Industry Project

ML Informed Drug Resistance in MTB

Gauranga Kumar Baishya (MDS202325)
Siddhesh Maheshwari (MDS202347)

November 23, 2024

Outline

1. Introduction: TB
2. Basics of Genetics
3. Drug Resistance
4. Pipeline
5. Methodology
6. Results and Conclusion

Introduction: MTB

- **What is TB?**
    - Tuberculosis (TB) is an infectious disease caused by bacteria called *Mycobacterium tuberculosis*.
    - It mostly affects the lungs but can spread to other parts of the body.
- **How is TB spread?**
    - TB spreads through the air when an infected person coughs, sneezes, or talks.
- **Why is it necessary to understand TB?**
    - TB is one of the top 10 causes of death worldwide.
    - Drug-resistant TB is becoming a major challenge, making treatment harder.
    - Early detection and treatment can save lives and prevent its spread.
- **WHO's End TB Strategy**
    - The World Health Organization (WHO) aims to reduce TB deaths by 90% and cases by 80% by 2030.
    - This requires better diagnosis, treatment, and understanding of the disease.

TB Spread

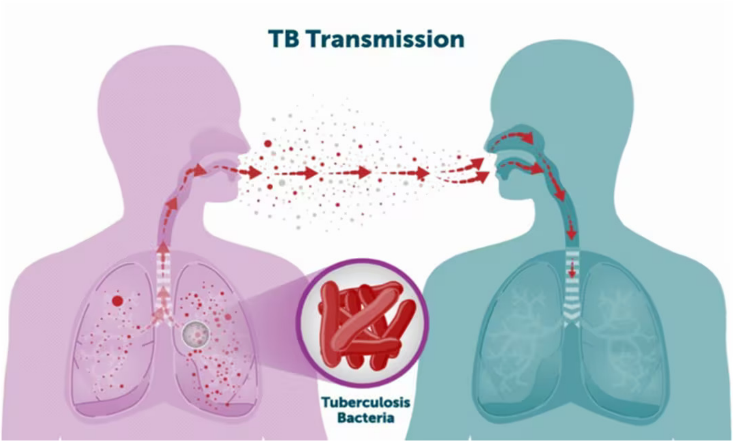
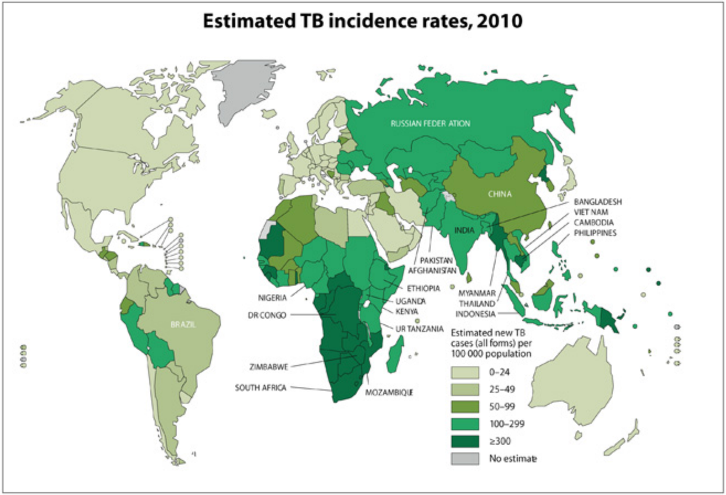


