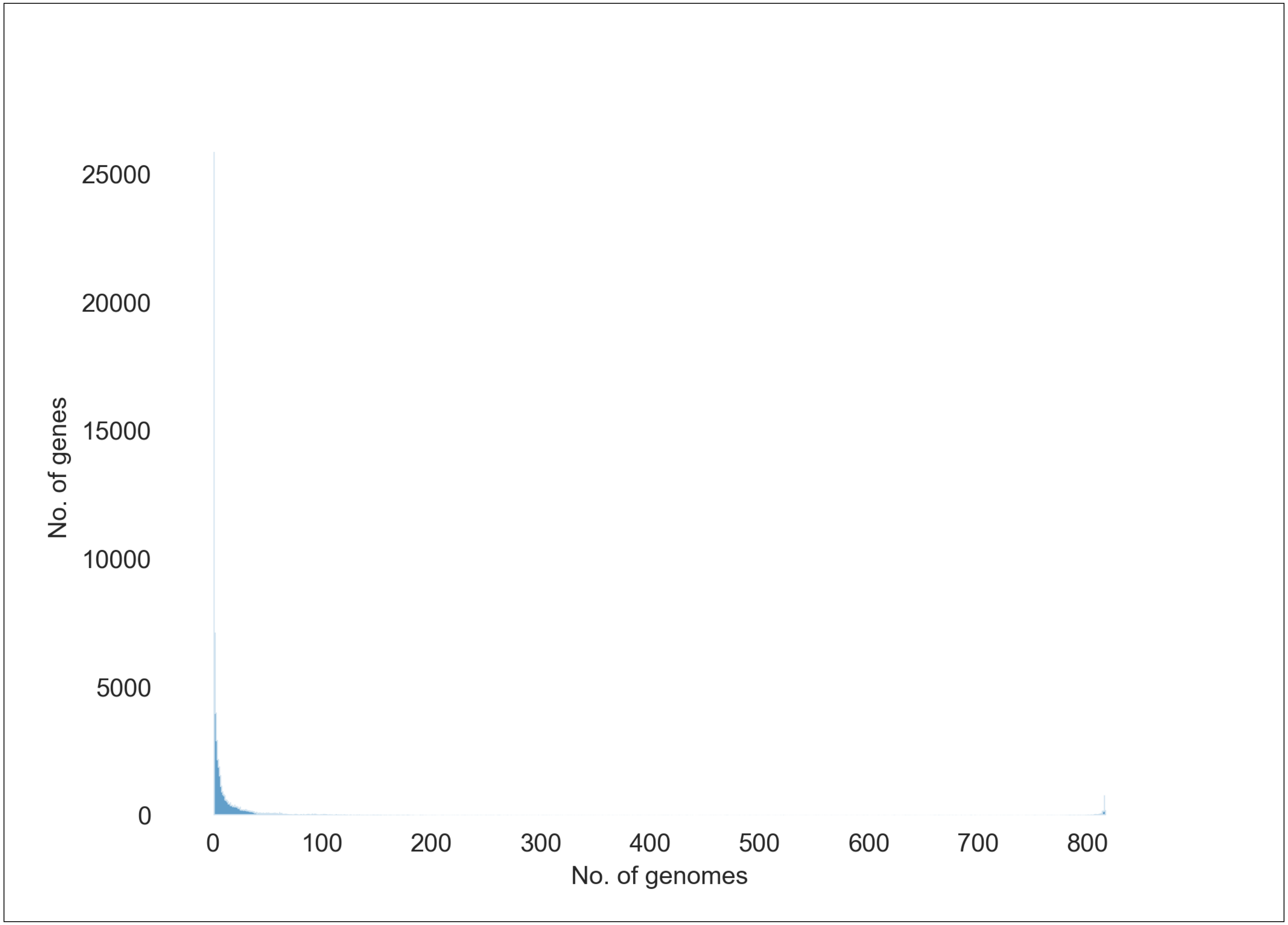
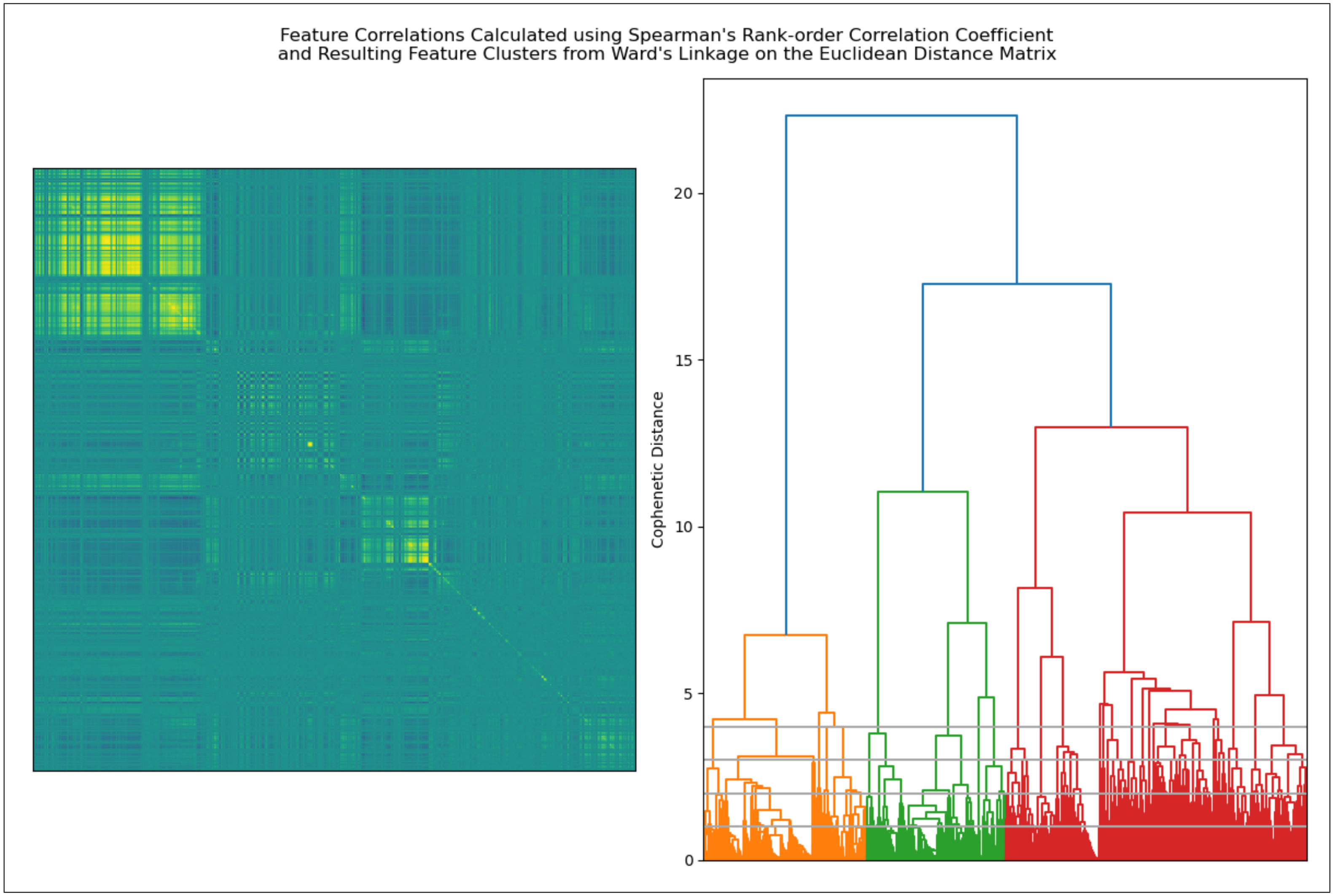


Figure 3.2: Count of genes annotated per genome.



Feature Correlations Calculated using Spearman's Rank-order Correlation Coefficient
and Resulting Feature Clusters from Ward's Linkage on the Euclidean Distance Matrix

Figure 3.3: At left, correlation matrix from Spearman's rank order coefficients. At right, the resulting hierarchical clustering using Ward's method.

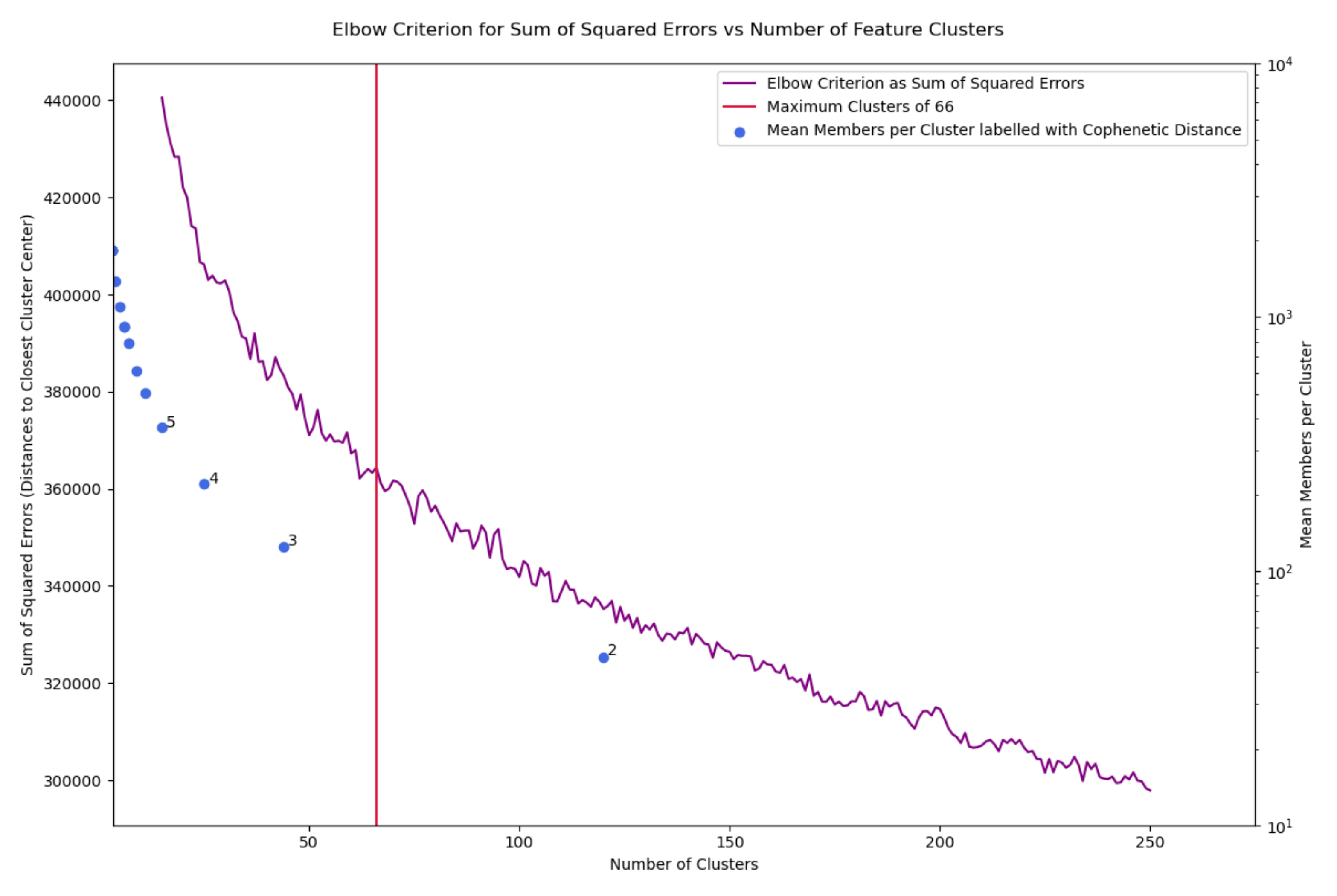


Figure 3.4: Visual representation of the elbow criterion approximation for determination of feature clusters using Ward's method.

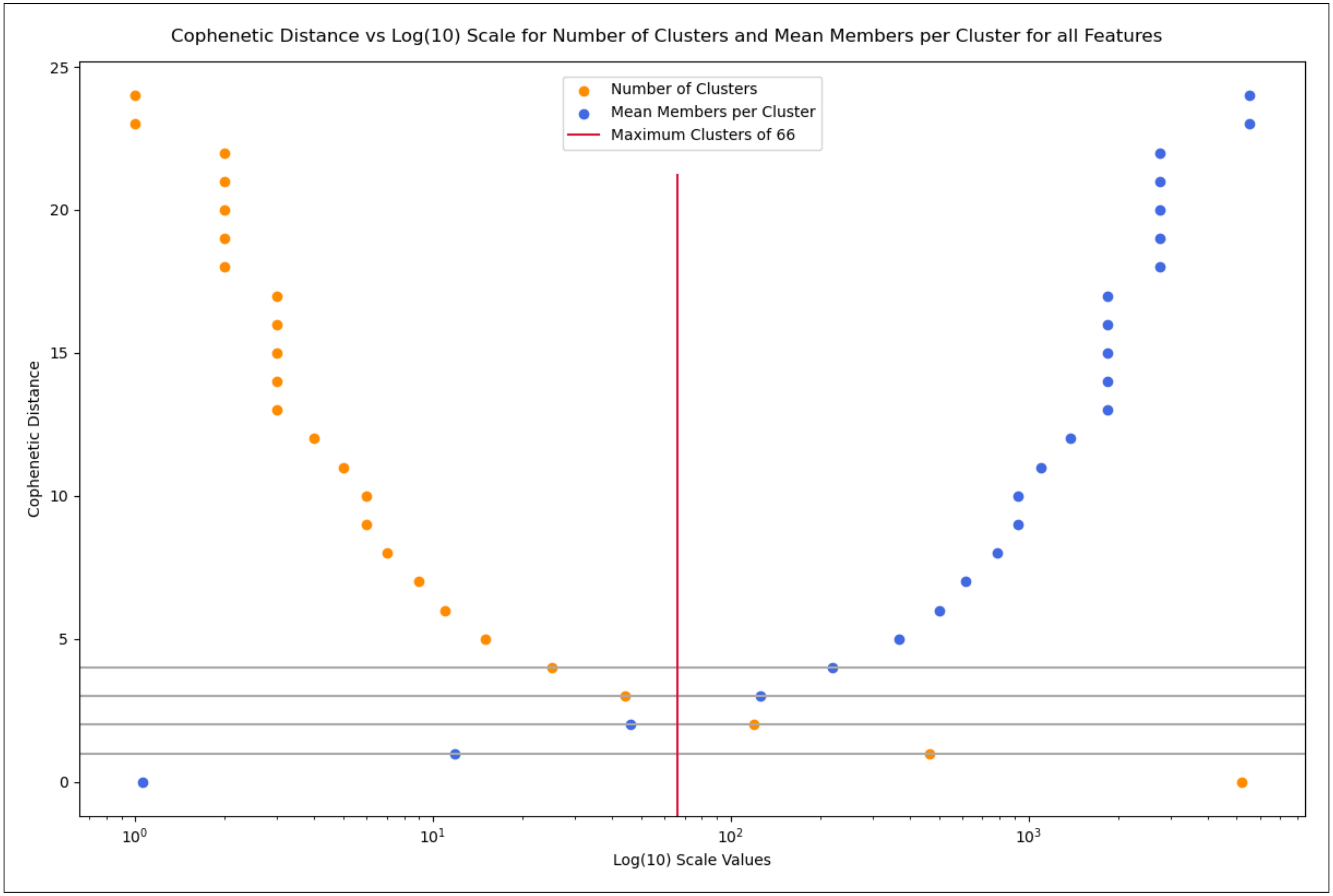


Figure 3.5: Clusters and mean members per cluster plotted on cophenetic distance and log base 10 axes. Where the orange and blue curves intersect, the mean members per cluster and total number of clusters are roughly equivalent. This approach maximizes the amount of information present in each cluster. The cluster count I implemented for this study (66) displays in red, close to the intersection point of the two curves.

