RUTGERS

Bioinformatic Methods for Analyzing Mutations in Bacterial Genomes

Howard Fan

May 14, 2025

RUTGERS

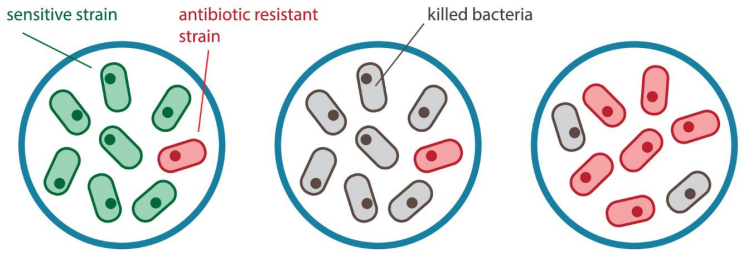
What is Mutation Analysis

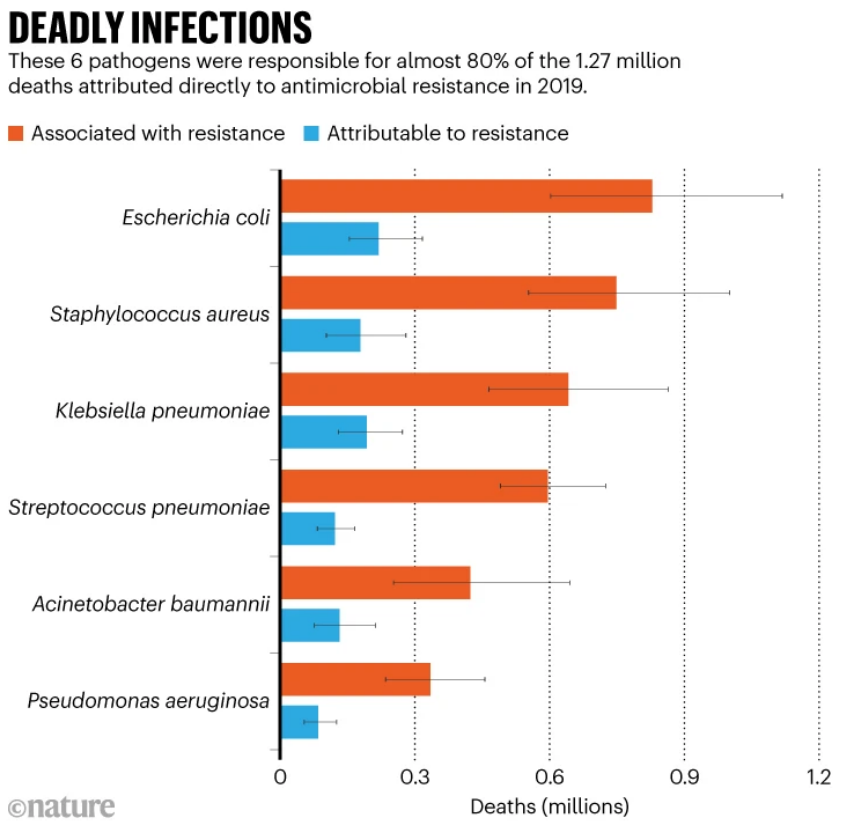
- Study of genetic alterations in genomes
- Focuses on point mutations, insertions, deletions, and structural changes
- Key to understanding evolution, antibiotic resistance, and pathogenicity

Why do we care about mutations in bacteria?

Mutations can cause bacteria to be more resistant to antibiotics, leading to:

