**Objectives**

*Prediction*

**Based on the AMR genotype, can we predict AMR phenotype?**

**Is there a characteristic of E. coli that we can use to predict the animal species of origin?**

To assess our ability to predict labels such as AMR phenotype or animal species of origin, we will use a multi output neural network model. Our choice of model is motivated by the large number of predictor variables relative to our sample size and our multiple output prediction space. Our predictor variables for the neural network are the presence-absence data for each gene across all identified genes in our dataset. Our response variable is a multi-output representation consisting of the gene resistance phenotype. We will split our data into training, testing, and validation sets at an 80:10:10 ratio respectively. A 5-fold cross validation scheme for parameter tuning.

*Inference*

**2. IS there a relationship between genotype and phenotype (yes/no)?**

Based on Feldgarden et al. 2018, a susceptibility test to measure the relationship between individual genes and resistance phenotype. Each test connected a single sample gene to its predicted resistance phenotype. A test returns a “positive” result if the resistance phenotype matches the predicted conferred resistance by a single gene. Performance is evaluated by the number of tests which correctly predict a positive (PPV) or negative (NPV) result.

**2a. What is the strength of the genotype / phenotype relationship?**

A mixed effects ordinal logistic regression model will be used to assess the strength of relationship between sample genotype and resistance phenotype. Models will only test for predicted relationships as presented by the Comprehensive Antimicrobial Resistance Database (CARD). The response variable is a 3-level ordinal measure of the sample’s resistance to an individual antimicrobial agent from Susceptible to Intermediate to Resistant (SIR), as determined by its recorded MIC values. The predictor variables are factors representing the presence or absence of all genes which are predicted to confer resistance to the response antimicrobial agent and are a part of the same gene functional family. By treating the predictor variables as factors, we aim to measure the effect of genes whose expression is dependent on a non-linear combination of genes working in tandem. For example, is resistance to ampicillin absolutely conferred by a single gene, is resistance a phenotype which increases when multiple genes are present, or is resistance conferred by several genes which must reach a certain threshold number to be expressed? In addition, we include animal origin as a fixed effect to account for differences in genotype-phenotype relationships between samples from different animal species. We evaluate the strength of relationship by reporting variables (genes) which significantly (alpha = 0.05) increase the log-odds of exhibiting the resistance phenotype.

**Based on the AMR gene data that we have, can we cluster profiles based on animal species of origin?**

We use a clustering approach to determine if there is an underlying structure to our sample genotype profile data. Each sample profile consists of gene presence absence across all unique antimicrobial resistance genes identified by AMR/resFinder in our dataset. Because our dataset is relative high dimension and only consists of binary variables, we will use a non-parametric k-modes clustering method. We evaluate clusters by comparing cluster labels to metadata labels using an adjusted Rand index. The adjusted Rand index is a measure from 0 to 1 of the similarity between two groups of clusters, while accounting for the cluster similarity by random chance. We hypothesize that cluster labels will have the highest agreement with animal species labels.

**Data**

**Methods**