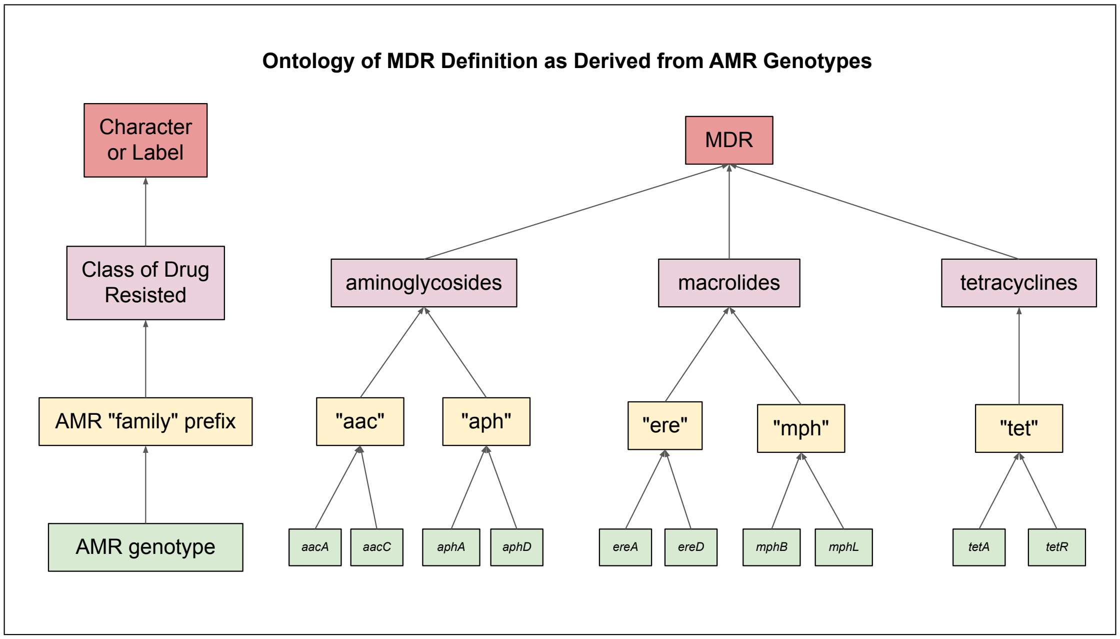


Figure 3.6: I defined binary MDR labels from annotated AMR genotypes. Red indicates the level of granularity used for the binary character matrix in Chapter 2. Pink represents the labels chosen for the machine learning investigation in Chapter 3. AMR prefixes in yellow for comparison against Table 2.1. In green are examples of the original AMR gene annotations from the metadata file.

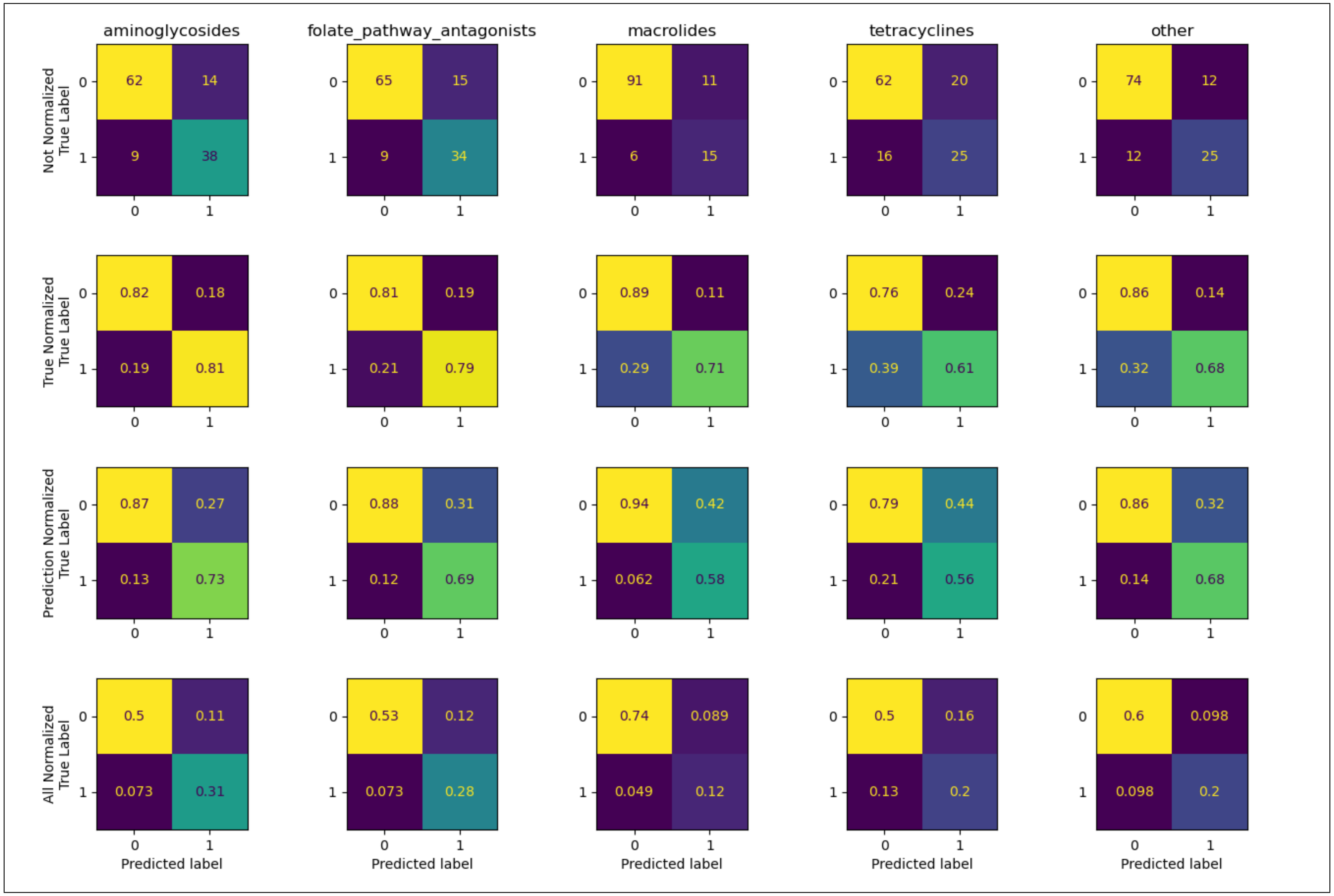


Figure 3.7: Comparison of confusion matrices for each of the five binary classifiers when applying a strategy of random oversampling of the minority class.

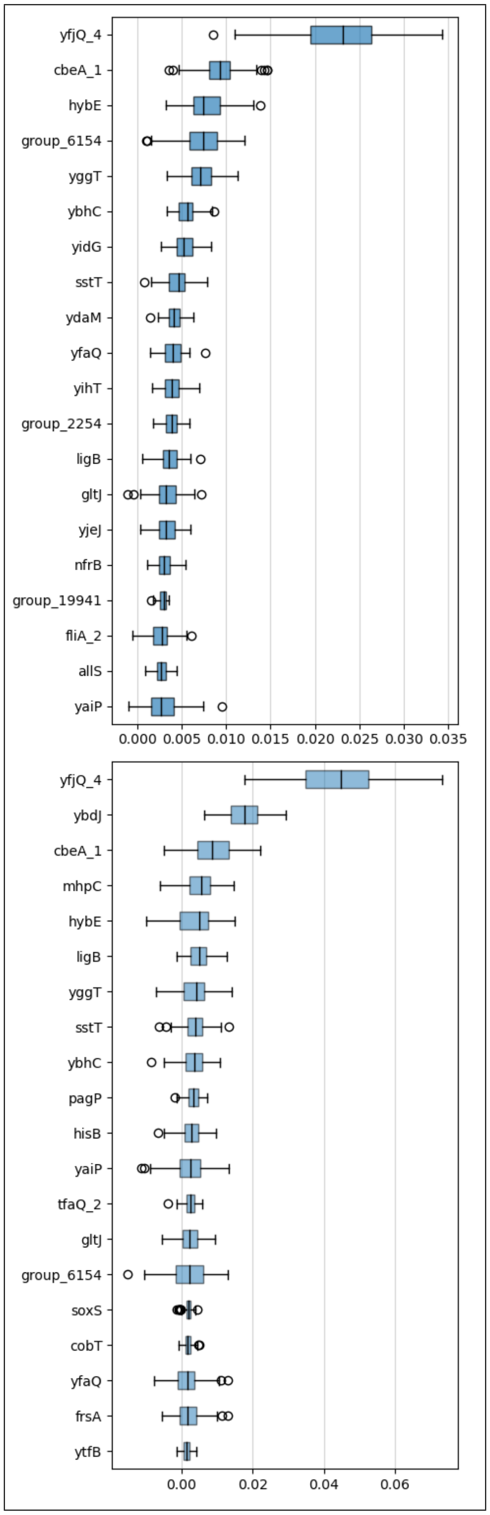


Figure 3.8: Box and whisker plots of permutation importance values (sorted descending) on training and validation data, top and bottom respectively, for the **aminoglycosides** classifier. Quartiles 2 and 3 are shaded with median indicated. Outlier values are represented by empty circles.

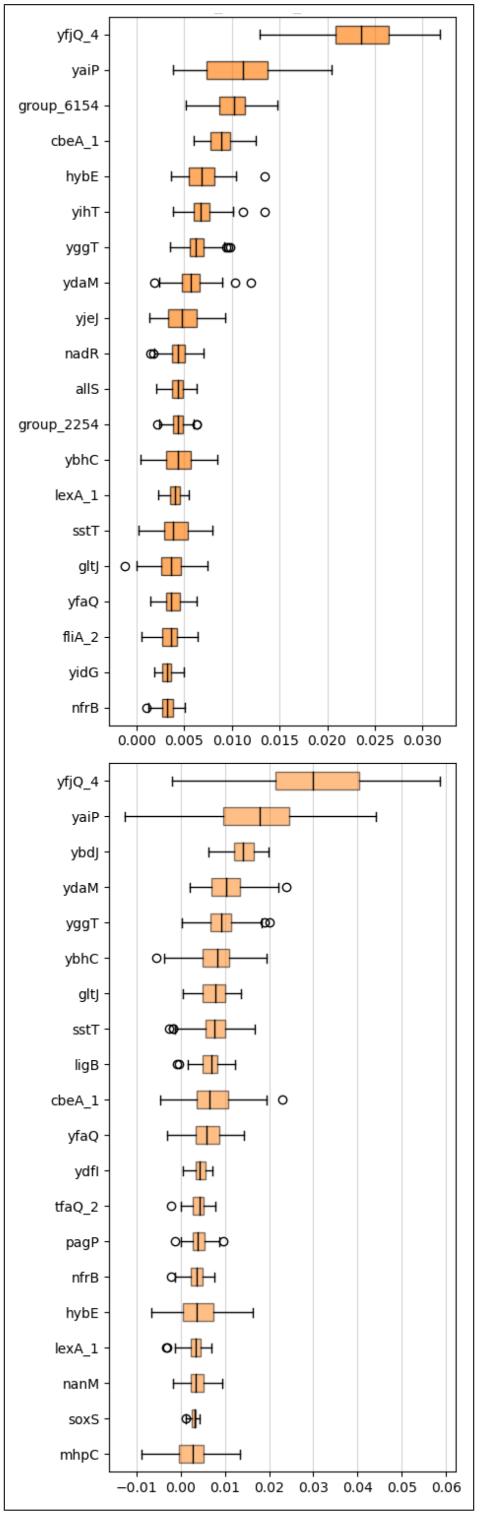


Figure 3.9: Box and whisker plots of permutation importance values (sorted descending) on training and validation data, top and bottom respectively, for the **folate pathway antagonists** classifier. Quartiles 2 and 3 are shaded with median indicated. Outlier values are represented by empty circles.

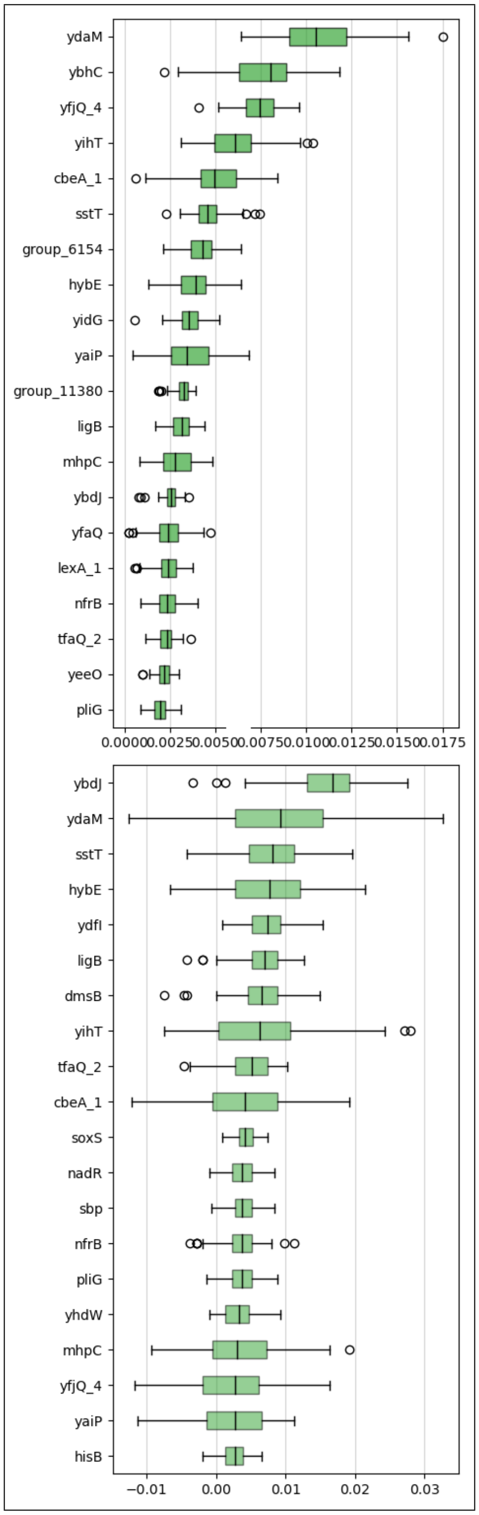


Figure 3.10: Box and whisker plots of permutation importance values (sorted descending) on training and validation data, top and bottom respectively, for the **macrolides** classifier. Quartiles 2 and 3 are shaded with median indicated. Outlier values are represented by empty circles.

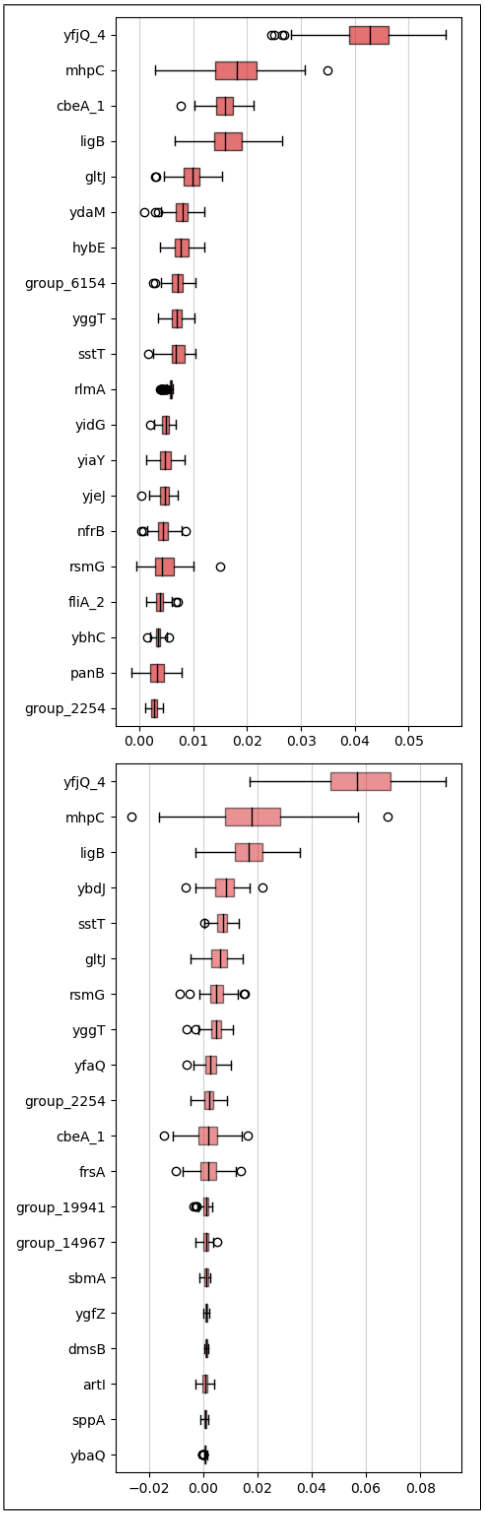


Figure 3.11: Box and whisker plots of permutation importance values (sorted descending) on training and validation data, top and bottom respectively, for the **tetracyclines** classifier. Quartiles 2 and 3 are shaded with median indicated. Outlier values are represented by empty circles.

