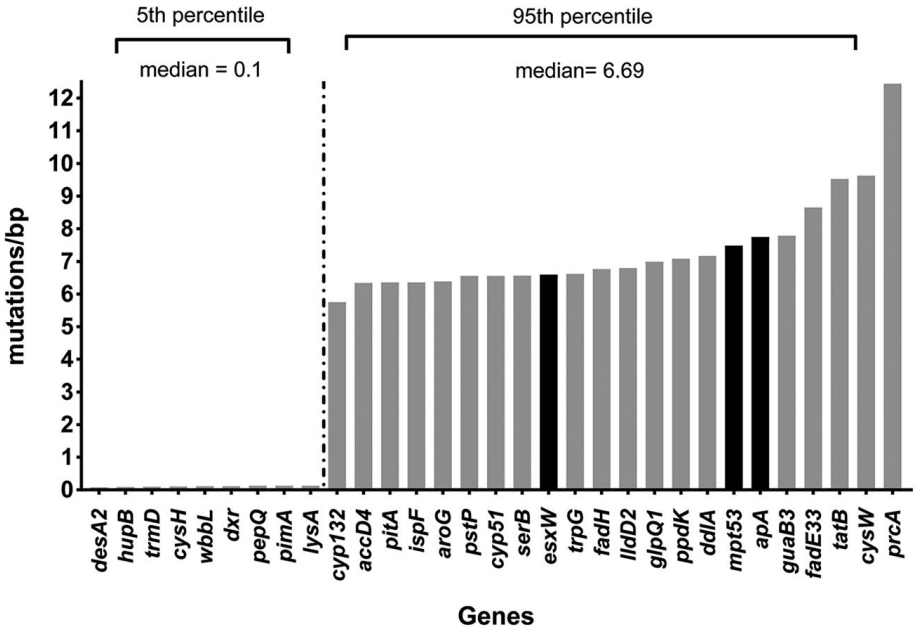
Intro to the SNP Project

Amelia Harrison

Presented: April 20, 2022

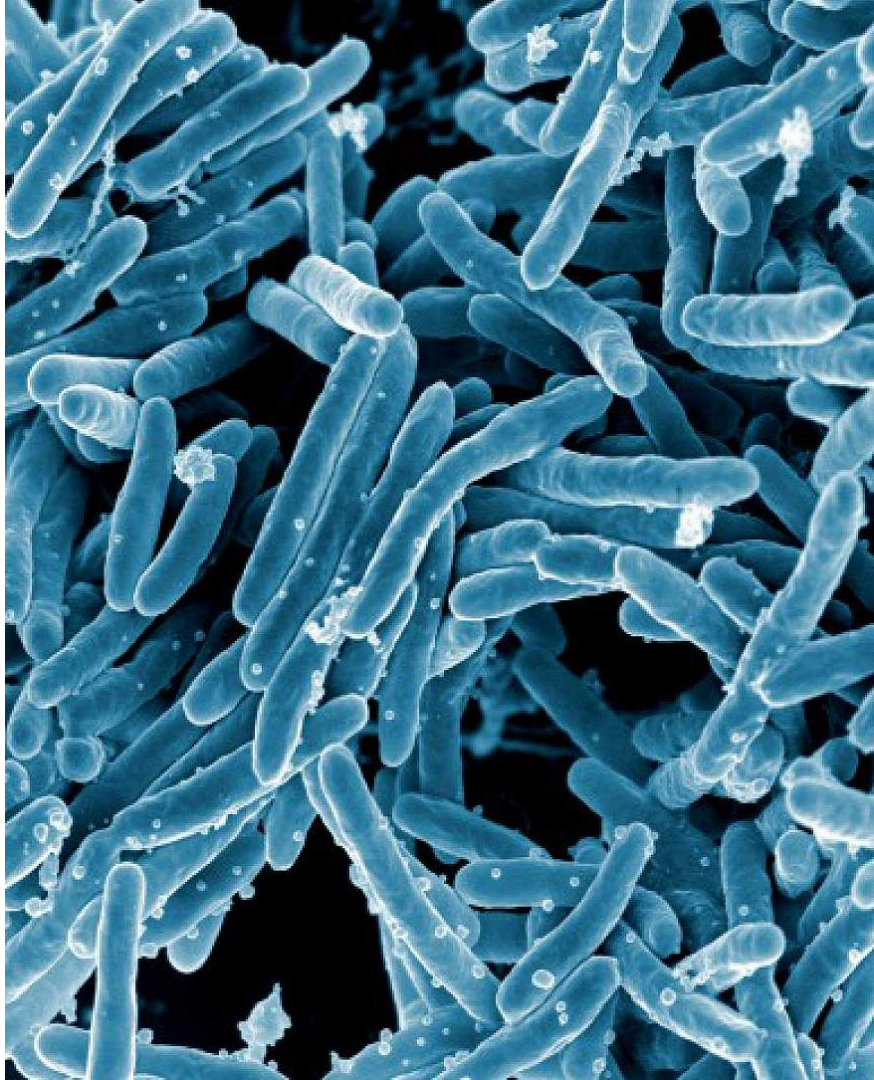
About the paper

- Mapping Gene-by-Gene Single-Nucleotide Variation in 8,535 *Mycobacterium tuberculosis* Genomes: a Resource To Support Potential Vaccine and Drug Development
- Analyzed whole-genome sequencing data from 17 studies to determine which genes are the most and least conserved in *M. tuberculosis*



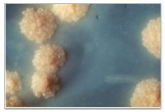
Mycobacterium tuberculosis

- Causative agent of tuberculosis (TB)
- High GC Gram+ Actinobacteria
- Genome is 4.4 million base pairs long and encodes for ~4,000 genes
    - https://doi.org/10.1038/31159



TAXONOMY

Was this helpful?