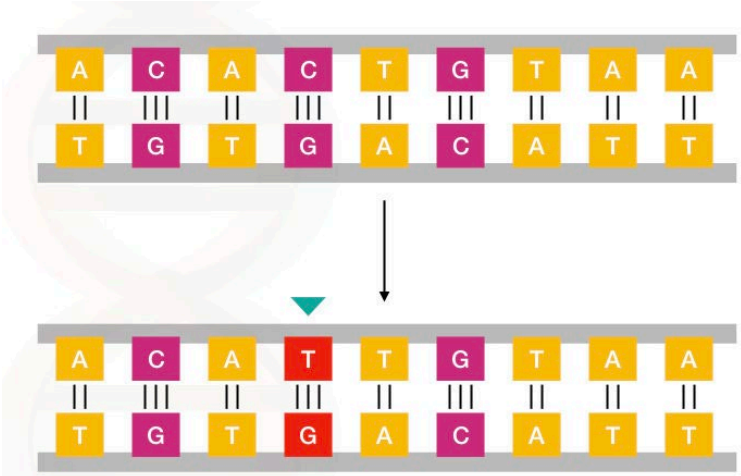
- **Reduced effectiveness of treatments**
  - Can lead to longer illnesses, more severe disease, and higher risk of death
- **Increased medical costs and burden**
  - Resistant infections often require more expensive or risky alternatives, longer hospital stays, and additional medical care.





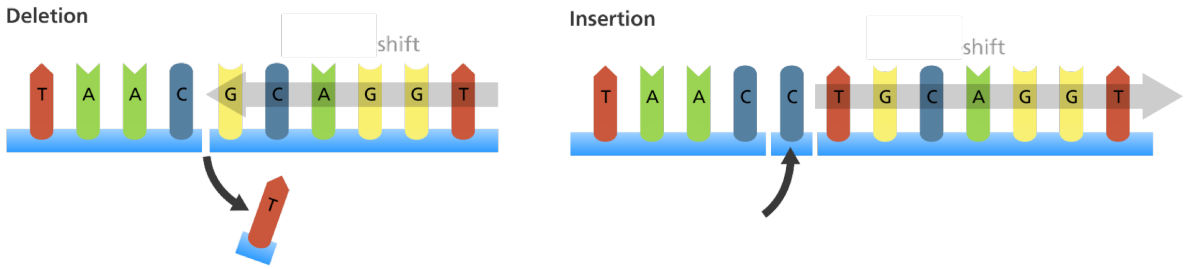
Types of Mutations

Single Nucleotide Polymorphisms



- Most common
- Also known as point mutations
- May be synonymous (silent) or non-synonymous (nonsense or missense)