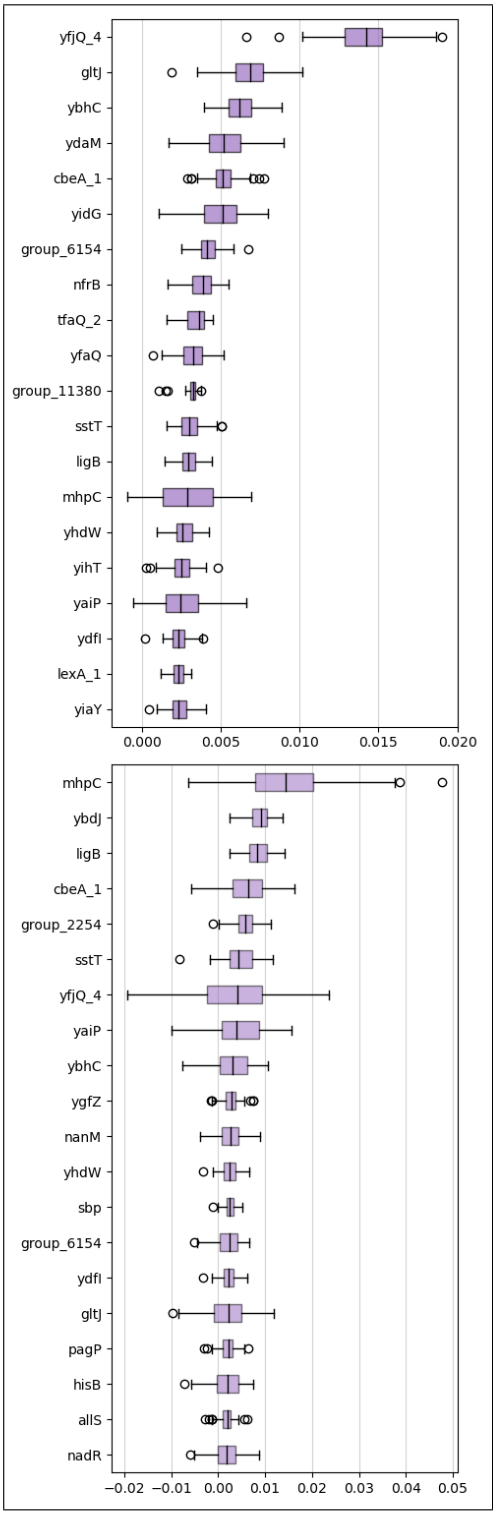


Figure 3.12: Box and whisker plots of permutation importance values (sorted descending) on training and validation data, top and bottom respectively, for the **others** classifier. Quartiles 2 and 3 are shaded with median indicated. Outlier values are represented by empty circles.

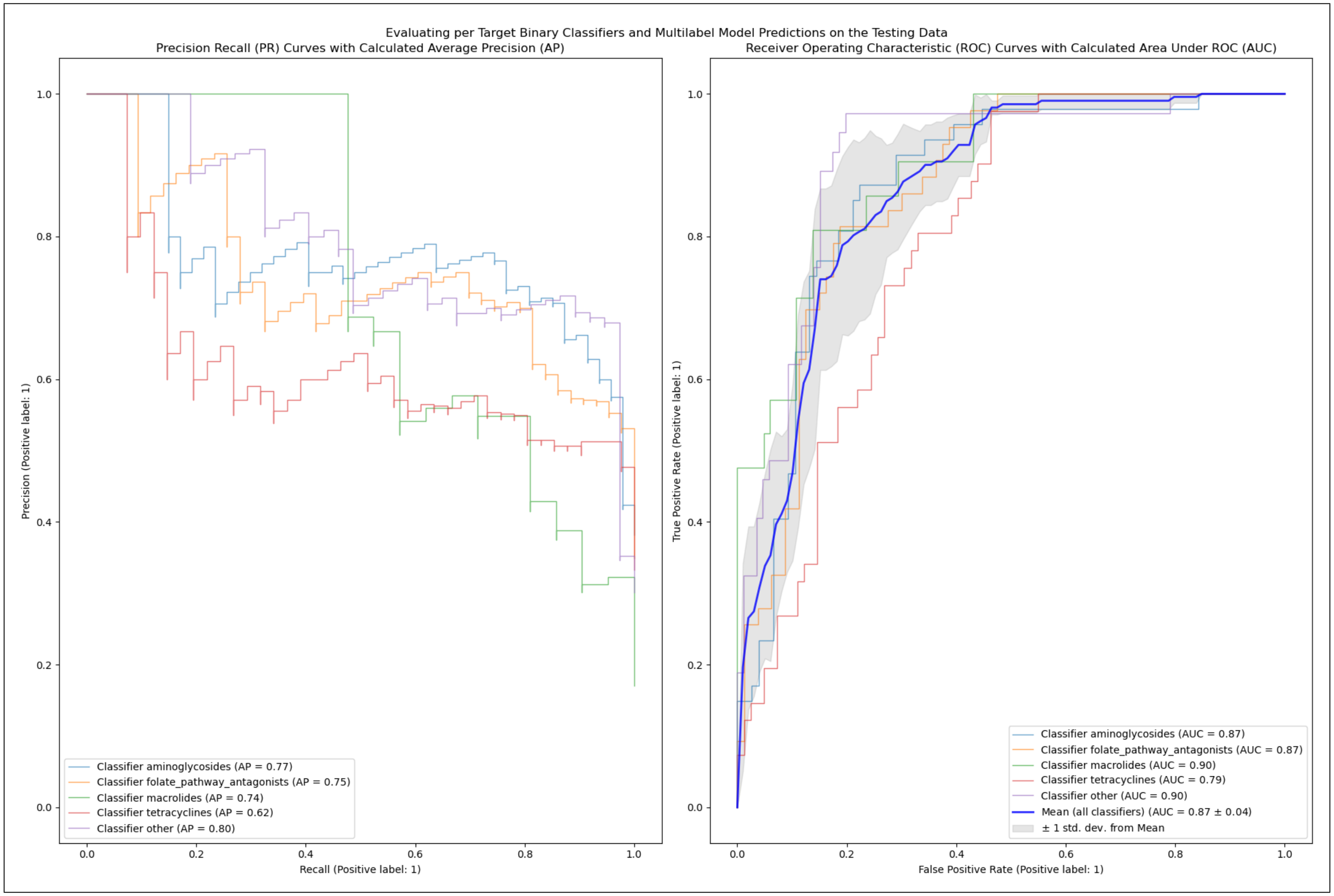


Figure 3.13: Evaluation of the five classifiers with precision recall (left) and ROC (right) curves on the testing data. Mean ROC curve represents the combined score of the binary classifiers when predicting the multilabel problem.

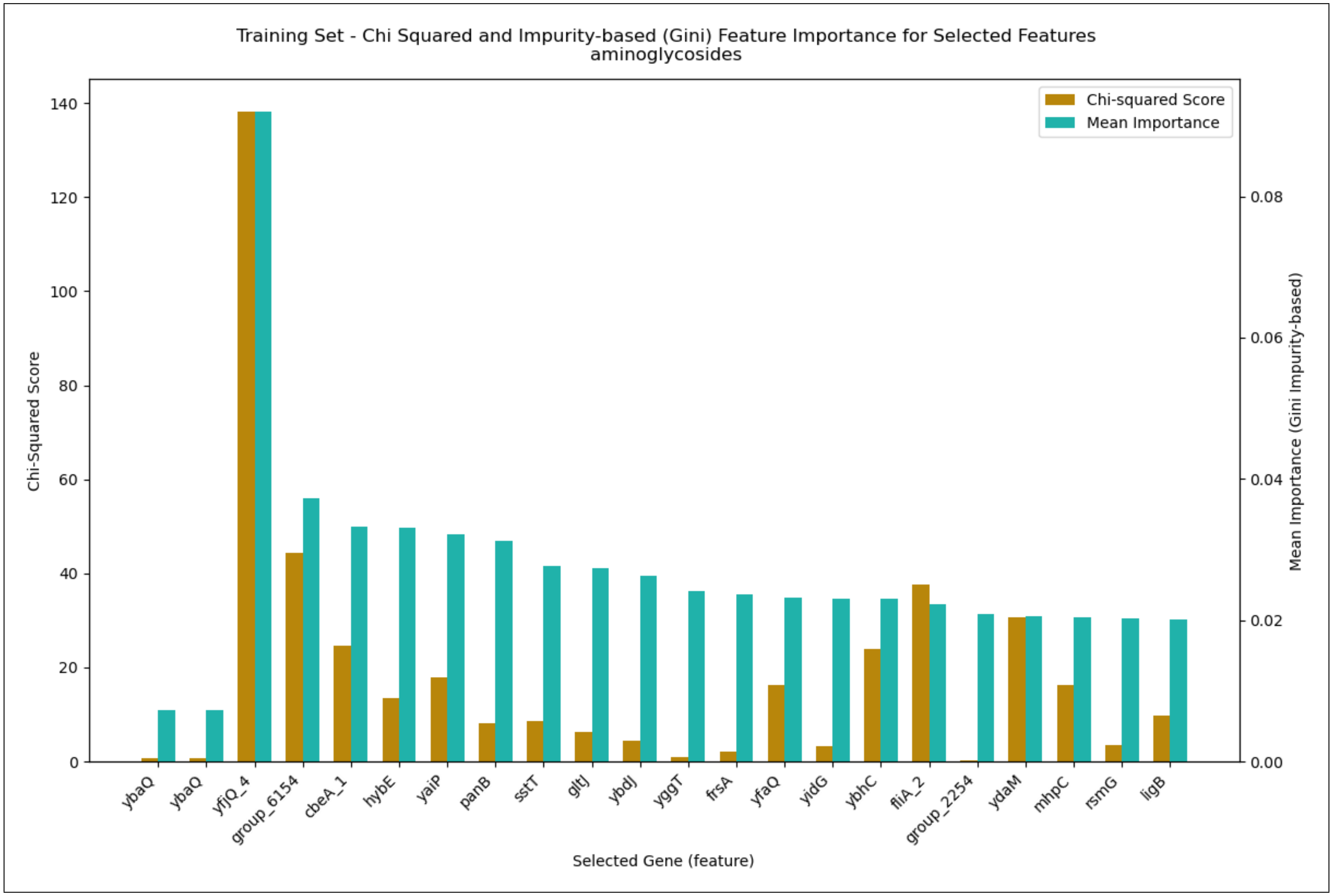


Figure 3.14: A bar plot representation of the difference between Chi-square Score (high correlation) and mean impurity-based (Gini) feature importance when identifying important features for the **aminoglycosides** model. This figure compares the box and whisker plots for permutation importance. Top 20 features sorted by decreasing mean importance (Gini). The left two bars represent the cluster (shown twice) once for mutS and again for the mutS paralog.

CHAPTER 4: EXAMINING SELECT IMPORTANT GENETIC PREDICTORS OF MDR IN *E. COLI*

4.1 Introduction

The ultimate goal of my dissertation is a more thorough understanding of the mechanisms driving the progress of the silent pandemic, via the proxy of end-stage AMR: MDR phenotypes. Initially in Chapter 2, I described the implementation of phylogenetic trees for understanding MDR in the context of MMR. I used phylogenetic comparative methods in Chapter 2 to identify dependent evolution between an MDR phenotype character and hypermutator genotype characters. I found those results unsatisfactory, supporting independent evolution between the traits, instead of revealing more useful genotype-to-phenotype hypotheses. To identify a new genetic explanation for MDR, I changed my methods.

In Chapter 3, I employed a supervised machine learning approach for uncovering genetic patterns that predict MDR. That analysis did not develop specific genotype-to-phenotype hypotheses of MDR, but instead corroborated the results of Chapter 2. The Chapter 3 results reveal that the ***clusters yfjQ_4*** and ***ydaM*** of annotated genes best predict MDR, as expressed by the five individual classifiers for different categories of AMR. Neither ***cluster yfjQ_4*** nor ***ydaM*** include the *mutS* gene indicated by LeClerc et al., 1996 as necessary for hypermutation [37]. Although of lower rank than ***cluster yfjQ_4***, the cluster ***cluster cbeA_1*** also appears in every the top five ranks of importance for all classifiers. I selected these three clusters for further analysis, as representative of the predictive ability of my five trained models. To improve interpretation of my results, I develop genotype-to-phenotype hypotheses of MDR, by examining the top ranked feature clusters, ***yfjQ_4***, ***ydaM***, and ***cbeA_1***, from Chapter 3 in fuller detail (see Table 3.1).

Effective models contextualize or simplify intractable problems, without resorting to self-evidence. On its own Chapter 3 contains two notable drawbacks: feature explainability and the potential for self-evident labeling. A valid criticism of the previous chapters is that my approach, training classifiers based on clustered annotations taken from a largely unshared subset of the pangenome, draws from overly complex inputs and outputs. To improve analyses when interpreting results, machine learning models should be examined by their inputs and outputs, instead of only by the logic of the classifier [177]. In addition, further criticism could logically assert that the best predictors of MDR phenotypes are highly likely to be known AMR gene determinants.

In this chapter, I aim to preemptively address both criticisms, by more thoroughly investigating the contents of the select predictive features, ***yfjQ_4***, ***ydaM***, and ***cbeA_1***, from the classifiers. Current research indicates a consideration of the biological context improves interpretation of machine learning models [144, 158, 154]. For the models trained in Chapter 3 to inform our understanding of the AMR pandemic, the identified feature clusters must supply evidence to support the development of MDR phenotypes. The features that best predict MDR phenotypes should 1) be in concordance with experimental results, and 2) not reduce the question of MDR phenotype to the simple presence or absence of specific AMR genes. Ideally, the best predictive features should include genes that underlie the proliferation of AMR, without expressly representing known AMR genes. The null hypothesis of this investigation asserts that select best predictors of MDR phenotypes identified in Chapter 3, specifically ***yfjQ_4***, ***ydaM***, and ***cbeA_1***, are composed of known AMR genes. My alternative hypothesis states that the most useful predictors of MDR phenotypes are instead genes directly involved in transmitting AMR genes.

4.2 Materials and Methods

When splitting paralogs, Roary assigns an internal identifier following the schema group_##### [106]. For example, although Prokka annotated the *mutS* gene, Roary identified a context dependent paralog of the *mutS* annotation, designated as *group_24522*. Roary found *mutS* 680 times and *group_24522* 136 times, matching the 816 identified sequences of *mutS* in Chapter 2 (missing only GCF_002843685 as noted). I carried out this process for all annotated genes found in less than 95% and greater than 10% of the pangenome, encompassing all of the shell and the top 5% of the cloud sections. During these assignments, Roary also standardizes uncertain language or annotations using the catchall term hypothetical protein [106]. Roary recorded the results of this process in the output *gene_presence_absence.csv*, the file which I then passed as input to the methods described in Chapter 3. While this approach cleans the data, it might also lead to overly simplified downstream analyses. To improve the explainability of the Chapter 3 models, I identify the annotation data collectively grouped under the hypothetical protein definition by Roary.

