

Phylogenetics and Machine Learning: Characterizing Antimicrobial Resistance in *Escherichia coli*



David Brown
Lab of Dr. Dan Janies
UNC Charlotte Department of Bioinformatics and Genomics

Second Committee Organizational Meeting

Purpose of Dissertation Investigations

To elucidate the evolutionary history and proliferation of global multidrug-resistance in *Escherichia coli*.

CHAPTER 1 - Introduction and Background

- Background

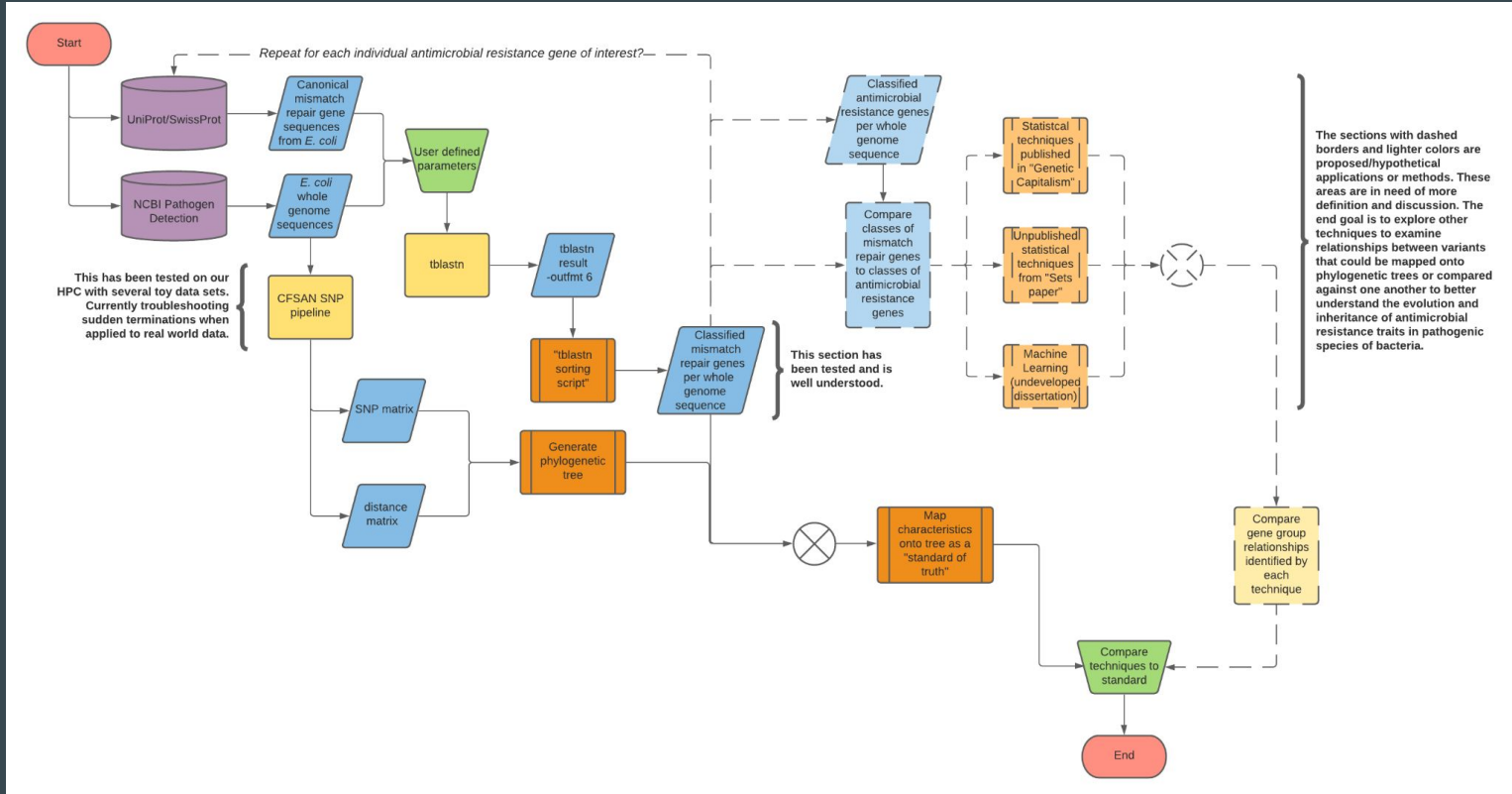
- Current silent pandemic of antibiotic resistance genes
- Circumstances are worsening due to:
 - Proliferation in number of resistance genes
 - Increased amount of resisted drug classes
 - Geographic spread
- Culminates in widespread multidrug-resistance
 - Defined as "resistant to at least one antibiotic in three or more drug classes."
 - <https://www.cdc.gov/narms/resources/glossary.html>