Image Credit: https://www.cdc.gov/tb/causes/index.html

TB Incidence



Estimated TB incidence rates, 2010

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

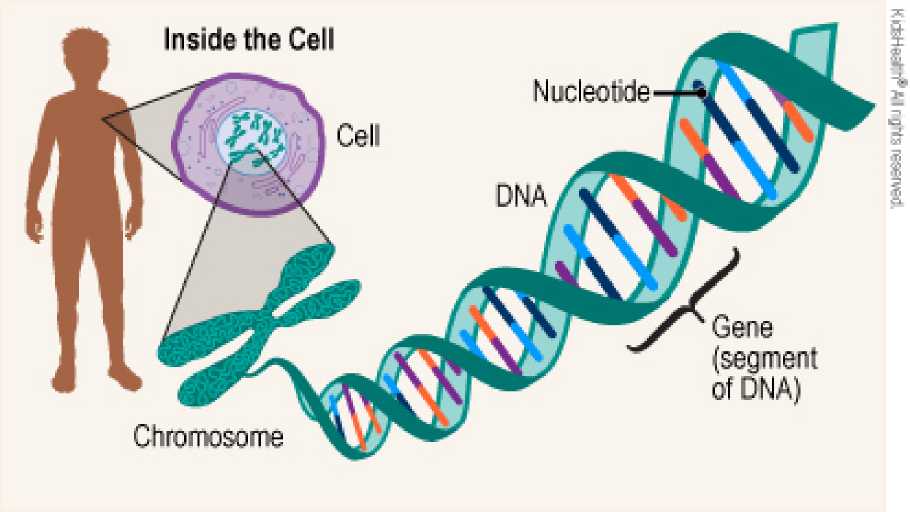
Source: *Global Tuberculosis Control 2011*. WHO, 2011.

World Health Organization

Basics of Genetics

- **DNA**
  - DNA (Deoxyribonucleic Acid) is the molecule that contains the genetic instructions for all living organisms.
  - It is like a code that tells cells how to grow, function, and reproduce.
- **Genes**
  - Genes are specific segments of DNA that carry the instructions for making proteins.
  - Proteins are essential for various functions in the body, such as building structures or fighting infections.

Basic Genetics



Basics of Genetics

- **Drug Resistance**
  - Drug resistance occurs when bacteria or other microbes evolve and develop the ability to survive despite the use of antibiotics.
  - This happens due to genetic mutations or acquiring resistance genes from other bacteria.
- **Common Antibiotics and Resistance**
  - First-line antibiotics are the primary drugs used to treat TB.
  - **Isoniazid**: A key first-line antibiotic for TB treatment
  - **Rifampicin**: Another first-line antibiotic
- **Example: Rifampicin Resistance and rpoB Gene**
  - Rifampicin is a key antibiotic for treating TB.
  - Resistance to rifampicin occurs due to mutations in the *rpoB* gene.
  - Mutations in *rpoB* (prevent rifampicin from binding effectively), allow the bacteria to survive.

Antimicrobial resistance with time

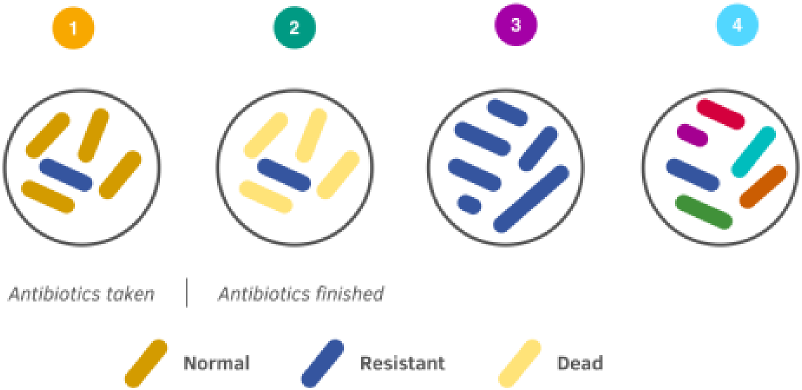
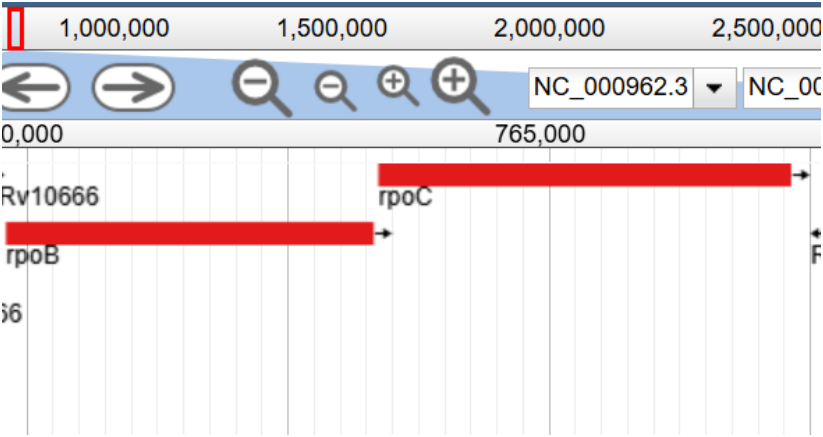