At the end of the processes described in Chapter 3, I collected the top five ranked feature clusters from each binary classifier (see Table 3.1). I recorded the annotated contents of each cluster, examined the annotations, and gathered all hypothetical protein annotations from the top five ranked features. I queried the previously described SQL database for the annotated gene name and added the first match to a large multiFASTA file in preparation for analysis via the Basic Local Alignment Search Tool (BLAST) method using a nucleotide query against a protein database (BLASTX) [178]. I separated the file into batches of 20 sequences and sent each batch to the BLASTX program implemented on the UniProt website as “blastx” to be searched against the “UniProtKB reference proteomes + Swiss-Prot” databases. I used the automatic settings (BLOSUM62, E-Threshold 10, gapped, no filtering) on the UniProt website, only restricting the results by taxonomy for *E. coli* (NCBI:txid562), Enterobacterales (NCBI:txid91347), viruses (NCBI:txid10239), and unclassified viruses (NCBI:txid12333) as found in the NCBI Taxonomy [179]. I included Enterobacterales and both viral taxonomies to account for the hypothetical protein annotations that Roary assigned, thinking at the time that closely related bacterial genes, lysogenic viruses, or prophages may be found in the data. I then manually curated the outputs of the BLAST program, noting the best matches by e-value, length, and percent identity. I took care to emphasize hits of similar length (both query and target), preferring to maximize length and identity while minimizing e-value. I also recorded sequences similar to the highest matches (of the same score). I performed the search against the Uniprot/TREMBL data set, but I placed emphasis on human annotated SwissProt sequences to maintain high confidence in my results. I then matched the resolved uncertainties with the results of Roary. Due to the larger size of the training data set, I derived my inferences about best predictors from the training and not the validation data, although I observed similar patterns between the two subsets.

4.3 Results

To reiterate the results of Chapter 3, the *mutS* gene is not important for predicting an MDR phenotype. However, I noticed that although Roary identified two *mutS* paralogs, the hierarchical clustering on Spearman's rank correlation coefficient grouped both paralogs into the same cluster (see Figure 3.14). I found this fact especially interesting, given that clustering placed paralogs of the *umuC* gene into several different feature clusters. The *umuC* gene participates in *E. coli* SOS mutagenesis and repair. The *umuC* annotation grouped with both *mutS* paralogs, while *umuC_2* associated with the top ranked permutation importance feature for predicting MDR (***cluster yfjQ_4***) as shown in Table 4.1. This observation indicates that 1) the models do not immediately cluster closely-related proteins and 2) highly similar proteins by function are not grouped together by default. The data processing and model training emphasized the importance of specific proteins, purely from the annotation of high-quality WGS.

Nearly all of the genes in the top ranked ***cluster yfjQ_4*** are from the protein families Tra and Trb families, which are involved with the horizontal transfer of genetic material. The potential exceptions are *umuC_2*, a pilin protein, and a putative CP4-57 prophage gene. I observed no AMR genes, either as annotated by Prokka and Roary nor matching the list of AMR gene families from Chapter 1 (see Table 2.1). The theme of ***cluster yfjQ_4*** is HGT, as shown by specific proteins in allowing for the formation of conjugative structures.

Table 4.1: All component annotations found within ***cluster yfjQ_4***, which ranked as most important for the **aminoglycosides**, **folate pathway antagonists**, **tetracyclines**, and **others** models. The first (*Gene*) column of the Roary output file *gene_presence_absence.csv* provides the identifying Gene name. The best annotation, protein name, or gene identifier is taken from Roary or BLAST results.

Gene (Roary)	**Computationally Predicted Protein Name (BLAST // Roary)**
group_16177	LPD25 domain-containing protein
group_10103	Protein TrbI
group_10424	TraK lipoprotein
group_12760	Protein TraQ
group_7878	Protein TraW
group_19502	Protein TraJ
group_4098	Protein TraB
group_24638	Terminase
group_9019	Protein TraE
group_11161	Protein TrbG
group_5897	Protein TraP
group_4701	Protein TrbJ
group_9337	Protein TraV
group_17382	Protein TraL
group_2543	Protein TrbB
group_4818	X polypeptide, ORF 19, ORF169, P19 protein
umuC_2	DUF4113 domain-containing protein // SOS mutagenesis and repair
traM	Relaxosome protein TraM
traY	Relaxosome protein TraY
group_2526	Protein TraC
ccdB	Uncharacterized protein YuaT
traA	Pilin, F-pilin
yfjQ_4	UPF0380 protein YubP
group_5903	Protein TrbD

Top ranked ***ydaM*** for the **macrolides model** is seen in Table 4.2. The presence

Table 4.2: Top ranked (highest permutation importance) ***cluster ydaM*** that best predicts resistance in the **macrolides** model. Indicates the best annotation, protein name, or gene identifier taken from BLAST results. *Non-unique Gene name*, the second column of the Roary output file *gene_presence_absence.csv*, indicates Roary split paralogous genes into individual annotations based on sequence context. Note *ydaM* and *group_11293* as well as *dbpA* and *group_25593*.

Gene (Roary)	**Non-unique Gene name (Roary)**	**Computationally Predicted Protein Name (BLAST)**
ydfN_2		ParB-like nuclease domain-containing protein YnaK
group_11781		Uncharacterized protein YdaU
group_31990		Uncharacterized protein YdaQ
dbpA		ATP-dependent RNA helicase DbpA, 3.6.4.13
sieB		Phage superinfection exclusion protein
ydaF		Uncharacterized protein YdaF
rzpR		Putative endopeptidase RzpR, 3.4.-.-, Putative Rz endopeptidase from lambdoid prophage Rac
group_11293	*ydaM*	Diguanylate cyclase DgcM, DGC, 2.7.7.65
group_25593	*dbpA*	ATP-dependent RNA helicase DbpA, 3.6.4.13
ydaV		Uncharacterized protein YdaV
trkG		Trk system potassium uptake protein TrkG
tfaR		Prophage tail fiber assembly protein homolog TfaR, Tail fiber assembly protein homolog from lambdoid prophage Rac
pinR		Serine recombinase PinR, 3.1.22.-, 6.5.1.-, DNA-invertase PinR, Putative DNA-invertase from lambdoid prophage Rac, Site-specific recombinase PinR
ydaM		Diguanylate cyclase DgcM, DGC, 2.7.7.65
ydaG		Uncharacterized protein YdaG
intR		Prophage integrase IntR, Putative lambdoid prophage Rac integrase
ydbH		Uncharacterized protein YdbH
racC		Protein RacC
ydaE		Uncharacterized protein YdaE
ynaE		Uncharacterized protein YnaE
recE		Exodeoxyribonuclease 8, 3.1.11.-, Exodeoxyribonuclease VIII, EXO VIII
ydaW		Protein YdaW
ydbK		Probable pyruvate-flavodoxin oxidoreductase, 1.2.7.-
ompN		Outer membrane porin N, Outer membrane protein N, Porin OmpN
ydaT		Uncharacterized protein YdaT
group_10760	*ompN*	Outer membrane porin N, Outer membrane protein N, Porin OmpN
kilR		Killing protein KilR, FtsZ inhibitor protein KilR
recT		Protein RecT, P33
group_10767	*ydbH*	Uncharacterized protein YdbH
group_37342	*racR*	Prophage repressor RacR, Rac prophage repressor
group_18678	*ydbK*	Probable pyruvate-flavodoxin oxidoreductase, 1.2.7.-
rcbA		Double-strand break reduction protein

of many prophage genes in this cluster is notable, again demonstrating a theme of HGT. In addition, both *ydaM* and it's paralog *group_ 11293* were placed in the same feature cluster.

Table 4.3: ***Cluster cbeA_ 1*** and its zero-based rank across all five classes of resistance: **aminoglycosides** (Amino.), **folate pathway antagonists** (Fol. Path. Antagonists), **macrolides** (Macro.), **tetracyclines** (Tetra.), and **others** (Others). *Non-unique Gene name*, the second column of the Roary output file *gene_ presence_ absence.csv*, indicates Roary split paralogous genes into individual annotations based on sequence context. Note *yfjQ_ 1* and *group_ 257* as well as *yeeP_ 1* and *group_ 1508.*

Gene (Roary)	Non-unique Gene name (Roary)	Amino. Rank	Fol. Rank	Macro. Rank	Tetra. Rank	Others Rank
group_ 13421		1	3	4	2	4
group_ 4945		1	3	4	2	4
group_ 22310		1	3	4	2	4
group_ 7475		1	3	4	2	4
group_ 9598		1	3	4	2	4
yfjJ		1	3	4	2	4
group_ 7384		1	3	4	2	4
group_ 5273		1	3	4	2	4
group_ 6844		1	3	4	2	4
group_ 13325		1	3	4	2	4
group_ 4579		1	3	4	2	4
cbeA_ 1		1	3	4	2	4
yeeS_ 1		1	3	4	2	4
php_ 1		1	3	4	2	4
group_ 684		1	3	4	2	4
group_ 257	*yfjQ_ 1*	1	3	4	2	4
group_ 1508	*yeeP_ 1*	1	3	4	2	4

Within the commonly occurring ***cluster cbeA_ 1***, the composition of observed genes changes as shown in Table 4.3. I observed many TA genes, along with several prophage genes and chemotaxis proteins demonstrated in Table 4.4. Interestingly, I also found a putative repair protein among the other annotated genes comprising the prophage. The theme of***cluster cbeA_ 1*** is also HGT, as indicated by specific proteins that allow for the selection of plasmids or prophages belonging to exclusive lineages (Inc regions).

Notably, while I found some specific AMR gene annotations in this data, those

Table 4.4: Contents and annotations for ***cluster cbeA_ 1***. Best annotation, protein name, or gene identifier taken from BLAST results. *Non-unique Gene name*, the second column of the Roary output file *gene_ presence_ absence.csv*, indicates Roary split paralogous genes into individual annotations based on sequence context. Note the number of toxin-antitoxin systems, lipoproteins, phospholipases, and repair proteins suggestive of phage interactions.

Gene (Roary)	**Non-unique Gene name (Roary)**	**Computationally Predicted Protein Name (BLAST)**
group_ 13421		Zinc ribbon domain-containing protein
group_ 4945		RelB antitoxin
group_ 22310		Uncharacterized protein
group_ 7475		DUF3310 domain-containing protein
group_ 9598		Chemotaxis protein
yfjJ		Uncharacterized protein YagK
group_ 7384		Regulatory protein
group_ 5273		Phospholipase
group_ 6844		C8 domain-containing protein
group_ 13325		Lipoprotein
group_ 4579		Inovirus Gp2 family protein
cbeA_ 1		CP4-44 prophage; antitoxin of the CbtA-CbeA toxin-antitoxin system
yeeS_ 1		CP4-44 prophage; predicted DNA repair protein
php_ 1		putative hydrolase
group_ 684		Dynamin-type G domain-containing protein
group_ 257	*yfjQ_ 1*	CP4-57 prophage; predicted protein
group_ 1508	*yeeP_ 1*	CP4-44 prophage; predicted GTP-binding protein

specific annotations are never in the top two ranked most important (best) predictors for any estimator. Annotations from Prokka, Roary, and subsequent BLAST analysis seem to agree on general themes via related annotations (genes with similar or related functions are grouped together often, but not exclusively grouped together). As demonstrated in Chapter 3, each model relies on different predictive features (see Table 3.1), and the clustered gene features are not solely drawn from groups related either by evolution or structure.

In the results and discussion of this fourth chapter, I focus on the highly ranked feature ***clusters yfjQ_4***, ***ydaM***, and ***cbeA_1*** for the sake of brevity and space. The full spreadsheet of joined Roary and BLAST annotations from the top five ranked best predictors for the five trained classifiers is available upon reasonable request of the author.

4.4 Discussion

First, the best predictors of each binary resistance class and therefore MDR phenotypes are genes associated with HGT. Each resistance class possesses an identifying set of clusters that best predicts resistance to aminoglycosides, folate pathway antagonists, macrolides, tetracyclines, and other antimicrobial compounds. The contents of each cluster are unique, so that while some clusters contain a few known AMR genes, I did not define the clusters seen in Tables 4.1, 4.2, or 4.4 as exclusively AMR. For these unique clusters, the selected feature clusters ***yfjQ_4***, ***ydaM***, and ***cbeA_1*** all show evidence of HGT.

Of note in the ***cluster ydaM*** are the genes attributed to the Rac prophage. Rac prophages are excisionable DNA regions [180] often associated with erythromycin resistance [181]. Erythromycin falls under the category of macrolide antimicrobials. This ***cluster ydaM*** ranks the highest for macrolide resistance and as rank 3 (0-based) for **others** resistance (see Table 3.1). The only time another Rac prophage gene appears among the top ranks is in ***cluster ligB*** which ranks 3 (0-based) for tetracycline resistance, where only a single gene "ydaV_2 Rac prophage; predicted DNA replication protein" appears. Most of the Rac prophage genes clustered together in ***cluster ydaM***, with the exception of the single gene. The macrolides model trained in this inquiry demonstrates that ***cluster ydaM*** is highly important for the prediction of macrolide resistance. This result concurs with both scientific literature and research, where the presence of the Rac prophage imparts resistance to macrolide antibiotics like erythromycin and fosfomycin [180].

Notably, in the case of erythromycin resistance, researchers point to overexpression of existing genes as contributing to AMR [181]. The models trained in Chapter 3 rely on the simple presence or absence of annotated genes within an *E. coli* sequence. When a gene is present in an isolate, the potential exists for the dosage of that gene to change, resulting in the overexpression or underexpression of that gene. Conversely, a bacterium cannot express genes that its genome lacks. In the methods of this chapter, I applied that concept to computationally identify MDR phenotypes without the need for susceptibility testing. Sequencing and proper annotation techniques may suffice to identify the capacity of bacteria for achieving MDR phenotypes. In this way, decision makers may leverage survey data describing AMR genes that currently circulate in a locale, to improve prescribing practices or food safety in a region [3].

The contents of each cluster are exclusive, so no single annotation occurs in more than one cluster. The Roary function of context-dependent paralog splitting adds granularity to the annotations initially assigned by Prokka. I engineered a binary presence or absence transformation of the annotation data using the Spearman's rank correlation coefficient to arrive at clustered groups of features. As evidenced by the selected feature clusters ***yfjQ_4***, ***ydaM***, and ***cbeA_1***, these clustered groups of features included highly similar genes by function or actual paralogs. However in some cases, paralogs are not found together within the same feature cluster. The separate grouping of paralogous genes opens an avenue for continuing research: to investigate sequence or context discrepancies between paralogs in order to better understand the emergent phenotype of MDR. Further, the exclusive nature of the clusters leads to an important conclusion: a binary representation of annotated accessory genes better predicts MDR phenotypes than the AMR genes usually attributed to imparting a resistance phenotype. Investigation of the annotations comprising these best predictors reveals currently known virulence genes and genes associated with HGT. The models developed in Chapter 3 and discussed in Chapter 4 explain the genetic determinants of the AMR silent pandemic in a biological context.

CHAPTER 5: CONCLUSIONS

The continued proliferation of AMR genes among the global bacterial population leads to MDR phenotypes. The problem of increasing MDR represents an opportunity for further research to support public health. Polygenic traits like MDR result from the complex interactions of multiple genes. Two main mechanistic hypotheses exist for explaining the rise in MDR bacterial populations: the generation of novel AMR genes through hypermutation or the population-scale proliferation of AMR genes via horizontal gene transfer. The main conclusions drawn from the completion of this research program are: 1) there is little to no observed support for a dependent evolutionary correlation between *mutS* and an MDR phenotype in *E. coli* as shown in these data, and 2) annotated genetic features associated with HGT best predict the MDR phenotype over both mutation-associated and AMR-associated genes which are traditionally understood to determine a bacterial resistance phenotype.