Mycobacterium tuberculosis

Mycobacterium tuberculosis is a species of high GC Gram+ in the family *Mycobacteriaceae*.

Taxonomy ID: 1773

Genomes
Browse and download

Literature

Bookshelf	1,914
MeSH	258
NLM Catalog	422
PubMed	84,512
PubMed Central	114,798

Genes

Gene	12,182
GEO DataSets	11,847
GEO Profiles	429,123
HomoloGene	5
PopSet	437

Proteins

Conserved Domains	331
Identical Protein Groups	800,290
Protein	21,789,153
Protein Family Models	1,934
Structure	3,849

Genomes

Assembly	7,144
BioCollections	0
BioProject	4,311
BioSample	150,649
Genome	6
Nucleotide	1,508,097
SRA	152,532
Taxonomy	1

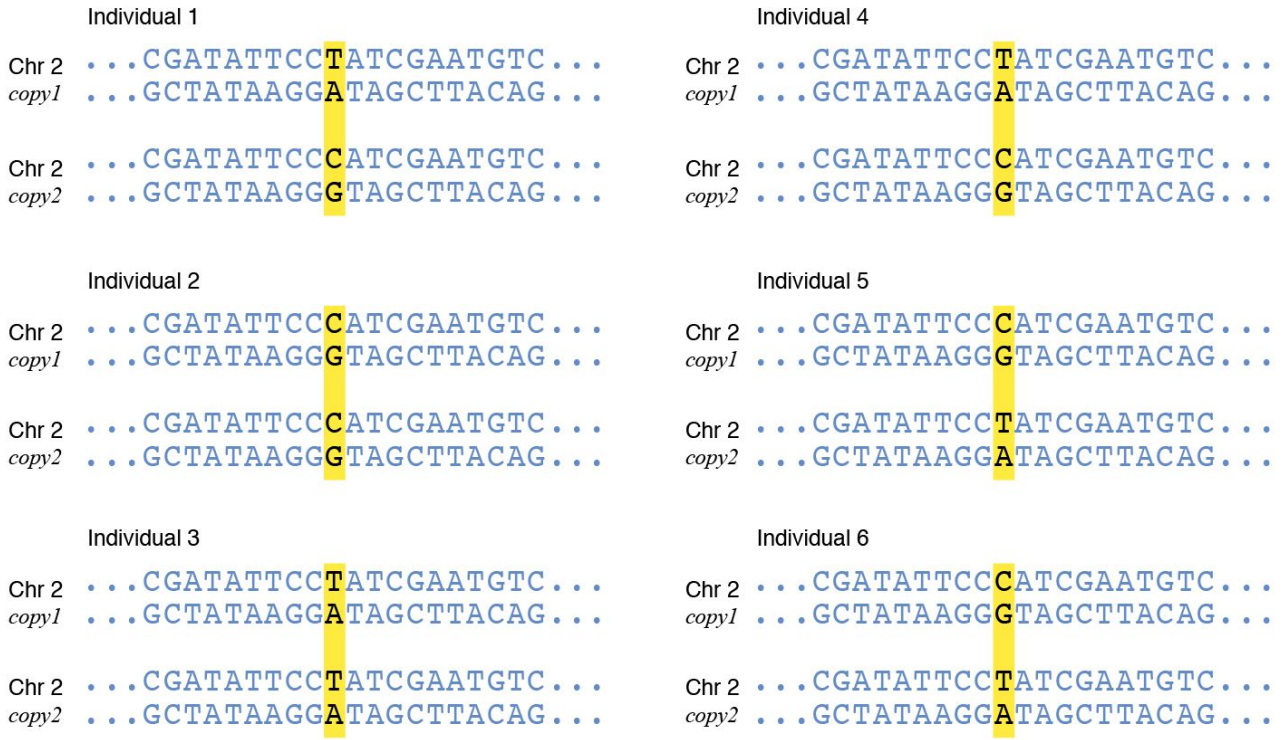
Clinical

ClinicalTrials.gov	608
ClinVar	21
dbGaP	4
dbSNP	0
dbVar	0
GTR	44
MedGen	45
OMIM	87

PubChem

BioAssays	20,487
Compounds	2
Pathways	292
Substances	26

Why look for SNPs?



General workflow

1. Download sequences from the SRA
2. Perform quality control and trimming of reads
3. Map reads to a reference genome (H37Rv)
4. Call variants
5. Filter variants
6. Compare variants across samples
7. Customize a phylogenetic tree

General workflow

1. **Download sequences from the SRA**
2. Perform quality control and trimming of reads
3. Map reads to a reference genome (H37Rv)
4. Call variants
5. Filter variants
6. Compare variants across samples
7. Customize a phylogenetic tree