Indels



- Insertion or deletion
- Often cause frameshifts if in coding regions

Identifying Variants from Bacterial Genomes

RUTGERS

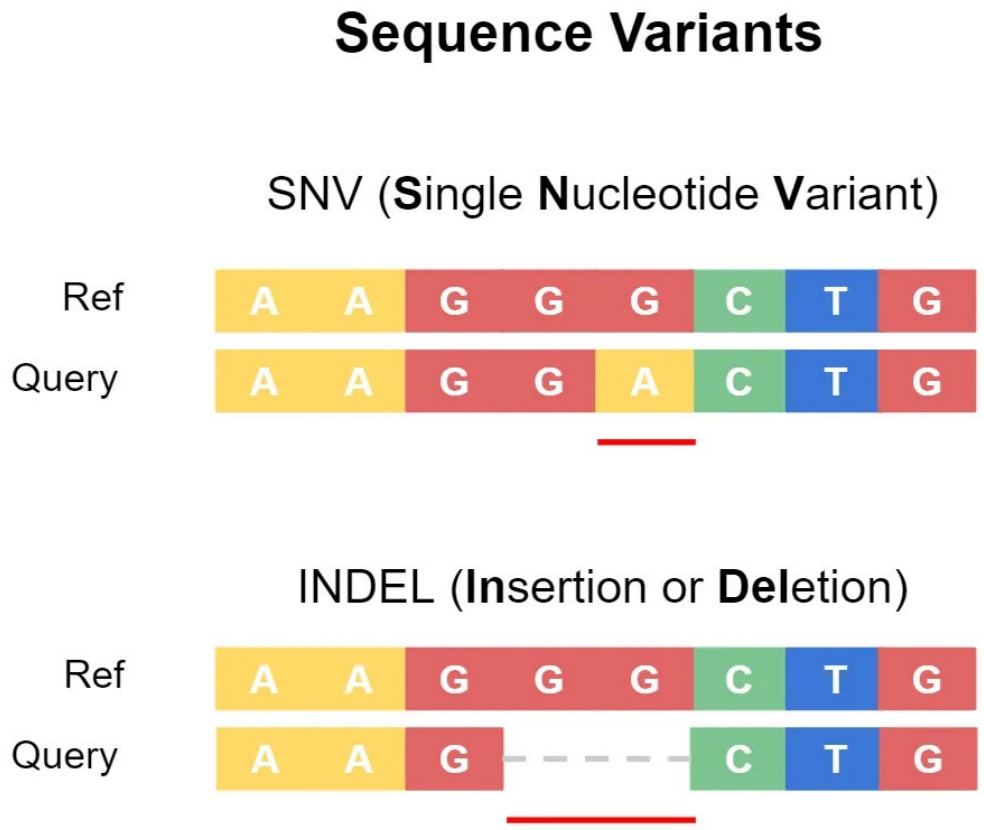
Bacterial Genome Workflow

1. Sample collection and DNA extraction
2. Sequencing (short-read vs long-read)
3. Quality control and preprocessing
4. Read alignment
5. Variant calling
6. Annotation
7. Downstream analysis

Variant Calling

We can detect variants from aligned sequences by comparing them to the reference sequence.

- SNP and Indel Callers
  - bcftools mpileup, GATK HaplotypeCaller
- Input requirements:
  - Aligned BAM files, reference genome
- Output:
  - VCF files listing variants, alleles, positions



Variant Filtering

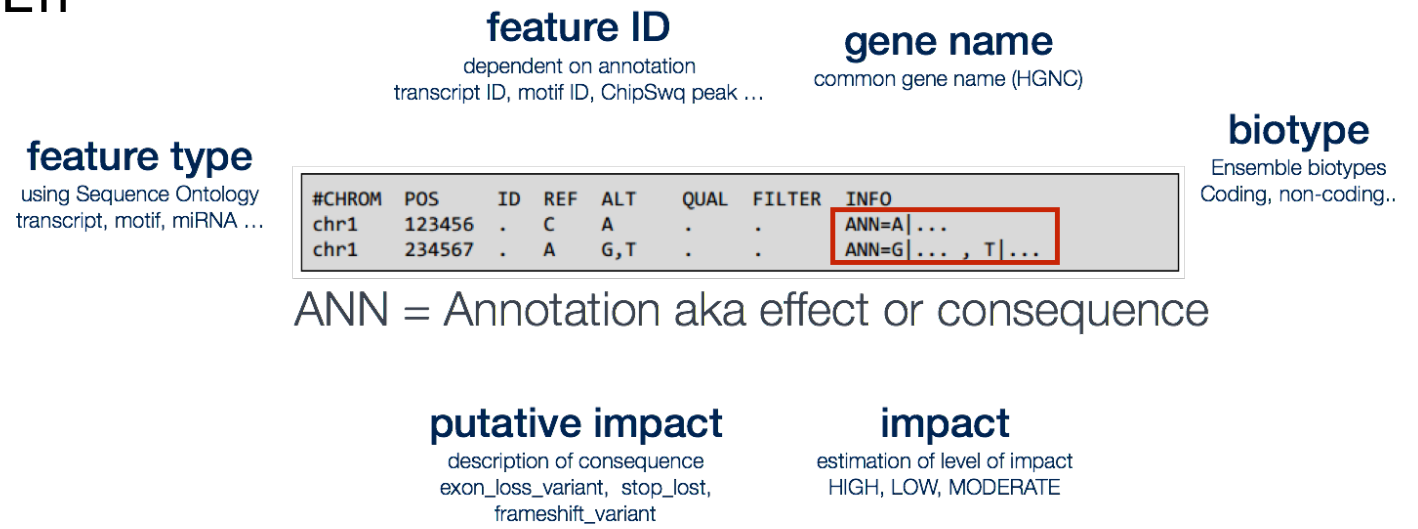
- Why filter?
  - Raw variant calls can include false positives
- Filtering strategies:
  - Depth of coverage (e.g., DP > 10)
  - Base quality (e.g., QUAL > 30)
  - Strand bias, allele frequency
- Tools:
  - bcftools filter, GATK VariantFiltration
  - Visual inspection with Integrative Genomics Viewer (IGV) if needed

RUTGERS

Variant Annotation

Variant annotation tools predict the functional impact of variants and add annotations such as gene names.