5.1 The *mutS* Gene and MDR Evolve Independently

In my first investigation, I used a data set of 817 high quality *E. coli* isolates to test an alternative hypothesis of dependent evolution between MDR and hypermutable phenotype *mutS*. I found insufficient evidence to invalidate the null hypothesis of independent evolution when comparing the AIC scores of models calculated with the `fitPagel` [117] implementation of Pagel's 1994 algorithm [116]. These findings do not agree with the paper by LeClerc et al., 1996 [37], as I was unable to observe the existence of the defective *mutS* gene defined by them. My results support independent evolution between variant amino acid positions in MutS and MDR phenotype *E. coli*.

Hypermutable phenotypes are not easily identifiable from WGS data if the traditional, quantifiable definition of hypermutator *mutS* (defective gene with a 221bp 3' deletion) [37] is unobservable. If instead, a more qualitative definition of hypermutation is chosen (e.g., variant positions in protein or gene sequences, temporary or induced hypermutable states, or overexpression of mismatch repair genes) it becomes challenging to represent the development of novel AMR mutations. To improve our understanding of the ongoing AMR pandemic, analyses of MDR must reconsider the role and definition of hypermutation.

5.1.1 Re-examining Hypermutation as Inducible

An emphasis on hypermutation overshadows the role of horizontal gene transfer in the proliferation of AMR genes into increasingly MDR phenotypes. As LeClerc et al., 1996 notes in the closing argument, they brought the hypermutator hypothesis to counter ideas around horizontal gene transfer.

> "Although they are counter to the current paradigm that antibiotic resistance is due to the acquisition of plasmids harboring multiple drug-resistant determinants, our data show that chromosomal mutations might explain at least some of the multiply drug-resistant organisms found clinically. These same mutator phenotypes could help explain how the plasmid-borne resistance determinants first became linked (relaxed recombination between disparate species) and are now so readily inherited." [37]

Thus, when I observe no evidence of a LeClercian hypermutator *mutS*, it is fair to consider the role of horizontal gene transfer. Horizontal gene transfer has been associated with AMR gene spread for decades, but as LeClerc noted HGT does not supply a mechanism for the development of novel AMR genes [37]. Therefore, although hypermutation (whether structural or induced) and horizontal gene transfer are well-documented phenomena, neither hypothesis fully explains the development and proliferation of AMR genes into MDR phenotypes. Hypermutation lacks a proliferative mechanism, while HGT lacks a developmental mechanism.

LeClerc's statements on hypermutation require nuance. I found no indication in my research that LeClerc's hypermutator *mutS* correlates with MDR, as originally suggested in the LeClerc et al., 1996 publication [37]. However, some AMR genes could be the result of increased mutation rates among bacteria, as LeClerc conjectured and as others demonstrated [79, 84, 37, 131, 130, 132, 133, 134]. The broader, inducible definition of hypermutation could drive AMR gene development [182, 84, 37, 133], while HGT plays the central role for the acquisition and inheritance of many AMR genes in *E. coli* [163, 5]. Expanding the definition of hypermutator to include inducible hypermutation certainly allows increased mutation rates to be associated with AMR genes, but as noted previously, overexpression of certain genes also imparts a resistance phenotype [183, 181]. To avoid confusion and arrive at a falsifiable hypothesis, I narrowly defined hypermutation in the original MutS deficient sense for the purposes of this research.

Other definitions of hypermutation, like inducible hypermutation, remain candidate hypotheses for novel AMR phenotypes, but based on the research I present in this manuscript, the traditional MutS deficient hypermutator phenotype does not correlate with MDR phenotypes at the population scale. LeClerc's original definition of hypermutator phenotype *mutS* is unobservable in my research data, so repair deficient *mutS* cannot explain the presence of the MDR phenotype.

A more apt understanding of LeClerc's paper should focus not only on the closing argument, but on an earlier statement. "Because MMR limits recombination between diverged sequences, inactivation of this system relaxes normal recombination barriers among species, offering a potential pathogen the opportunity to inherit, by horizontal transmission, useful genes from the reservoir of commensal and pathogenic bacteria at large. Promiscuity, then, might drive selection of these kinds of mutators among successful pathogens." [37] Recombination and promiscuity could drive MDR, an idea which I investigated next.

5.1.2 What Does MDR Require?

Multiple environmental factors interact when initiating a temporary hypermutable state in a bacterium [84, 80, 184] or when activating HGT processes [185, 72, 61, 186]. The food safety survey data investigated here does not include information about potential antimicrobial exposure for each sample nor gene expression data. Given the nature of the genomic surveillance data used, *in silico* modeling of these data will not address variable levels of gene expression. Gene expression data may be necessary to understand the mechanisms by which drug resistance develops in the short term and is fixed in the long term [5].

I chose to consider as much of the *E. coli* genome as possible without focusing only on select genes associated with the MMR system or from a notion of the core genome. This approach takes into account not just hypermutation or HGT, but other potentially novel hypotheses for understanding AMR proliferation. Future attempts to understand the evolutionary history of MDR should begin *de novo*, by testing for correlations between many annotated genes and MDR phenotypes, instead of focusing only on those genes involved in the evolution and acquisition of traits (like MMR associated genes or conjugation associated genes). Once my Chapter 2 analysis showed no support for dependent evolution between hypermutation and MDR, I attempted to identify which genetic determinants best predict MDR phenotypes in *E. coli.*

5.2 Best Predictors of MDR are not Known AMR or MMR Genes

Using a trained random forest classifier, I identified genetic patterns that predict MDR phenotypes from combinations of five unique, commonly screened antimicrobial compound classes. The five trained classifiers exhibit high accuracy, precision, and recall, standard metrics for benchmarking predictive models. In addition, structuring the investigation as a multilabel supervised machine learning problem enables a more granular understanding of MDR phenotypes, in contrast to the single binary phylogenetic character I used in Chapter 2. The use of gene annotations from the shell portion of the pangenome also expands on the previous analysis, by discovering useful patterns in genetic information that is neither shared nor derived in a phylogenetic sense.

5.2.1 Caveats of Predictive Features

Aside from the potential technical challenges in developing a kit, a major caveat is self-evident: **what if the best predictors of MDR phenotypes are simply known AMR genes?** Initial investigation of the best predictors revealed a large number of genes involved in horizontal gene transfer, with little to no occurrence of proteins involved in mismatch repair (specifically MutHLS) or genes involved in hypermutation (*uvrD*). While the important clusters include known prophages and transmissible genetic features, these clusters might best be described as virulence genes, or potential pathogenicity islands. On their own these results agree with current understandings of AMR proliferation. However, current food safety survey practices already annotate known AMR genes and suspected virulence genes through the PGAP [119, 120, 121]. If the genetic determinants of an MDR phenotype are simply the component AMR genes, then this research demonstrates nothing besides a previously understood correlation. With Chapter 4, I investigated and explained the contents of important feature clusters identified in the Chapter 3 inquiry to further understand the genetic predictors of MDR phenotypes.

5.3 HGT Mechanisms Best Predict MDR Phenotypes

To further validate the results of Chapter 3, I examined the hypothetical protein annotation that Roary [106] applies when standardizing inputs. I queried each hypothetical protein annotation using BLASTX to find the best matches and identify all component annotations in the top performing predictors. In the top predictors, I did not identify the presence of any genes that explicitly imparted resistance to an antimicrobial compound. Instead, nearly all annotated genes had some role in horizontal gene transfer, further validating the initial assessment of the clusters from Chapter 3. Explicit AMR genes are not as important for predicting AMR phenotypes or by extension MDR. Other researchers have pointed to pathogenicity islands (PAIs) as responsible for increasing the virulence of bacterial phenotypes [187, 188, 189]. My results concur with those researchers, that horizontal transmission of virulence and accessory genes drives acquisition of MDR phenotypes, not hypermutation or the AMR genes themselves. The models of Chapter 4 support that when addressing MDR, HGT mechanisms require more focus than hypermutation. HGT drives the proliferation of AMR genes into MDR phenotypes, given the evidence gathered in this study.

5.3.1 Potential Overlap with Pathogenicity Islands

If the clusters identified in Chapter 3 are pathogenicity islands, my research indicates that *E. coli* achieve MDR phenotypes through a combination of pathogenicity islands, each island targeted towards specific categories of antimicrobial compounds. More research is needed to verify these clusters as pathogenicity islands, specifically in regards to DNA segments flanking the clusters, the continuity and proximity of all component cluster genes, and G/C counts from the annotations. For now, the main conclusion of my research indicates that there is not a single genotype that results in MDR, but instead multiple pathways to broadly drug resistant *E. coli.* However, when examining resistance to specific categories of AMR compounds, I observed that each category of AMR compound is predicted by a single set of genes. As such, I interpret that *E. coli* strains appear to exhibit a modular, flexible genetic toolkit for adapting to evolutionary pressures brought by antimicrobial compounds. This view agrees with current understandings of bacterial evolution [80, 39], AMR spread via HGT [185, 72, 61], and the mobilome [9, 190, 17, 43].

5.3.2 High Importance of F Plasmid in AMR Proliferation and Resulting MDR Phenotypes

I find it concerning that the feature which best predicts resistance in four of the five categories is the same (***cluster yfjQ_4***). My concern increases when confronted by the fact that this cluster includes transfer genes (*tra*) associated with the F plasmid. The F plasmid enables bacterial conjugation, allowing an F plasmid positive bacterial cell to convert an F plasmid negative bacterial cell. The relatively high rank of ***cluster yfjQ_4*** for predicting the category of macrolide resistance leads me to question if there exists some selective pressure on *E. coli* to develop a single maximum fitness MDR strain by leveraging the F plasmid. My models fail to indicate the existence of that hypothetical maximum fitness *E. coli* strain. Each of the resistance categories varies from one another when comparing the most predictive clusters in second through fifth place (0-based ranks 1-4). I attributed a unique pattern of clusters, a gene set, to each resistance category. The specificity of each gene set drives the high accuracy, precision, and recall of the models trained in Chapter 3. I interpret such variation to mean that the combined genetic and resource burden to successfully resist all categories of antimicrobial compounds is not yet manageable for *E. coli.* However, if such a maximum fitness, pan-resistance genotype evolves, the predictive models that I developed in this research program can detect that threatening genotype.

5.4 Weak Mutators, HGT, and MDR

As noted previously, variable definitions for hypermutation exist, often centered around the relative strength of the mutator phenotype indicated by the number of mutations. Strong mutators differ from weak mutators, although both may fall under the broader definition of hypermutation that I acknowledge but do not use for this research. In this dissertation, I investigated a very specific definition of hypermutation that can be referred to as a strong mutator phenotype, the deficient MutS protein as described by LeClerc et al., 1996 [37]. My investigations 1) found no evidence that LeClerc's hypermutator phenotype existed in these data, 2) identified little to no dependent correlation between the MDR trait and MutS variant residues, and 3) noted that other features besides the *mutS* gene annotations best predict the MDR trait.

However, while I demonstrate that the canonical strong mutator is not associated with MDR, I did not rule out the effect of weak mutators, notably genes like *umuC* involved in SOS mutagenesis and repair [191]. Instead, *umuC* is present in the feature ***cluster yfjQ_4***, the top ranked best predictor in four of the five trained AMR classifiers. The exception is for macrolides, where ***cluster yfjQ_4*** ranks third (see Table 3.1). My evidence agrees with and builds upon the work of previous researchers who indicate the oppositional roles of the *E. coli* MMR and SOS systems in controlling both rates of mutation and interspecies recombination [131, 81]. Given the results of my research, weak mutator genes, HGT, and interspecies recombinant mechanisms work together to arrive at MDR bacterial phenotypes in response to the environmental stress caused by anthropogenic antimicrobials.

5.5 Intellectual Merit

My dissertation validates ML techniques to preprocess genetic data for explainable results. The feature engineering techniques demonstrated here should apply when supplementing phylogenetic analyses. This research shows that annotation of simple gene presence or absence predicts specific classes of AMR resulting in different MDR phenotypes. The traits that drive MDR phenotypes are polygenic. To identify the general principles behind the genetic mechanisms that proliferate resistance a sufficiently large data set, like the global survey data used here, is required.

The main conclusion of this research program is that the subversion of the MDR pandemic requires tracking and predicting the biological mechanisms of proliferation, instead of the capacity for hypermutation or the AMR genes themselves. When adapting into increasingly drug resistant phenotypes, *E. coli* bacteria prioritize a genetic capacity for transmitting genes instead of evolving novel resistance. As I demonstrate in Chapter 2, LeClerc's traditionally accepted definition of a hypermutator phenotype as a repair deficient *mutS* gene [37], shows little to no support for a dependent evolutionary correlation between variant positions in the MutS protein and the MDR trait. Further, in Chapter 3, when training random forests to classify resistance, I found that neither *mutS* nor its paralog occur in the feature clusters (***cluster yfjQ_ 4*** and ***ydaM***) that best predict resistance to five different types of antimicrobial compounds. To ensure that the best predictors from Chapter 3 were biologically relevant, I investigated three select predictive features, ***clusters yfjQ_ 4***, ***ydaM***, and ***cbeA_ 1*** in Chapter 4. All three clusters contain HGT-associated genes, not AMR genes or MMR genes like *mutS*. My research confirms that the mechanisms for developing novel resistance traits and even the resistance genes themselves are not as useful for predicting current AMR. Shifting focus from the mechanisms of resistance to the mechanisms of transmission allows researchers to better address the AMR pandemic [180].

Currently available data sets processed with machine learning techniques identify the genetic determinants of MDR. Investigating these determinants shall lead to novel drugs, new treatment strategies, and more efficient use of therapeutics. For example in the Ford et al., 2020 paper, we noted that AMR genes of various classes evolved under stabilizing selection (gained and lost resistance traits) or genetic capitalism (increasing accumulation of resistance) over time [5]. This dissertation expands on those results by identifying what I call gene sets, demonstrated in this manuscript as features of clustered genes that best predict MDR phenotypes. Those gene sets describe specific genetic annotations whose presence or absence indicates the mechanisms by which the pattern of genetic capitalism operates. The feature cluster gene sets identified by the models in Chapter 4 demonstrate how processed survey data predicts the characteristics of potentially epidemic MDR strains before they begin to dominate local environmental niches. The research emphasis on the MDR pandemic, the end stage of AMR proliferation, must focus on genetic mechanisms of transmission (like HGT) and not mechanisms of development (like hypermutation).

5.6 Broader Impacts

As opposed to the results of Chapter 2, the findings from this investigation improve public health in a survey context. Sequencing costs have decreased over time, but sequencing technologies are not standard practice worldwide. A cost barrier to the implementation of personalized medicine often exists at the local level. Even in the United States, it is rare to sequence the resident microflora of a sick individual with NGS techniques. Health care providers still employ methods like cell culture, staining, microscopy techniques, and test kits when making diagnoses.

The ongoing COVID-19 pandemic demonstrated the effectiveness of over the counter and take home rapid antigen testing for the purposes of public health and safety. In this context, it becomes possible to theorize the development of kits for the detection of the precursors of MDR. Kits could be developed using primers for the genetic sequences found within clusters of high importance identified by my research. While PCR primer technology is not available for personal home use, the development of such a kit may prove useful when targeted to primary care physicians, hospital systems, or local health departments. Before prescribing antimicrobial drugs, primary healthcare providers could use a kit that incubates a sick patient's microflora with specific PCR primers. Those primers recognize the genetic sequences predictive of individual classes of MDR phenotype. Such hypothetical kits should improve patient health outcomes and lower healthcare costs, by encouraging more efficacious prescription practices.