Image Credit: https://studymind.co.uk/notes/drug-resistance-antivirals-and-antiseptics/

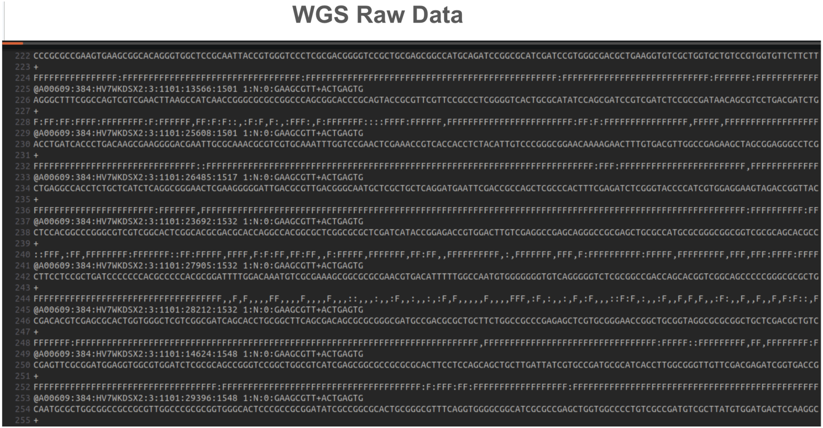
Compensatory Mutations

- **What are Compensatory Mutations?**
  - When bacteria develop resistance to antibiotics, some mutations can weaken their growth or survival.
  - Compensatory mutations help the bacteria recover from these negative effects, allowing them to survive better.
- **Example: Rifampicin Resistance and the rpoB Gene**
  - Mutations in the *rpoB* gene give resistance to rifampicin.
  - Compensatory mutations in *rpoC* genes restore bacteria survival while maintaining resistance!
- **Why are Compensatory Mutations Important?**
  - Understanding these mutations can help develop better treatment strategies to combat drug-resistant TB.