- Data Validity

- Food-associated bacteria are widely multidrug resistant
- Represents a diversity of sample sources
 - Globally
 - Clinical and/or environmental
- *Escherichia coli*
 - History as a model organism in biology
 - Current research methods focus on *Salmonella*

CLSI Class	Antimicrobial Agent
Aminoglycosides	Amikacin
	Gentamicin
	Kanamycin
	Streptomycin†
β-lactam combination agents	Amoxicillin-clavulanic acid
Cephems	Cefoxitin
	Ceftiofur
	Ceftriaxone‡
	Cephalothin
Folate pathway antagonists	Sulfamethoxazole
	Sulfisoxazole
	Trimethoprim-sulfamethoxazole
Macrolides	Azithromycin§
Penems	Meropenem
Penicillins	Ampicillin
Phenicol	Chloramphenicol
Quinolones	Ciprofloxacin**
	Nalidixic acid
Tetracyclines	Tetracycline

General Workflow Process



CHAPTER 2 - Mapping Traits to WGS Phylogeny

- Introduction

- Bacteria with deficient mismatch repair mechanisms have been described as having a "hypermutable phenotype". This highly adaptive phenotype is suggested to drive the acquisition of antimicrobial resistance. A main component of the methyl-directed mismatch repair system in bacteria is Mutator S.

- Research Questions

- What Mutator S variants are identified in the data?
- Which of those Mutator S variants are "hypermutable"?
- Are there correlations with hypermutable phenotypes and multidrug-resistance phenotypes?

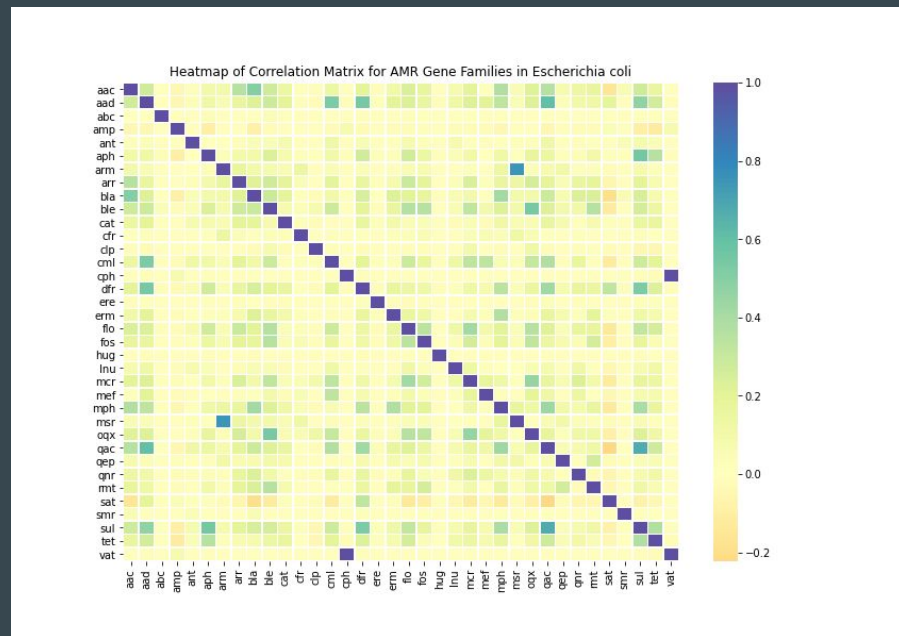
CHAPTER 2 - Mapping Traits to WGS Phylogeny - cont.