- Functional impacts:
  - Synonymous, nonsynonymous, frameshift
  - Premature stop codons, splice site mutations
- Tools: SnpEff



SnpEff

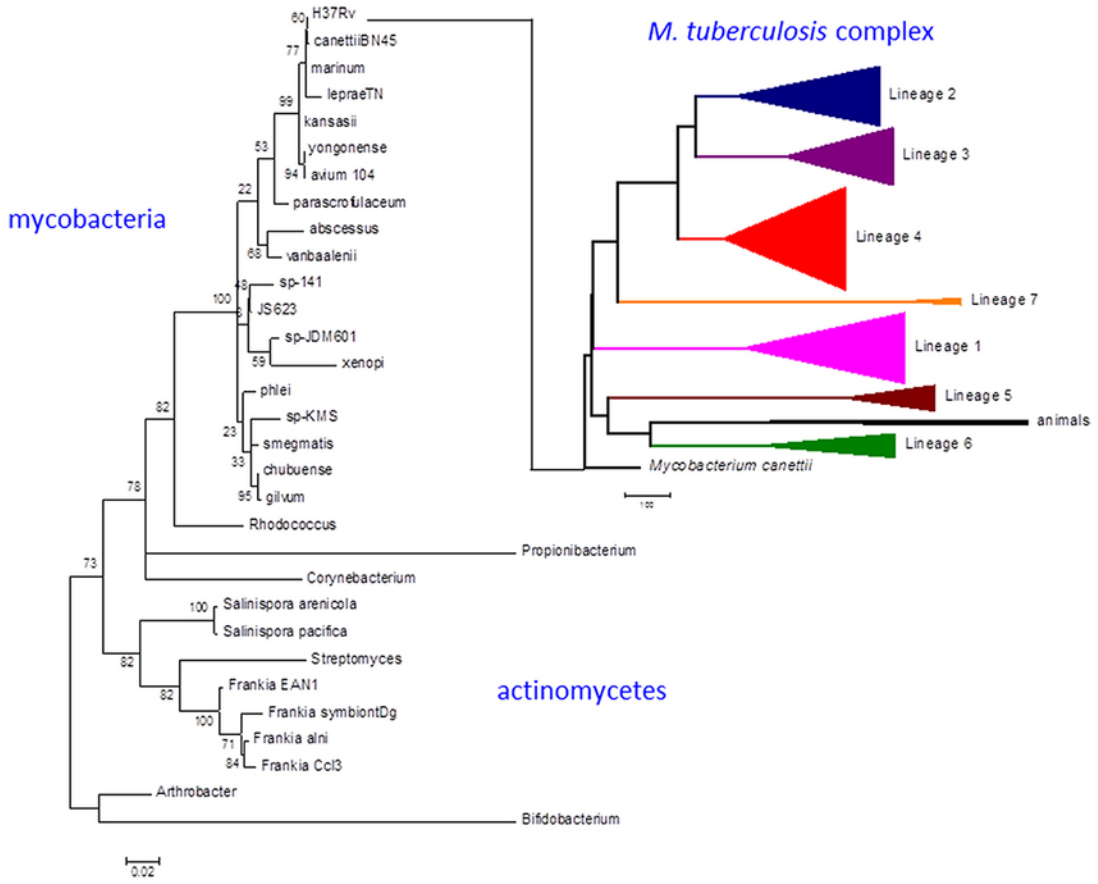
Genetic variant annotation and effect prediction toolbox.

Bacterial Phylogeny

Bacterial phylogeny provides insights on how bacteria evolves, spreads, and adapts.