Compensatory Mutations: *rpoB*, *rpoC*



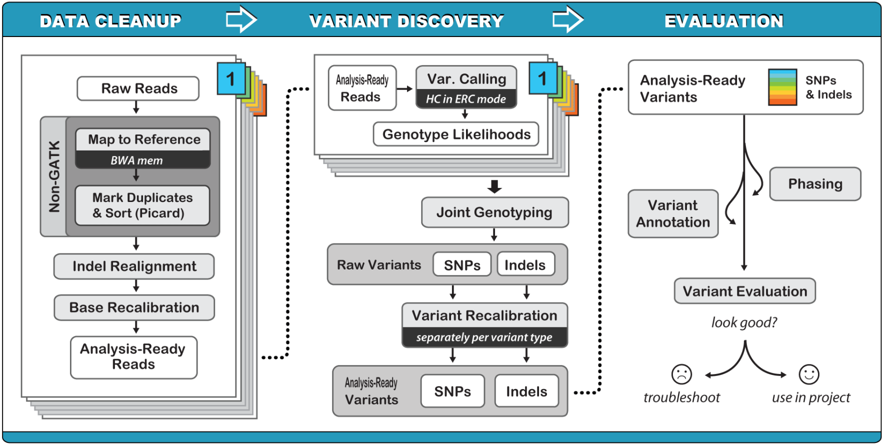
RAW Whole Genome Sequencing Data



WGS Raw Data

```
CCCGCGCCGAAGTGAAGCGGCACAGGGTGGCTCCGCAATTACCGTGGGTCCCTCGCGACGGGGTCCGCTGCGAGCGGCCATGCAGATCCGGCGCATCGATCCGTGGGCGACGCTGAAGGTGTCGCTGGTGCTGTCCGTGGTGTTCTTCTT
+
FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF:FFFFFFFFFFFF
@A00609:384:HV7WKDSX2:3:1101:13566:1501 1:N:0:GAAGCGTT+ACTGAGTG
AGGGCTTTCGGCCAGTCGTCGAACTTAAGCCATCAACCGGGCGCGCCGGCCCAGCGGCACCCGCAGTACCGCGTTCGTTCCGCCCTCGGGGTCACTGCGCATATCCAGCGATCCGTCGATCTCCGCCGATAACAGCGTCCTGACGATCTG
+
F:FF:FF:FFFF:FFFFFFFF:F:FFFFFF,FF:F:F::,:F:F,F:,:FFF:,F:FFFFFFF::::FFFF:FFFFFF,FFFFFFFFFFFFFFFFFFFFFFFF:FF:F:FFFFFFFFFFFFFFFF,FFFFF,FFFFFFFFFFFFFFFFFF
@A00609:384:HV7WKDSX2:3:1101:25608:1501 1:N:0:GAAGCGTT+ACTGAGTG
ACCTGATCACCCTGACAAGCGAAGGGGACGAATTGCGCAAACGCGTCGTGCAAATTTGGTCCGAACTCGAAACCGTCACCACCTCTACATTGTCCCGGGCGGAACAAAAGAACTTTGTGACGTTGGCCGAGAAGCTAGCGGAGGGCCTCG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF::FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF:FFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFF
@A00609:384:HV7WKDSX2:3:1101:26485:1517 1:N:0:GAAGCGTT+ACTGAGTG
CTGAGGCCACCTCTGCTCATCTCAGGCGGGAACTCGAAGGGGGATTGACGCGTTGACGGGCAATGCTCGCTGCTCAGGATGAATTCGACCGCCAGCTCGCCCACTTTCGAGATCTCGGGTACCCCATCGTGGAGGAAGTAGACCGGTTAC
+
FFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF:FF
@A00609:384:HV7WKDSX2:3:1101:23692:1532 1:N:0:GAAGCGTT+ACTGAGTG
CTCCACGGCCCGGGCGTCGTCGGCACTCGGCACGCGACGCACCAGGCCACGGCGCTCGGCGCGCTCGATCATACCGGAGACCGTGGACTTGTCGAGGCCGAGCAGGGCCGCGAGCTGCGCCATGCGCGGGCGGCGGTCGCGCAGCACGCC
+
::FFF,:FF,FFFFFFFF:FFFFFFF::FF:FFFFF,FFFF,F:F:FF,FF:FF,,F:FFFFF,FFFFFFF,FF:FF,,FFFFFFFFFF,:,FFFFFFF,FFF,F:FFFFFFFFFF:FFFFF,FFFFFFFFF,FFF,FFF:FFFF:FFFF
@A00609:384:HV7WKDSX2:3:1101:27905:1532 1:N:0:GAAGCGTT+ACTGAGTG
CTTCCTCCGCTGATCCCCCCCACGCCCCCACGCGGATTTTGGACAAATGTCGCGAAAGCGGCGCGCGAACGTGACATTTTTGGCCAATGTGGGGGGGTGTCAGGGGGTCTCGCGGCCGACCAGCACGGTCGGCAGCCCCCGGGCGCGCTG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,,F,F,,,,FF,,,,F,,,,F,,,::,,,:,,:F,,:,,:,:F,F,,,,,F,,,,FFF,:F,:,,:,F,:F,,,::F:F,:,,:F,,F,F,F,,:F:,,F,,F,,F,F:F::,F
@A00609:384:HV7WKDSX2:3:1101:28212:1532 1:N:0:GAAGCGTT+ACTGAGTG
CGACACGTCGAGCGCACTGGTGGGCTCGTCGGCGATCAGCACCTGCGGCTTCAGCGACAGCGCGCGGGCGATGCCGACGCGCTGCTTCTGGCCGCCCGAGAGCTCGTGCGGGAACCGGCTGCGGTAGGCGCGCGGCTGCTCGACGCTGTC
+
FFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF::FFFFFFFFF,FF,FFFFFFFF:F
@A00609:384:HV7WKDSX2:3:1101:14624:1548 1:N:0:GAAGCGTT+ACTGAGTG
CGAGTTCGCGGATGGAGGTGGCGTGGATCTCGCGCAGCCGGGTCCGGCTGGCGTCATCGAGCGGCGCCGCGCGCACTTCCTCCAGCAGCTGCTTGATTATCGTGCCGATGCGCATCACCTTGGCGGGTTGTTCGACGAGATCGGTGACCG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F:FFF:FF:FFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00609:384:HV7WKDSX2:3:1101:29396:1548 1:N:0:GAAGCGTT+ACTGAGTG
CAATGCGCTGGCGGCCGCCGCGTTGGCCCGCGCGGTGGGCACTCCCGCCGCGGATATCGCCGGCGCACTGCGGGCGTTTCAGGTGGGGCGGCATCGCGCCGAGCTGGTGGCCCCTGTCGCCGATGTCGCTTATGTGGATGACTCCAAGGC
+
```