- Hypothesis

- Certain Mutator S variants and/or their lineages are correlated with specific, grouped classes of antimicrobial resistance genes.

- Methods

- Make a WGS SNP matrix
 - CFSAN SNP pipeline
- Build a phylogenetic tree from matrix
 - RAxML
 - TNT, PAUP
 - POY, Mesquite, R packages
- Overlay characters for:
 - Mutator S variants
 - UniProt
 - Multidrug-resistance
 - CDC, CARD



CHAPTER 3 - Functional Assessment of WGS for Prediction of Multidrug-resistant Phenotypes

- Introduction

- Similar data has been demonstrated to predict host species for *Salmonella* isolates. The US FDA has called for more research into improving predictive models based on WGS data. Multidrug-resistance phenotypes can be thought of as similar to host specificity, due to required underlying genes for host colonization, pathogenicity, and virulence.

- Research Questions

- Can machine learning algorithms predict multidrug-resistance from WGS data?
- What are the most important WGS SNPs?
- Are WGS SNPs indicative of compensatory mutations or are they limited to genes for resistance, pathogenicity, and virulence?


CHAPTER 3 - Functional Assessment of WGS for Prediction of Multidrug-resistant Phenotypes - cont.

- Hypothesis

- Certain patterns of WGS SNPs can both define and predict specific types of multidrug-resistance using supervised machine learning.

- Methods

- Same WGS SNP matrix
 - CFSAN SNP pipeline
- Same character classes
 - Multidrug-resistance
 - CDC, CARD
- Predict multidrug-resistance character from WGS SNP data (supervised machine learning)
 - sklearn multiclass classification algorithms
 - Random Forest, Naive Bayes, k-NN



DOMAIN 6 EVOLUTION AND GENOMICS

***Salmonella* Genomics in Public Health and Food Safety**

ERIC W. BROWN,^a REBECCA BELL,^a GUODONG ZHANG,^a
RUTH TIMME,^a JIE ZHENG,^a THOMAS S. HAMMACK,^a
AND MARC W. ALLARD^a

^aCenter for Food Safety and Applied Nutrition, U.S. Food and Drug Administration,
College Park, Maryland, USA

CHAPTER 4 - Combined Phylogenetic and Machine Learning Techniques for Predicting MDR Associations

- Introduction

- Large sample sizes and high genetic variation represent major challenges for big data genomics at scale. Reticulate evolution and horizontal gene transfer further complicate analyses, especially for highly similar samples. Techniques agnostic of evolutionary models could therefore be useful.

- Research Questions

- Can unsupervised machine learning "preprocess" genomic data into more computationally manageable clusters prior to phylogenetic analysis?
- Are the resulting clusters useful for phylogenetic inferences?
- Do the clusters represent functional (or phenetic) information that could associate multidrug-resistance with metadata like host, region, or sample type?

CHAPTER 4 - Combined Phylogenetic and Machine Learning Techniques for Predicting MDR Associations - cont.

- Hypothesis
 - Functional clusters (identified via unsupervised machine learning) can be used to clarify phylogenetic hypotheses while remaining agnostic of evolutionary models.
- Methods
 - Same WGS SNP matrix
 - CFSAN SNP pipeline
 - Same character classes
 - Multidrug-resistance
 - CDC, CARD
 - Create exclusive clusters from WGS SNP data and/or multidrug-resistance characters (unsupervised machine learning)
 - Hierarchical or distance-based clustering
 - sklearn
 - Build phylogenetic trees within clusters
 - TNT, PAUP



Significance

- **Intellectual Merit**
 - Refining current understandings of multidrug-resistance in enteric bacteria
 - Demonstrating newly available tools and data sets
 - CFSAN SNP Pipeline, NCBI Pathogen Detection
- **Broader Impacts**
 - Improvements for current disease surveillance
 - Insights for anticipating other pandemic zoonoses
 - Prolong the efficacy of current antibiotics

Questions

Thank you for your kind attention.