5.7 Outcomes

As the COVID-19 pandemic demonstrates, effective actions must be employed at the national level [3]. To escape the silent pandemic of AMR, we should also develop programs and tools to help physicians make intelligent, informed decisions about antimicrobial prescribing [6]. Whole genome sequencing is a useful tool, but at the same time, concessions must be made at the local level because of current costs. It is not currently feasible to equip local clinics with full sequencing technologies. The implementation of test kits and the subsequent rise in their use during the COVID-19 pandemic demonstrates a potential appetite for local care. This inquiry did not develop a kit or assess the financial feasibility of such an idea, but a test kit for MDR is technologically possible. A nucleic acid amplification test (NAAT) of PCR primers for genes from different categories of resistance would enable local health departments or clinicians to make informed treatment decisions in response to the resistant strains found in a geographic area, essentially an antibiogram for an entire region.

The methods described here represent a potential improvement over susceptibility testing, in that the model-based method is predictive and not reactive. Modeling indicates the capacity of a strain to develop resistance to more classes of antimicrobial compounds. Phylogenetic investigation of high importance model features will yield hypotheses describing the stepwise acquisition of mutations that accumulate AMR genes into MDR phenotypes. By understanding the sets of genes that enable proliferation, we will anticipate and subvert the characteristics of strains that drive the silent pandemic.

REFERENCES

[1] Centers for Disease Control and Prevention (U.S.), "Antibiotic resistance threats in the united states, 2019," tech. rep., National Center for Emerging Zoonotic and Infectious Diseases (U.S.), Nov. 2019.

[2] Interagency Coordination Group on Antimicrobial Resistance, "No time to wait: Securing the future from drug-resistant infections." `https://www.who.int/publications/i/item/no-time-to-wait-securing-the-future-from-drug-resistant-infections`, Apr. 2019.

[3] O. Cars, S. J. Chandy, M. Mpundu, A. Q. Peralta, A. Zorzet, and A. D. So, "Resetting the agenda for antibiotic resistance through a health systems perspective," *Lancet Glob Health*, vol. 9, pp. e1022–e1027, July 2021.

[4] D. Chaudhry and P. Tomar, "Antimicrobial resistance: the next BIG pandemic," *Int J Community Med Public Health*, vol. 4, pp. 2632–2636, July 2017.

[5] C. T. Ford, G. L. Zenarosa, K. B. Smith, D. C. Brown, J. Williams, and D. Janies, "Genetic capitalism and stabilizing selection of antimicrobial resistance genotypes in escherichia coli," *Cladistics*, vol. 36, pp. 348–357, Aug. 2020.

[6] A. Gautam, "Antimicrobial resistance: The next probable pandemic," *JNMA J. Nepal Med. Assoc.*, vol. 60, pp. 225–228, Feb. 2022.

[7] R. Laxminarayan, "The overlooked pandemic of antimicrobial resistance," *Lancet*, vol. 399, pp. 606–607, Feb. 2022.

[8] E. W. Brown, R. Bell, G. Zhang, R. Timme, J. Zheng, T. S. Hammack, and M. W. Allard, "Salmonella genomics in public health and food safety," *EcoSal Plus*, p. eESP00082020, June 2021.

[9] J. M. Bello-López, O. A. Cabrero-Martínez, G. Ibáñez-Cervantes, C. Hernández-Cortez, L. I. Pelcastre-Rodríguez, L. U. Gonzalez-Avila, and G. Castro-Escarpulli, "Horizontal gene transfer and its association with antibiotic resistance in the genus aeromonas spp," *Microorganisms*, vol. 7, Sept. 2019.

[10] E. Broaders, C. G. M. Gahan, and J. R. Marchesi, "Mobile genetic elements of the human gastrointestinal tract: potential for spread of antibiotic resistance genes," *Gut Microbes*, vol. 4, pp. 271–280, July 2013.

[11] G. Gregova and V. Kmet, "Antibiotic resistance and virulence of escherichia coli strains isolated from animal rendering plant," *Sci. Rep.*, vol. 10, p. 17108, Oct. 2020.

[12] J. S. Iramiot, H. Kajumbula, J. Bazira, E. P. de Villiers, and B. B. Asiimwe, "Whole genome sequences of multi-drug resistant escherichia coli isolated in a pastoralist community of western uganda: Phylogenomic changes, virulence and resistant genes," *PLoS One*, vol. 15, p. e0231852, May 2020.

[13] K. Hamelin, G. Bruant, A. El-Shaarawi, S. Hill, T. A. Edge, J. Fairbrother, J. Harel, C. Maynard, L. Masson, and R. Brousseau, "Occurrence of virulence and antimicrobial resistance genes in escherichia coli isolates from different aquatic ecosystems within the st. clair river and detroit river areas," *Appl. Environ. Microbiol.*, vol. 73, pp. 477–484, Jan. 2007.

[14] E. Peterson and P. Kaur, "Antibiotic resistance mechanisms in bacteria: Relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens," *Front. Microbiol.*, vol. 9, p. 2928, Nov. 2018.

[15] A. R. Mahoney, M. M. Safaee, W. M. Wuest, and A. L. Furst, "The silent pandemic: Emergent antibiotic resistances following the global response to SARS-CoV-2," *iScience*, vol. 24, p. 102304, Apr. 2021.

[16] E. Denamur and I. Matic, "Evolution of mutation rates in bacteria," *Mol. Microbiol.*, vol. 60, pp. 820–827, May 2006.

[17] M. R. Gillings, "Evolutionary consequences of antibiotic use for the resistome, mobilome and microbial pangenome," *Front. Microbiol.*, vol. 4, p. 4, Jan. 2013.

[18] A. Khaledi, M. Schniederjans, S. Pohl, R. Rainer, U. Bodenhofer, B. Xia, F. Klawonn, S. Bruchmann, M. Preusse, D. Eckweiler, A. Dötsch, and S. Häussler, "Transcriptome profiling of antimicrobial resistance in pseudomonas aeruginosa," *Antimicrob. Agents Chemother.*, vol. 60, pp. 4722–4733, Aug. 2016.

[19] H. W. Stokes and M. R. Gillings, "Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into gram-negative pathogens," *FEMS Microbiol. Rev.*, vol. 35, pp. 790–819, Sept. 2011.

[20] E. Stubberfield, M. AbuOun, E. Sayers, H. M. O'Connor, R. M. Card, and M. F. Anjum, "Use of whole genome sequencing of commensal escherichia coli in pigs for antimicrobial resistance surveillance, united kingdom, 2018," *Euro Surveill.*, vol. 24, Dec. 2019.

[21] I. Martin-Loeches, A. Torres, M. Rinaudo, S. Terraneo, F. de Rosa, P. Ramirez, E. Diaz, L. Fernández-Barat, G. L. Li Bassi, and M. Ferrer, "Resistance patterns and outcomes in intensive care unit (ICU)-acquired pneumonia. validation of european centre for disease prevention and control (ECDC) and the centers for disease control and prevention (CDC) classification of multidrug resistant organisms," *J. Infect.*, vol. 70, pp. 213–222, Mar. 2015.

[22] M. P. Weinstein, *Performance Standards for Antimicrobial Susceptibility Testing.* Clinical and Laboratory Standards Institute, 31 ed., 2021.

[23] "Antibiotics tested by NARMS." `https://www.cdc.gov/narms/antibiotics-tested.html`, Apr. 2019. Accessed: 2023-3-14.

[24] J. Davies and D. Davies, "Origins and evolution of antibiotic resistance," *Microbiol. Mol. Biol. Rev.*, vol. 74, pp. 417–433, Sept. 2010.

[25] S. A. Ochoa, A. Cruz-Córdova, V. M. Luna-Pineda, J. P. Reyes-Grajeda, V. Cázares-Domínguez, G. Escalona, M. E. Sepúlveda-González, F. López-Montiel, J. Arellano-Galindo, B. López-Martínez, I. Parra-Ortega, S. Giono-Cerezo, R. Hernández-Castro, D. de la Rosa-Zamboni, and J. Xicohtencatl-Cortes, "Multidrug- and extensively Drug-Resistant uropathogenic escherichia coli clinical strains: Phylogenetic groups widely associated with integrons maintain high genetic diversity," *Front. Microbiol.*, vol. 7, p. 2042, Dec. 2016.

[26] H. A. Khan, A. Ahmad, and R. Mehboob, "Nosocomial infections and their control strategies," *Asian Pac. J. Trop. Biomed.*, vol. 5, pp. 509–514, July 2015.

[27] K. Inweregbu, J. Dave, and A. Pittard, "Nosocomial infections," *Contin Educ Anaesth Crit Care Pain*, vol. 5, pp. 14–17, Feb. 2005.

[28] D. R. Jenkins, "Nosocomial infections and infection control," *Medicine*, vol. 45, pp. 629–633, Oct. 2017.

[29] P. P. Khil, A. Dulanto Chiang, J. Ho, J.-H. Youn, J. K. Lemon, J. Gea-Banacloche, K. M. Frank, M. Parta, R. A. Bonomo, and J. P. Dekker, "Dynamic emergence of mismatch repair deficiency facilitates rapid evolution of Ceftazidime-Avibactam resistance in pseudomonas aeruginosa acute infection," *MBio*, vol. 10, Sept. 2019.

[30] H. H. Mehta, A. G. Prater, K. Beabout, R. A. L. Elworth, M. Karavis, H. S. Gibbons, and Y. Shamoo, "The essential role of hypermutation in rapid adaptation to antibiotic stress," *Antimicrob. Agents Chemother.*, vol. 63, July 2019.

[31] M. C. Orencia, J. S. Yoon, J. E. Ness, W. P. Stemmer, and R. C. Stevens, "Predicting the emergence of antibiotic resistance by directed evolution and structural analysis," *Nat. Struct. Biol.*, vol. 8, pp. 238–242, Mar. 2001.

[32] P. Dadgostar, "Antimicrobial resistance: Implications and costs," *Infect. Drug Resist.*, vol. 12, pp. 3903–3910, Dec. 2019.

[33] D. R. Evans, M. P. Griffith, A. J. Sundermann, K. A. Shutt, M. I. Saul, M. M. Mustapha, J. W. Marsh, V. S. Cooper, L. H. Harrison, and D. Van Tyne, "Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital," *Elife*, vol. 9, Apr. 2020.

[34] S. R. Partridge, S. M. Kwong, N. Firth, and S. O. Jensen, “Mobile genetic elements associated with antimicrobial resistance,” *Clin. Microbiol. Rev.*, vol. 31, Oct. 2018.

[35] R. A. Weingarten, R. C. Johnson, S. Conlan, A. M. Ramsburg, J. P. Dekker, A. F. Lau, P. Khil, R. T. Odom, C. Deming, M. Park, P. J. Thomas, NISC Comparative Sequencing Program, D. K. Henderson, T. N. Palmore, J. A. Segre, and K. M. Frank, “Genomic analysis of hospital plumbing reveals diverse reservoir of bacterial plasmids conferring carbapenem resistance,” *MBio*, vol. 9, Feb. 2018.

[36] F. Labat, O. Pradillon, L. Garry, M. Peuchmaur, B. Fantin, and E. Denamur, “Mutator phenotype confers advantage in escherichia coli chronic urinary tract infection pathogenesis,” *FEMS Immunol. Med. Microbiol.*, vol. 44, pp. 317–321, June 2005.

[37] J. E. LeClerc, B. Li, W. L. Payne, and T. A. Cebula, “High mutation frequencies among escherichia coli and salmonella pathogens,” *Science*, vol. 274, pp. 1208–1211, Nov. 1996.

[38] J. A. Eisen, “Horizontal gene transfer among microbial genomes: new insights from complete genome analysis,” *Curr. Opin. Genet. Dev.*, vol. 10, pp. 606–611, Dec. 2000.

[39] C. R. Woese, “On the evolution of cells,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, pp. 8742–8747, June 2002.

[40] H. Ochman, J. G. Lawrence, and E. A. Groisman, “Lateral gene transfer and the nature of bacterial innovation,” *Nature*, vol. 405, pp. 299–304, May 2000.

[41] S. Porwollik and M. McClelland, “Lateral gene transfer in salmonella,” *Microbes Infect.*, vol. 5, pp. 977–989, Sept. 2003.

[42] T. Wirth, D. Falush, R. Lan, F. Colles, P. Mensa, L. H. Wieler, H. Karch, P. R. Reeves, M. C. J. Maiden, H. Ochman, and M. Achtman, “Sex and virulence in escherichia coli: an evolutionary perspective,” *Mol. Microbiol.*, vol. 60, pp. 1136–1151, June 2006.

[43] B. Rusconi, F. Sanjar, S. S. K. Koenig, M. K. Mammel, P. I. Tarr, and M. Eppinger, “Whole genome sequencing for Genomics-Guided investigations of escherichia coli O157:H7 outbreaks,” *Front. Microbiol.*, vol. 7, p. 985, June 2016.

[44] M. Barlow, “What antimicrobial resistance has taught us about horizontal gene transfer,” *Methods Mol. Biol.*, vol. 532, pp. 397–411, 2009.

[45] M. J. Bottery, A. J. Wood, and M. A. Brockhurst, “Selective conditions for a multidrug resistance plasmid depend on the sociality of antibiotic resistance,” *Antimicrob. Agents Chemother.*, vol. 60, pp. 2524–2527, Apr. 2016.

[46] M. J. Bottery, *The Sociality and Evolution of Plasmid-Mediated Antimicrobial Resistance.* PhD thesis, University of York, Sept. 2017.

[47] H. Wang, L. Hou, Y. Liu, K. Liu, L. Zhang, F. Huang, L. Wang, A. Rashid, A. Hu, and C. Yu, "Horizontal and vertical gene transfer drive sediment antibiotic resistome in an urban lagoon system," *J. Environ. Sci.*, vol. 102, pp. 11–23, Apr. 2021.

[48] C. R. Linder, B. M. E. Moret, L. Nakhleh, and T. Warnow, "Network (reticulate) evolution: biology, models, and algorithms," in *The Ninth Pacific Symposium on Biocomputing (PSB)*, 2004.

[49] M.-E. Böhm, M. Razavi, N. P. Marathe, C.-F. Flach, and D. G. J. Larsson, "Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities," *Microbiome*, vol. 8, p. 41, Mar. 2020.

[50] C. O. Vrancianu, L. I. Popa, C. Bleotu, and M. C. Chifiriuc, "Targeting plasmids to limit acquisition and transmission of antimicrobial resistance," *Front. Microbiol.*, vol. 11, p. 761, May 2020.

[51] G. Alvarez-Uria, S. Gandra, S. Mandal, and R. Laxminarayan, "Global forecast of antimicrobial resistance in invasive isolates of escherichia coli and klebsiella pneumoniae," *Int. J. Infect. Dis.*, vol. 68, pp. 50–53, Mar. 2018.

[52] T. J. Johnson, C. M. Logue, J. R. Johnson, M. A. Kuskowski, J. S. Sherwood, H. J. Barnes, C. DebRoy, Y. M. Wannemuehler, M. Obata-Yasuoka, L. Spanjaard, and L. K. Nolan, "Associations between multidrug resistance, plasmid content, and virulence potential among extraintestinal pathogenic and commensal escherichia coli from humans and poultry," *Foodborne Pathog. Dis.*, vol. 9, pp. 37–46, Jan. 2012.

[53] F. Pérez-Rodríguez and B. Mercanoglu Taban, "A State-of-Art review on Multi-Drug resistant pathogens in foods of animal origin: Risk factors and mitigation strategies," *Front. Microbiol.*, vol. 10, p. 2091, Sept. 2019.

[54] R. A. Cook, W. B. Karesh, and S. A. Osofsky, "About us the manhattan principles," Sept. 2004.

[55] "Mission statement." `https://onehealthinitiative.com/mission-statement/`. Accessed: 2021-12-2.

[56] A. G. P. Ross, S. M. Crowe, and M. W. Tyndall, "Planning for the next global pandemic," *Int. J. Infect. Dis.*, vol. 38, pp. 89–94, Sept. 2015.

[57] M. W. Allard, E. Strain, D. Melka, K. Bunning, S. M. Musser, E. W. Brown, and R. Timme, "Practical value of food pathogen traceability through building a Whole-Genome sequencing network and database," *J. Clin. Microbiol.*, vol. 54, pp. 1975–1983, Aug. 2016.