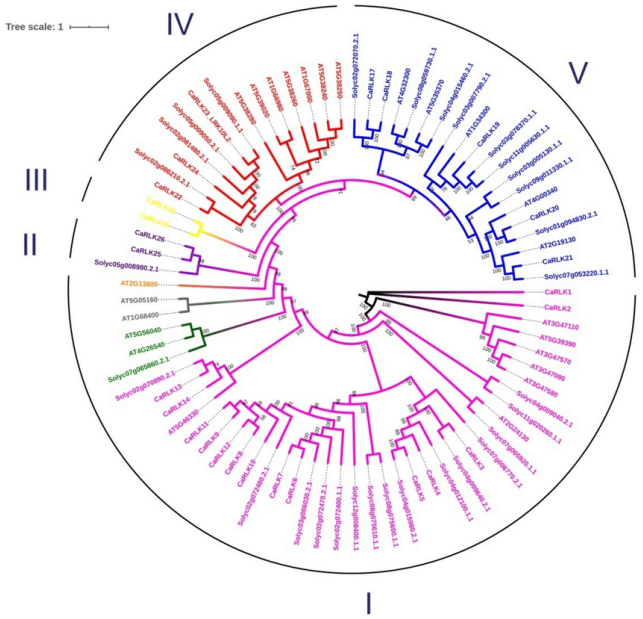
- For example, mycobacterium tuberculosis has 7 known lineages that vary in virulence and drug resistance.

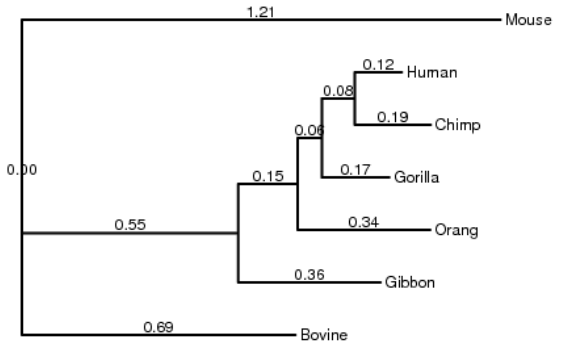
Hence, when studying bacterial variants, it is important to consider phylogeny.



Bacterial Phylogeny: Tree Building Tools

Tool	Purpose	Features
IQ-TREE	Maximum likelihood phylogenetic inference	Fast, model selection, bootstrapping
RAxML/FastTree	Tree building from alignments	FastTree for large trees, RAxML for complex models





Newick Format

Statistical and Bioinformatic Methods for Studying Genetic Variants

RUTGERS

Statistical Association Analysis

Regression models can be used to identify significant variants associated with traits of interest.

- For binary traits (e.g., disease presence, mortality):
  - Logistic regression

$$\log\left(\frac{P(trait = 1)}{P(trait = 0)}\right) = \beta_0 + \beta_1 \cdot SNP + \beta_2 \cdot Covariates$$

- For quantitative traits (e.g., weight, blood pressure, WBC count)
  - Linear regression

$$Trait = \beta_0 + \beta_1 \cdot SNP + \beta_2 \cdot Covariates + \epsilon$$

The results typically include a p-value and effect size.