Pipeline



DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
Raw Reads
Non-GATK
Map to Reference
BWA mem
Mark Duplicates & Sort (Picard)
Indel Realignment
Base Recalibration
Analysis-Ready Reads
1
Analysis-Ready Reads
Var. Calling
HC in ERC mode
Genotype Likelihoods
1
Joint Genotyping
Raw Variants
SNPs
Indels
Variant Recalibration
separately per variant type
Analysis-Ready Variants
SNPs
Indels
Analysis-Ready Variants
SNPs & Indels
Phasing
Variant Annotation
Variant Evaluation
look good?
☹
troubleshoot
☺
use in project

Mapping to Reference and Marking Duplicates

- **Mapping with bwa mem**
  - Aligns sequencing reads to a reference genome to find their matching locations.
  - This helps determine where each read originates in the genome.
- **Sorting and Converting**
  - Converts the alignment file to a compressed BAM format for storage and processing.
  - Then, sort the reads by their positions in the genome for easier analysis.
- **Marking Duplicates using Picard**
  - Identifies and removes duplicate reads created during PCR amplification.
  - This ensures the results are not biased by extra copies of the same read.

Base Recalibration and Analysis-Ready Reads

- **Base Recalibration**
  - Fixes errors in the quality scores assigned to each base by the sequencer (machine error).
  - Uses known reference variants to adjust and improve base quality accuracy.
  - Ensures that base quality scores better reflect the true confidence in the sequencing data.
- **Generating Analysis-Ready Reads**
  - After recalibration, creates a high-quality BAM file that is ready for analysis.
  - These reads are more accurate and reliable for downstream tasks like variant calling.

Variant Calling

- **Variant Calling Process with GATK**
  - Identifies SNPs (single nucleotide polymorphisms) and Indels (insertions and deletions) from recalibrated BAM files.
  - Generates a VCF (Variant Call Format) file containing the detected variants.
- **Refining Variants**
  - Filters variants to remove low-quality or false-positive calls.
  - Ensures the final dataset contains only high-confidence SNPs and Indels.
- **Analysis-Ready Variants**
  - Produces a final set of variants for downstream interpretation.
  - Facilitates accurate genome analysis and better biological insights.