[58] B. Brown, M. Allard, M. C. Bazaco, J. Blankenship, and T. Minor, "An economic evaluation of the whole genome sequencing source tracking program in the U.S," *PLoS One*, vol. 16, p. e0258262, Oct. 2021.

[59] M. AbuOun, H. M. O'Connor, E. J. Stubberfield, J. Nunez-Garcia, E. Sayers, D. W. Crook, R. P. Smith, and M. F. Anjum, "Characterizing antimicrobial resistant escherichia coli and associated risk factors in a Cross-Sectional study of pig farms in great britain," *Front. Microbiol.*, vol. 11, p. 861, May 2020.

[60] T. Ferenci, "What is driving the acquisition of muts and rpos polymorphisms in escherichia coli?," *Trends Microbiol.*, vol. 11, pp. 457–461, Oct. 2003.

[61] N. Frazão, A. Sousa, M. Lässig, and I. Gordo, "Horizontal gene transfer overrides mutation in escherichia coli colonizing the mammalian gut," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, pp. 17906–17915, Sept. 2019.

[62] R. Jayaraman, "Mutators and hypermutability in bacteria: the escherichia coli paradigm," *J. Genet.*, vol. 88, pp. 379–391, Dec. 2009.

[63] A. J. McAdam, "Enterobacteriaceae? enterobacterales? what should we call enteric Gram-Negative bacilli? a Micro-Comic strip," *J. Clin. Microbiol.*, vol. 58, Jan. 2020.

[64] R. P. Maharjan and T. Ferenci, "The impact of growth rate and environmental factors on mutation rates and spectra in escherichia coli," *Environ. Microbiol. Rep.*, vol. 10, pp. 626–633, Dec. 2018.

[65] C. J. Thorns, "Bacterial food-borne zoonoses," *Rev. Sci. Tech.*, vol. 19, pp. 226–239, Apr. 2000.

[66] A. E. Dewar, J. L. Thomas, T. W. Scott, G. Wild, A. S. Griffin, S. A. West, and M. Ghoul, "Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range," *Nat Ecol Evol*, Nov. 2021.

[67] O. Ehuwa, A. K. Jaiswal, and S. Jaiswal, "Salmonella, food safety and food handling practices," *Foods*, vol. 10, Apr. 2021.

[68] M. K. Fatica and K. R. Schneider, "Salmonella and produce: survival in the plant environment and implications in food safety," *Virulence*, vol. 2, pp. 573–579, Nov. 2011.

[69] M. K. Glynn, C. Bopp, W. Dewitt, P. Dabney, M. Mokhtar, and F. J. Angulo, "Emergence of multidrug-resistant salmonella enterica serotype typhimurium DT104 infections in the united states," *N. Engl. J. Med.*, vol. 338, pp. 1333–1338, May 1998.

[70] T. Humphrey, "Salmonella, stress responses and food safety," *Nat. Rev. Microbiol.*, vol. 2, pp. 504–509, June 2004.

[71] B. Rowe, L. R. Ward, and E. J. Threlfall, “Multidrug-resistant salmonella typhi: a worldwide epidemic,” *Clin. Infect. Dis.*, vol. 24 Suppl 1, pp. S106–9, Jan. 1997.

[72] I. L. Brito, “Examining horizontal gene transfer in microbial communities,” *Nat. Rev. Microbiol.*, vol. 19, pp. 442–453, July 2021.

[73] G. Cabot, L. Zamorano, B. Moyà, C. Juan, A. Navas, J. Blázquez, and A. Oliver, “Evolution of pseudomonas aeruginosa antimicrobial resistance and fitness under low and high mutation rates,” *Antimicrob. Agents Chemother.*, vol. 60, pp. 1767–1778, Jan. 2016.

[74] “Glossary of terms related to antibiotic resistance.” `https://www.cdc.gov/narms/resources/glossary.html`, Mar. 2019. Accessed: 2023-3-19.

[75] W. Hennig, “Phylogenetic systematics,” *Annu. Rev. Entomol.*, vol. 10, pp. 97–116, Jan. 1965.

[76] D. Janies, “Phylogenetic concepts and tools applied to epidemiologic investigations of infectious diseases,” *Microbiol Spectr*, vol. 7, July 2019.

[77] B. Henrichfreise, I. Wiegand, W. Pfister, and B. Wiedemann, “Resistance mechanisms of multiresistant pseudomonas aeruginosa strains from germany and correlation with hypermutation,” *Antimicrob. Agents Chemother.*, vol. 51, pp. 4062–4070, Nov. 2007.

[78] M. D. Maciá, D. Blanquer, B. Togores, J. Sauleda, J. L. Pérez, and A. Oliver, “Hypermutation is a key factor in development of multiple-antimicrobial resistance in pseudomonas aeruginosa strains causing chronic lung infections,” *Antimicrob. Agents Chemother.*, vol. 49, pp. 3382–3386, Aug. 2005.

[79] D. Brégeon, I. Matic, M. Radman, and F. Taddei, “Inefficient mismatch repair: genetic defects and down regulation,” *J. Genet.*, vol. 78, pp. 21–28, Apr. 1999.

[80] A. Giraud, I. Matic, O. Tenaillon, A. Clara, M. Radman, M. Fons, and F. Taddei, “Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut,” *Science*, vol. 291, pp. 2606–2608, Mar. 2001.

[81] I. Matic, C. Rayssiguier, and M. Radman, “Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species,” *Cell*, vol. 80, pp. 507–515, Feb. 1995.

[82] A.-L. Prunier and R. Leclercq, “Role of muts and mutl genes in hypermutability and recombination in staphylococcus aureus,” *J. Bacteriol.*, vol. 187, pp. 3455–3464, May 2005.

[83] N. Woodford and M. J. Ellington, “The emergence of antibiotic resistance by mutation,” *Clin. Microbiol. Infect.*, vol. 13, pp. 5–18, Jan. 2007.

[84] G. M. Eliopoulos and J. Blázquez, "Hypermutation as a factor contributing to the acquisition of antimicrobial resistance," *Clin. Infect. Dis.*, vol. 37, pp. 1201–1209, Nov. 2003.

[85] E. Denamur, S. Bonacorsi, A. Giraud, P. Duriez, F. Hilali, C. Amorin, E. Bingen, A. Andremont, B. Picard, F. Taddei, and I. Matic, "High frequency of mutator strains among human uropathogenic escherichia coli isolates," *J. Bacteriol.*, vol. 184, pp. 605–609, Jan. 2002.

[86] R. J. Willems, J. Top, D. J. Smith, D. I. Roper, S. E. North, and N. Woodford, "Mutations in the DNA mismatch repair proteins MutS and MutL of oxazolidinone-resistant or -susceptible enterococcus faecium," *Antimicrob. Agents Chemother.*, vol. 47, pp. 3061–3066, Oct. 2003.

[87] A. Jolivet-Gougeon, B. Kovacs, S. Le Gall-David, H. Le Bars, L. Bousarghin, M. Bonnaure-Mallet, B. Lobel, F. Guillé, C.-J. Soussy, and P. Tenke, "Bacterial hypermutation: clinical implications," *J. Med. Microbiol.*, vol. 60, pp. 563–573, May 2011.

[88] M. W. Allard and E. W. Brown, "Epidemiology needs more interdisciplinary teams with expertise in molecular systematics, public health and food safety," *Cladistics*, vol. 36, pp. 345–347, Aug. 2020.

[89] K. M. Osman, A. D. Kappell, M. Elhadidy, F. ElMougy, W. A. A. El-Ghany, A. Orabi, A. S. Mubarak, T. M. Dawoud, H. A. Hemeg, I. M. I. Moussa, A. M. Hessain, and H. M. Y. Yousef, "Poultry hatcheries as potential reservoirs for antimicrobial-resistant escherichia coli: A risk to public health and food safety," *Sci. Rep.*, vol. 8, p. 5859, Apr. 2018.

[90] B. Tolar, L. A. Joseph, M. N. Schroeder, S. Stroika, E. M. Ribot, K. B. Hise, and P. Gerner-Smidt, "An overview of PulseNet USA databases," *Foodborne Pathog. Dis.*, vol. 16, pp. 457–462, July 2019.

[91] "Methods description for ftp://ncbi.nlm.nih.gov/pathogen/," Jan. 2017.

[92] "README for NCBI pathogen detection," Dec. 2021.

[93] "Transcript for VitalSigns teleconference: Antibiotic resistant germs." `https://www.cdc.gov/media/releases/2018/t0403-antibiotic-resistant-germs.html`, Apr. 2019. Accessed: 2023-3-19.

[94] A. Montana, A. Dekhtyar, E. Neal, M. Black, and C. Kitts, "Chronology-Sensitive hierarchical clustering of pyrosequenced DNA samples of e. coli: A case study," in *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 155–159, Nov. 2011.

[95] A. Barlaam, A. Parisi, E. Spinelli, M. Caruso, P. D. I. Taranto, and G. Normanno, "Global emergence of Colistin-Resistant escherichia coli in food chains

and associated food safety implications: A review," *J. Food Prot.*, vol. 82, pp. 1440–1448, Aug. 2019.

[96] F. Baquero, "From pieces to patterns: evolutionary engineering in bacterial pathogens," *Nat. Rev. Microbiol.*, vol. 2, pp. 510–518, June 2004.

[97] F. Baquero, T. M. Coque, and R. Canton, "Antibiotics, complexity, and Evolution-Perspective-Antibiotic usage increases disorder at different biological levels, promoting the emergence of alternative orders in the microbiosphere," *ASM News-American Society for Microbiology*, vol. 69, no. 11, pp. 547–552, 2003.

[98] F. Baquero, A. P. Tedim, and T. M. Coque, "Antibiotic resistance shaping multi-level population biology of bacteria," *Front. Microbiol.*, vol. 4, p. 15, Mar. 2013.

[99] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual: (Python Documentation Manual Part 2).* CreateSpace Independent Publishing Platform, Mar. 2009.

[100] I. Anaconda, "Anaconda software distribution," 2020.

[101] S. R. Eddy, "Profile hidden markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[102] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, p. 119, Mar. 2010.

[103] F. A. Matsen, R. B. Kodner, and E. V. Armbrust, "pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree," *BMC Bioinformatics*, vol. 11, p. 538, Oct. 2010.

[104] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Res.*, vol. 25, pp. 1043–1055, July 2015.

[105] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, pp. 2068–2069, July 2014.

[106] A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J. A. Keane, and J. Parkhill, "Roary: rapid large-scale prokaryote pan genome analysis," *Bioinformatics*, vol. 31, pp. 3691–3693, Nov. 2015.

[107] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, pp. 1312–1313, May 2014.

[108] F. Lemoine and O. Gascuel, “Gotree/Goalign: toolkit and go API to facilitate the development of phylogenetic workflows,” *NAR Genom Bioinform*, vol. 3, p. lqab075, Sept. 2021.

[109] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, “Biopython: freely available python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, pp. 1422–1423, June 2009.

[110] B. P. Alcock, A. R. Raphenya, T. T. Y. Lau, K. K. Tsang, M. Bouchard, A. Edalatmand, W. Huynh, A.-L. V. Nguyen, A. A. Cheng, S. Liu, S. Y. Min, A. Miroshnichenko, H.-K. Tran, R. E. Werfalli, J. A. Nasir, M. Oloni, D. J. Speicher, A. Florescu, B. Singh, M. Faltyn, A. Hernandez-Koutoucheva, A. N. Sharma, E. Bordeleau, A. C. Pawlowski, H. L. Zubyk, D. Dooley, E. Griffiths, F. Maguire, G. L. Winsor, R. G. Beiko, F. S. L. Brinkman, W. W. L. Hsiao, G. V. Domselaar, and A. G. McArthur, “CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database,” *Nucleic Acids Res.*, vol. 48, pp. D517–D525, Jan. 2020.

[111] Microgenomics, “Tutorials/pangenome.md at master Â· microgenomics/tutorials,” Apr. 2018.

[112] M. L. McHugh, “The chi-square test of independence,” *Biochem. Med.*, vol. 23, no. 2, pp. 143–149, 2013.

[113] K. Katoh, K. Misawa, K.-I. Kuma, and T. Miyata, “MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform,” *Nucleic Acids Res.*, vol. 30, pp. 3059–3066, July 2002.

[114] A. J. Page, B. Taylor, A. J. Delaney, J. Soares, T. Seemann, J. A. Keane, and S. R. Harris, “SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments,” *Microb Genom*, vol. 2, p. e000056, Apr. 2016.

[115] W. P. Maddison and D. R. Maddison, “Mesquite: A modular system for evolutionary analysis,” 2021.

[116] M. Pagel, “Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters,” *Proc. Biol. Sci.*, vol. 255, no. 1342, pp. 37–45, 1994.

[117] L. J. Revell, “phytools: an R package for phylogenetic comparativebiology (and other things),” *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 217–223, 2012.

[118] R Core Team, “R: A language and environment for statistical computing,” 2021.

[119] D. H. Haft, M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O’Neill, W. Li, F. Chitsaz, M. K. Derbyshire, N. R. Gonzales, M. Gwadz, F. Lu, G. H.

Marchler, J. S. Song, N. Thanki, R. A. Yamashita, C. Zheng, F. Thibaud-Nissen, L. Y. Geer, A. Marchler-Bauer, and K. D. Pruitt, "RefSeq: an update on prokaryotic genome annotation and curation," *Nucleic Acids Res.*, vol. 46, pp. D851–D860, Jan. 2018.

[120] W. Li, K. R. O'Neill, D. H. Haft, M. DiCuccio, V. Chetvernin, A. Badretdin, G. Coulouris, F. Chitsaz, M. K. Derbyshire, A. S. Durkin, N. R. Gonzales, M. Gwadz, C. J. Lanczycki, J. S. Song, N. Thanki, J. Wang, R. A. Yamashita, M. Yang, C. Zheng, A. Marchler-Bauer, and F. Thibaud-Nissen, "RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation," *Nucleic Acids Res.*, vol. 49, pp. D1020–D1028, Jan. 2021.

[121] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell, "NCBI prokaryotic genome annotation pipeline," *Nucleic Acids Res.*, vol. 44, pp. 6614–6624, Aug. 2016.

[122] "Escherichia coli str. K-12 substr. MG1655 strain K-12 chromosome, comp - nucleotide - NCBI." `https://www.ncbi.nlm.nih.gov/nuccore/CP025268.1?from=2861895&amp;to=2861963`. Accessed: 2023-3-14.

[123] M. H. Lamers, A. Perrakis, J. H. Enzlin, H. H. Winterwerp, N. de Wind, and T. K. Sixma, "The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch," *Nature*, vol. 407, pp. 711–717, Oct. 2000.

[124] G. Obmolova, C. Ban, P. Hsieh, and W. Yang, "Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA," *Nature*, vol. 407, pp. 703–710, Oct. 2000.

[125] M. Kang, K. Kim, D. Choe, S. Cho, S. C. Kim, B. Palsson, and B.-K. Cho, "Inactivation of a Mismatch-Repair system diversifies genotypic landscape of escherichia coli during adaptive laboratory evolution," *Front. Microbiol.*, vol. 10, p. 1845, Aug. 2019.

[126] B. Li, H.-C. T. Tsui, J. E. LeClerc, M. Dey, M. E. Winkler, and T. A. Cebula, "Molecular analysis of muts expression and mutation in natural isolates of pathogenic escherichia coli," *Microbiology*, vol. 149, pp. 1323–1331, May 2003.

[127] C. Rayssiguier, D. S. Thaler, and M. Radman, "The barrier to recombination between escherichia coli and salmonella typhimurium is disrupted in mismatch-repair mutants," *Nature*, vol. 342, pp. 396–401, Nov. 1989.