Multiple Testing Correction

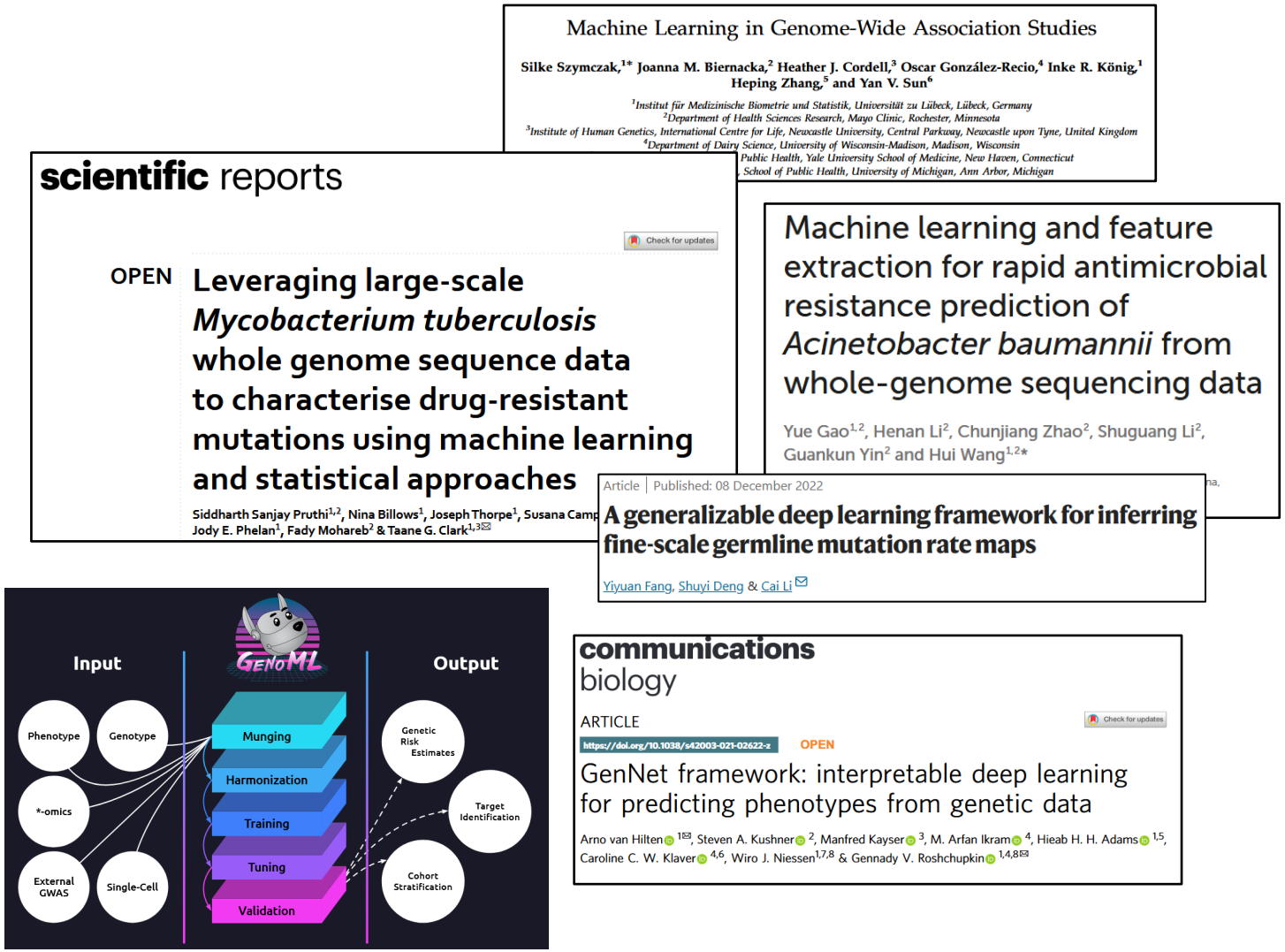
Because we are often testing many SNPs (possibly millions) at once, we need to control the false positive rate.

- Bonferroni Correction
  - Set p-value cut-off as $\alpha = \frac{0.05}{number\ of\ SNPs}$
- False Discovery Rate (FDR)
  - Benjamini-Hochberg procedure
  - Controls the expected proportion of false positives among all significant results
- Permutation Testing
  - Calculate empirical p-values based on phenotype permutation

Identifying Significant Variants: Bioinformatic Tools

Tool	Best For	Features
PLINK	General GWAS analysis	Fast; supports linear/logistic regression
GEMMA	Mixed models	Linear mixed models to correct for population structure and relatedness
SAIGE	Case-control studies with imbalance	Scalable generalized mixed models; controls for unbalanced case-control ratios and relatedness
RVTESTS	Rare variant analysis	Single variant and burden tests; useful for rare disease and sequencing data

Machine Learning in Genomic Studies



- Machine learning has become an essential tool in genomic studies due to its ability to process and extract patterns from massive, complex datasets.
- Open-source projects and data sharing have made incorporation of machine learning easier.

Model Training Workflow

1) Preparing the data

Sample	SNP1	SNP2	SNP3	SNP4	SNP5	...
DRR034340	1	0	0	0	0	...
DRR034341	0	0	0	1	0	...
DRR034342	0	1	0	0	0	...
DRR034343	1	0	0	1	0	...
...	...	...	...	...	...	...