Methodology

- Goal: Predict Interrelationship Between **rpoB** and **rpoC** Mutations

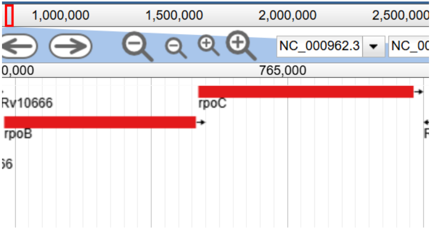
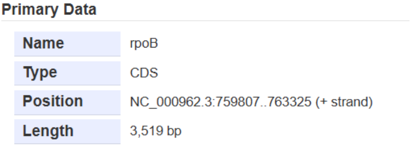


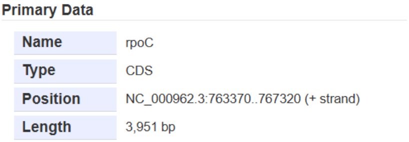
Figure: Position Of RpoB and RpoC in refrence Genome



Primary Data

Name	rpoB
Type	CDS
Position	NC_000962.3:759807..763325 (+ strand)
Length	3,519 bp

Figure: RpoB



Primary Data

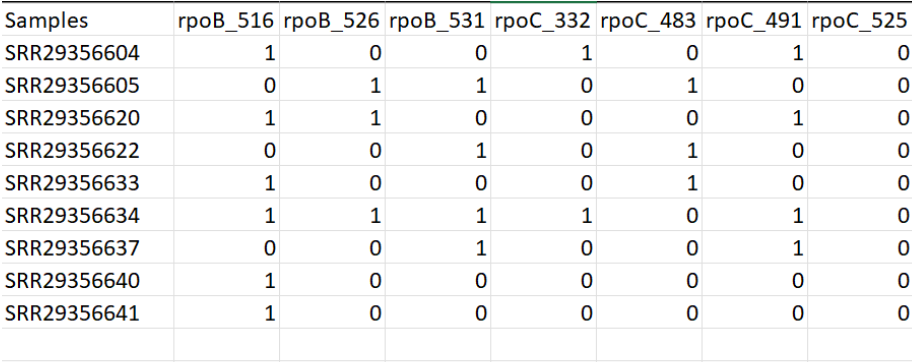
Name	rpoC
Type	CDS
Position	NC_000962.3:763370..767320 (+ strand)
Length	3,951 bp

Figure: RpoC

Methodology

- Goal: Predicting **rpoC** Mutations Based on **rpoB** Mutations.
- The mutations for which we need to predict relationship
  - rpoB_516 rpoB_526 rpoB_531
  - rpoC_332 rpoC_483 rpoC_491 rpoC_525
- Data collection
  - PHLTA, Israel (233)
  - Argentina (117)
- Preprocessing.
  - Ongoing.....

Modeling



Samples	rpoB_516	rpoB_526	rpoB_531	rpoC_332	rpoC_483	rpoC_491	rpoC_525
SRR29356604	1	0	0	1	0	1	0
SRR29356605	0	1	1	0	1	0	0
SRR29356620	1	1	0	0	0	1	0
SRR29356622	0	0	1	0	1	0	0
SRR29356633	1	0	0	0	1	0	0
SRR29356634	1	1	1	1	0	1	0
SRR29356637	0	0	1	0	0	1	0
SRR29356640	1	0	0	0	0	0	0
SRR29356641	1	0	0	0	0	0	0

Figure: Data Preprocessed

Modeling

- Models will be used for predicting resistance (tentative):
    - Association Rules
    - Logistic Regression
    - Random Forest
    - Gradient Boosting
    - Neural Networks.

Results and Conclusion

- On the way . . .