[128] H. Sheng, J. Huang, Z. Han, M. Liu, Z. Lü, Q. Zhang, J. Zhang, J. Yang, S. Cui, and B. Yang, "Genes and proteomes associated with increased mutation frequency and multidrug resistance of naturally occurring mismatch Repair-Deficient salmonella hypermutators," *Front. Microbiol.*, vol. 11, p. 770, May 2020.

[129] M.-R. Baquero, A. I. Nilsson, M. d. C. Turrientes, D. Sandvang, J. C. Galán, J. L. Martínez, N. Frimodt-Møller, F. Baquero, and D. I. Andersson, "Polymorphic mutation frequencies in escherichia coli: emergence of weak mutators in clinical isolates," *J. Bacteriol.*, vol. 186, pp. 5538–5542, Aug. 2004.

[130] I. Matic, M. Radman, F. Taddei, B. Picard, C. Doit, E. Bingen, E. Denamur, and J. Elion, "Highly variable mutation rates in commensal and pathogenic escherichia coli," *Science*, vol. 277, pp. 1833–1834, Sept. 1997.

[131] R. C. Massey and A. Buckling, "Environmental regulation of mutation rates at specific sites," *Trends Microbiol.*, vol. 10, pp. 580–584, Dec. 2002.

[132] W. A. Rosche and P. L. Foster, "The role of transient hypermutators in adaptive mutation in escherichia coli," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, pp. 6862–6867, June 1999.

[133] R. M. Schaaper and M. Radman, "The extreme mutator effect of escherichia coli mutd5 results from saturation of mismatch repair by excessive DNA replication errors," *EMBO J.*, vol. 8, pp. 3511–3516, Nov. 1989.

[134] J. D. Tompkins, J. L. Nelson, J. C. Hazel, S. L. Leugers, J. D. Stumpf, and P. L. Foster, "Error-prone polymerase, DNA polymerase IV, is responsible for transient hypermutation during adaptive mutation in escherichia coli," *J. Bacteriol.*, vol. 185, pp. 3469–3472, June 2003.

[135] M. Lynch, M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster, "Genetic drift, selection and the evolution of the mutation rate," *Nat. Rev. Genet.*, vol. 17, pp. 704–714, Oct. 2016.

[136] M. Lynch and B. Trickovic, "A theoretical framework for evolutionary cell biology," *J. Mol. Biol.*, vol. 432, pp. 1861–1879, Mar. 2020.

[137] S. Wielgoss, J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R. E. Lenski, and D. Schneider, "Mutation rate inferred from synonymous substitutions in a Long-Term evolution experiment with escherichia coli," *G3*, vol. 1, pp. 183–186, Aug. 2011.

[138] S. Wielgoss, J. E. Barrick, O. Tenaillon, M. J. Wiser, W. J. Dittmar, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R. E. Lenski, and D. Schneider, "Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, pp. 222–227, Jan. 2013.

[139] A. Bohr and K. Memarzadeh, "Chapter 2 - the rise of artificial intelligence in healthcare applications," in *Artificial Intelligence in Healthcare* (A. Bohr and K. Memarzadeh, eds.), pp. 25–60, Academic Press, Jan. 2020.

[140] C. Krittanawong, "The rise of artificial intelligence and the uncertain future for physicians," *Eur. J. Intern. Med.*, vol. 48, pp. e13–e14, Feb. 2018.

[141] V. Patel and M. Shah, "Artificial intelligence and machine learning in drug discovery and development," *Intelligent Medicine*, vol. 2, pp. 134–140, Aug. 2022.

[142] Q. Waymel, S. Badr, X. Demondion, A. Cotten, and T. Jacques, "Impact of the rise of artificial intelligence in radiology: What do radiologists think?," *Diagn. Interv. Imaging*, vol. 100, pp. 327–336, June 2019.

[143] A. Yakimovich, "msphere of influence: the rise of artificial intelligence in infection biology," *mSphere*, vol. 4, June 2019.

[144] X. Deng, S. Cao, and A. L. Horn, "Emerging applications of machine learning in food safety," *Annu. Rev. Food Sci. Technol.*, vol. 12, pp. 513–538, Mar. 2021.

[145] A. L. Samuel, "Some studies in machine learning using the game of checkers. II—Recent progress," *IBM J. Res. Dev.*, vol. 11, pp. 601–617, Nov. 1967.

[146] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology," *Mol. Biosyst.*, vol. 5, pp. 1593–1605, Dec. 2009.

[147] A. L. Tarca, V. J. Carey, X.-W. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS Comput. Biol.*, vol. 3, p. e116, June 2007.

[148] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Mol. Syst. Biol.*, vol. 12, p. 878, July 2016.

[149] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[150] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees.* CRC Press, 2017.

[151] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[152] X. Deng, H. C. den Bakker, and R. S. Hendriksen, "Genomic epidemiology: Whole-Genome-Sequencing-Powered surveillance and outbreak investigation of foodborne bacterial pathogens," *Annu. Rev. Food Sci. Technol.*, vol. 7, pp. 353–374, Jan. 2016.

[153] S. Li, Y. He, D. A. Mann, and X. Deng, "Global spread of salmonella enteritidis via centralized sourcing and international trade of poultry breeding stocks," *Nat. Commun.*, vol. 12, p. 5109, Aug. 2021.

[154] S. Zhang, S. Li, W. Gu, H. den Bakker, D. Boxrud, A. Taylor, C. Roe, E. Driebe, D. M. Engelthaler, M. Allard, E. Brown, P. McDermott, S. Zhao, B. B. Bruce, E. Trees, P. I. Fields, and X. Deng, "Zoonotic source attribution of salmonella enterica serotype typhimurium using genomic surveillance data, united states," *Emerg. Infect. Dis.*, vol. 25, pp. 82–91, Jan. 2019.

[155] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning: Methods and Applications* (C. Zhang and Y. Ma, eds.), pp. 307–323, Boston, MA: Springer US, 2012.

[156] "User guide: contents." `https://scikit-learn.org/1.0/user_guide.html`. Accessed: 2021-12-2.

[157] "Sklearn.Ensemble.RandomForestClassifier." `https://scikit-learn.org/1.0/modules/generated/sklearn.ensemble.RandomForestClassifier.html`. Accessed: 2023-3-19.

[158] A. Drouin, S. Giguère, M. Déraspe, M. Marchand, M. Tyers, V. G. Loo, A.-M. Bourgault, F. Laviolette, and J. Corbeil, "Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons," *BMC Genomics*, vol. 17, p. 754, Sept. 2016.

[159] M. Galardini, "Roary/contrib/roary_plots at master Â· Sanger-Pathogens/Roary," 2015.

[160] "2.3. clustering." `https://scikit-learn.org/1.0/modules/clustering.html`. Accessed: 2021-12-2.

[161] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Am. Stat. Assoc.*, vol. 58, pp. 236–244, Mar. 1963.

[162] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: fundamental algorithms for scientific computing in python," *Nat. Methods*, vol. 17, pp. 261–272, Mar. 2020.

[163] E. A. Cummins, A. E. Snaith, A. McNally, and R. J. Hall, "The role of potentiating mutations in the evolution of pandemic escherichia coli clones," *Eur. J. Clin. Microbiol. Infect. Dis.*, Nov. 2021.

[164] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Comput. Surv.*, vol. 27, pp. 326–327, Sept. 1995.

[165] X. Ying, "An overview of overfitting and its solutions," *J. Phys. Conf. Ser.*, vol. 1168, p. 022022, Feb. 2019.

[166] T. J. Bradberry, “Trent-B/iterative-stratification: Scikit-learn cross validators for iterative stratification of multilabel data,” 2018.

[167] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[168] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” June 2011.

[169] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” June 2004.

[170] G. E. A. P. Batista, A. L. C. Bazzan, and M. C. Monard, “Balancing training data for automated annotation of keywords: a case study.”

[171] G. Menardi and N. Torelli, “Training and assessing classification rules with imbalanced data,” *Data Min. Knowl. Discov.*, vol. 28, pp. 92–122, Jan. 2014.

[172] Y. Che, Y. Yang, X. Xu, K. Břinda, M. F. Polz, W. P. Hanage, and T. Zhang, “Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, Feb. 2021.

[173] Y. He, X. Zhou, Z. Chen, X. Deng, A. Gehring, H. Ou, L. Zhang, and X. Shi, “PRAP: Pan resistome analysis pipeline,” *BMC Bioinformatics*, vol. 21, p. 20, Jan. 2020.

[174] E. Shoeb, U. Badar, J. Akhter, H. Shams, M. Sultana, and M. A. Ansari, “Horizontal gene transfer of stress resistance genes through plasmid transport,” *World J. Microbiol. Biotechnol.*, vol. 28, pp. 1021–1025, Mar. 2012.

[175] B. G. Spratt, L. D. Bowler, Q. Y. Zhang, J. Zhou, and J. M. Smith, “Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal neisseria species,” *J. Mol. Evol.*, vol. 34, pp. 115–125, Feb. 1992.

[176] X. Zhuge, Y. Ji, F. Tang, Y. Sun, M. Jiang, W. Hu, Y. Wu, F. Xue, J. Ren, W. Zhu, and J. Dai, “Population structure and antimicrobial resistance traits of avian-origin mcr-1-positive escherichia coli in eastern china, 2015 to 2017,” *Transbound. Emerg. Dis.*, vol. 66, pp. 1920–1929, Sept. 2019.

[177] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, “Manifold: A Model-Agnostic framework for interpretation and diagnosis of machine learning models,” *IEEE Trans. Vis. Comput. Graph.*, vol. 25, pp. 364–373, Jan. 2019.

[178] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, “BLAST+: architecture and applications,” *BMC Bioinformatics*, vol. 10, p. 421, Dec. 2009.

[179] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, and I. Karsch-Mizrachi, "NCBI taxonomy: a comprehensive update on curation, resources and tools," *Database*, vol. 2020, Jan. 2020.

[180] X. Liu, Y. Li, Y. Guo, Z. Zeng, B. Li, T. K. Wood, X. Cai, and X. Wang, "Physiological function of rac prophage during biofilm formation and regulation of rac excision in escherichia coli K-12," *Sci. Rep.*, vol. 5, p. 16074, Nov. 2015.

[181] V. W. C. Soo, P. Hanson-Manful, and W. M. Patrick, "Artificial gene amplification reveals an abundance of promiscuous resistance determinants in escherichia coli," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, pp. 1484–1489, Jan. 2011.

[182] I. Chopra, A. J. O'Neill, and K. Miller, "The role of mutators in the emergence of antibiotic-resistant bacteria," *Drug Resist. Updat.*, vol. 6, pp. 137–145, June 2003.

[183] J.-C. Galán, M.-C. Turrientes, M.-R. Baquero, M. Rodríguez-Alcayna, J. Martínez-Amado, J.-L. Martínez, and F. Baquero, "Mutation rate is reduced by increased dosage of mutl gene in escherichia coli K-12," *FEMS Microbiol. Lett.*, vol. 275, pp. 263–269, Oct. 2007.

[184] C. Witzany, R. R. Regoes, and C. Igler, "Assessing the relative importance of bacterial resistance, persistence and hyper-mutation for antibiotic treatment failure," *Proc. Biol. Sci.*, vol. 289, p. 20221300, Nov. 2022.

[185] B. J. Arnold, I.-T. Huang, and W. P. Hanage, "Horizontal gene transfer and adaptive evolution in bacteria," *Nat. Rev. Microbiol.*, vol. 20, pp. 206–218, Apr. 2022.

[186] H. Hasegawa, E. Suzuki, and S. Maeda, "Horizontal plasmid transfer by transformation in escherichia coli: Environmental factors and possible mechanisms," *Front. Microbiol.*, vol. 9, p. 2365, Oct. 2018.

[187] Y. Deng, H. Xu, Y. Su, S. Liu, L. Xu, Z. Guo, J. Wu, C. Cheng, and J. Feng, "Horizontal gene transfer contributes to virulence and antibiotic resistance of vibrio harveyi 345 based on complete genome sequence analysis," *BMC Genomics*, vol. 20, p. 761, Oct. 2019.

[188] J. R. Johnson and A. L. Stell, "Extended virulence genotypes of escherichia coli strains from patients with urosepsis in relation to phylogeny and host compromise," *J. Infect. Dis.*, vol. 181, pp. 261–272, Jan. 2000.

[189] D. S. Mohammad, "On the role of genomic islands in bacterial pathogenicity and antimicrobial resistance," *Cloud Publ. Int. J. Adv. Biotechnol. Bioeng*, vol. 2, pp. 18–36, 2014.

[190] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint, “Mobile genetic elements: the agents of open source evolution,” *Nat. Rev. Microbiol.*, vol. 3, pp. 722–732, Sept. 2005.

[191] B. T. Smith and G. C. Walker, “Mutagenesis and more: umuDC and the escherichia coli SOS response,” *Genetics*, vol. 148, pp. 1599–1610, Apr. 1998.

APPENDIX A: HARDWARE SPECIFICATIONS

Code development and testing took place on a desktop Mac provided by the UNC Charlotte Department of Bioinformatics and Genomics.

- iMac (27-inch, Late 2012)
  - Processor: 2.9 GHz Quad-Core Intel Core i5
  - Memory: 16 GB 1600 MHz DDR3

Research computing resources provided by the UNC Charlotte Research Computing group. I used the HPC cluster, which runs Red Hat Enterprise Linux and is managed by SLURM (https://oneit.charlotte.edu/urc/our-environment) [Internet]. Resource requests were for one or multiple of the following nodes:

- Dual 18-Core Intel Xeon Gold 6154 CPU @ 3.00GHz (36 cores / node)
- 388GB RAM (10.7GB / core)
- 100Gbit EDR Infiniband Interconnect

APPENDIX B: ANACONDA ENVIRONMENTS

Fourteen files (suffix ".yml") for creating the various Anaconda environments used in this research. Both the full list of dependencies for the UNC Charlotte UPC architecture and the historical requests (parameter `--from-history`) to create them are included.

- For snp-sites
  - 290B hpc_binf_snp-sites_environment_from_history.yml
  - 5.7K hpc_binf_snp-sites_environment_full.yml
- For CheckM
  - 213B hpc_checkm_environment_from_history.yml
  - 4.4K hpc_checkm_environment_full.yml
- Base Python 3 bioinformatics environment for data exploration and manipulation
  - 1.3K hpc_galick_gun_environment_from_history.yml
  - 2.6K hpc_galick_gun_environment_full.yml
- For Gotree
  - 160B hpc_gotree_environment_from_history.yml
  - 841B hpc_gotree_environment_full.yml
- For machine learning on imbalanced data with scikit-learn
  - 229B hpc_ml_imba_environment_from_history.yml
  - 5.2K hpc_ml_imba_environment_full.yml
- For Prokka and Roary

  - 18K hpc_prokka_environment_from_history.yml
  - 11K hpc_prokka_environment_full.yml

- For R
  - 300B hpc_r_for_vs_code_env_environment_from_history.yml
  - 8.6K hpc_r_for_vs_code_env_environment_full.yml

APPENDIX C: EXAMPLE RAxML SUBMISSION TO SLURM

Contains a single file (suffix “.slurm”) of a RAxML submission script to SLURM on the UNC Charlotte HPC architecture. Includes settings for SLURM resource requests and RAxML parameters.

- 5.1K hpc_raxml_settings.slurm

APPENDIX D: SUPPLEMENTARY DATA

Contains three files from this investigation: a phylogenetic tree (suffix ".tree") based on the core genome of 817 *E. coli* isolates as determined by Roary (without splitting paralogs) in Chapter 2, a spreadsheet (suffix ".csv") of gene annotations created by Roary (with split paralogs) in Chapter 3, and a spreadsheet (suffix ".tsv") of outputs from the `fitPagel` function in phytools.

- Phylogenetic tree from Chapter 2
  - 27K RAxML_rootedTree.900_ns_core_opt.ROOTED.tree
- Roary annotations from Chapter 3
  - 242M gene_presence_absence.csv
- `fitPagel` outputs from Chapter 2
  - 29K hpc_output_pagel_full_adjusted.tsv