51,229 samples x 9,643 SNPs

Test set (20%)

Training set (80%)

2) Training the model

Tune Parameters

- Random Forest
  - Max_depth, min_samples_leaf, min_samples_split, n_estimators
- Gradient-Boosted Trees
  - N_estimators, learning_rate, max_depth

Perform cross validation

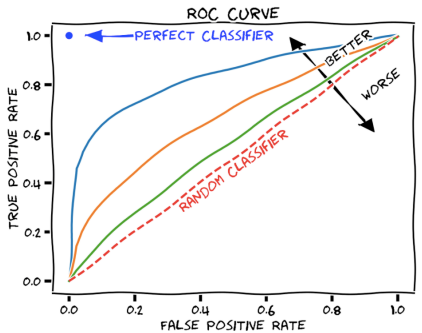
- Stratified K-Fold cross-validator
- Test model with multiple subsets of the training data using varying thresholds

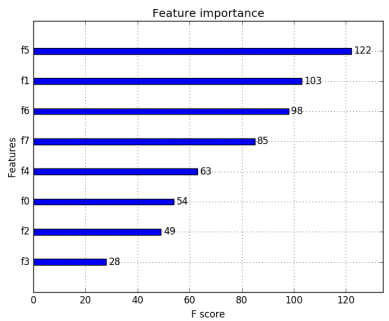
Find optimal threshold

- Test thresholds between 0.3 and 0.8
- Choose threshold with best accuracy score

Trained model

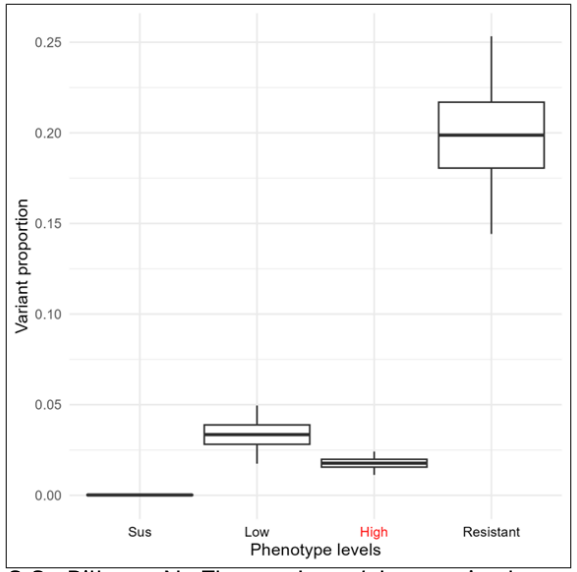
3) Evaluating the model

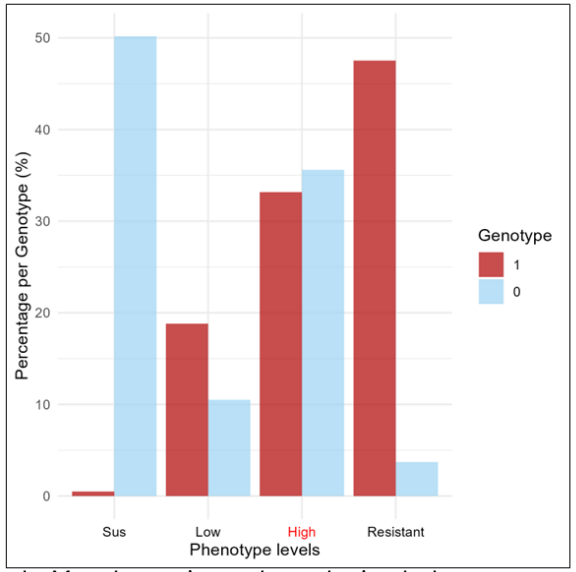


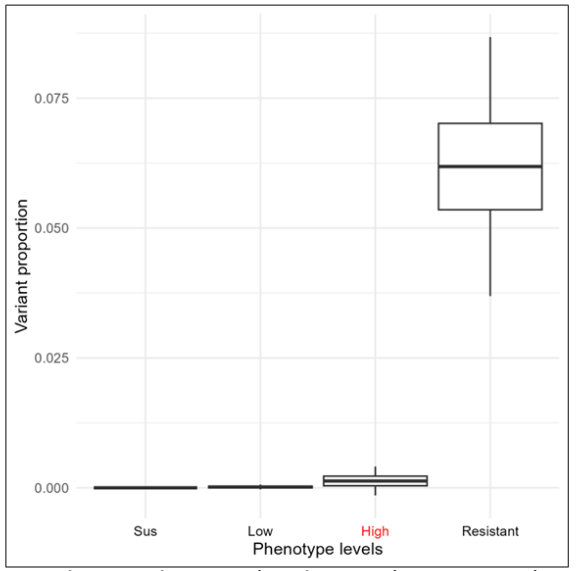


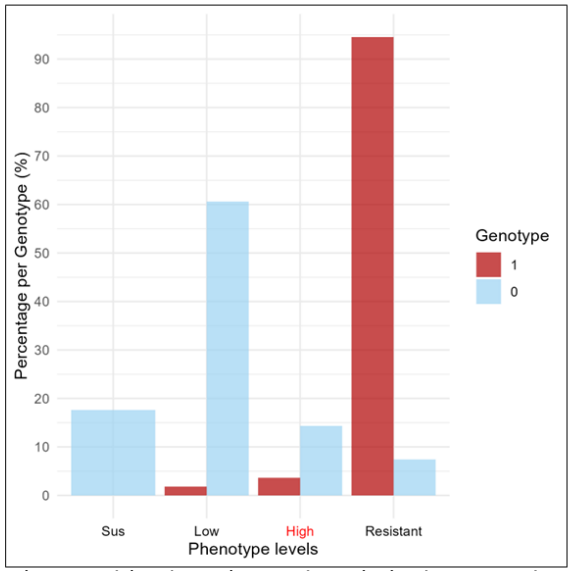
Genomic Associations with MIC Phenotypes

Drug	Position (gene)	Variant*	Suscept. Coeff.	Low Coeff.	Resist. Coeff.	Resist. *P*-value	Resist. RR
RIF	763,555 (*rpoB*)	230 C > T	− 0.358	0.474	1.913	$< 3 \times 10^{-7}$	6.773
INH	1,673,423 (*inhA*)	− 779G > T**	− 4.843	1.050	4.192	$< 10^{-6}$	66.168
EMB	4,248,002 (*embB*)	1489 C > A	− 4.584	− 1.930	2.164	$< 3 \times 10^{-10}$	8.707
AMK	1,473,246 (*Rv1313c*)	− 3741T > C **	− 6.422	− 4.975	2.320	$< 2 \times 10^{-14}$	10.174
KAN	1,473,246 (*Rv1313c*)	− 3741T > C **	− 3.684	− 3.063	3.915	$< 10^{-6}$	50.165
ETH	1,674,263 (*inhA*)	62T > C	− 14.794	0.380	3.457	$< 2 \times 10^{-6}$	31.708
LEV	7581 (*gyrA*)	280G > A	− 2.933	− 3.774	2.746	$< 10^{-6}$	15.575
MOX	7581 (*gyrA*)	280G > A	− 3.340	− 3.694	2.221	$< 10^{-6}$	9.215









Pruthi, S.S., Billows, N., Thorpe, J. *et al.* Leveraging large-scale *Mycobacterium tuberculosis* whole genome sequence data to characterise drug-resistant mutations using machine learning and statistical approaches. *Sci Rep* **14**, 27091 (2024). https://doi.org/10.1038/s41598-024-77947-w

RUTGERS

Summary

- Genetic variants in bacteria can result in antibiotic resistance, which can hinder clinical treatments and increase risk of morality.
- Variant calling (GATK, bcftools) and annotation (SnpEff) tools are used to identify and interpret genetic variants in query sequences.
- Statistical association tests, bioinformatic tools, and machine learning help to identify key variants associated with traits of interest.

Thank you for listening!