Dataset name: Strain and Host Phenotype

Description: Each line represents a single patient isolate and is labeled with a unique id, with w prefix indicating that whole genome sequence data is available
Subject: Medicine, Health and Life Sciences
Keyword: antibiotic resistance, geography

Number of columns=38
Number of rows=2024

Variable (column name)	Values
StrainID[1]	Integer or w+integer
StrainName	Integer
ResistanceType[2]	S, LDR, MDR or XDR
IsoniazidDST	s, r, or i
EthionamideDST	s, r, or i
RifampicinDST	s, r, or i
RifabutinDST	s, r, or i
EthambutolDST	s, r, or i
PyrazinamideDST	s, r, or i
StreptomycinDST	s, r, or i
AmikacinDST	s, r, or i
CapreomycinDST	s, r, or i
KanamycinDST	s, r, or i
OfloxacinDST	s, r, or i
CiprofloxacinDST	s, r, or i
LevofloxacinDST	s, r, or i
MoxifloxacinDST	s, r, or i
GatifloxacinDST	s, r, or i
ParaAminoSalicylicAcidDST	s, r, or i
CycloserineDST	s, r, or i
ThioacetazoneDST	s, r, or i
ProthionamideDST	s, r, or i
ClofazimineDST	s, r, or i
ClarithromycinDST	s, r, or i
AmoxicillinClavulanateDST	s, r, or i
LinezolidDST	s, r, or i
Date	Date isolate was obtained
Country	Country name in English
City	City name in English
Setting	Description of isolation setting
Source	Name of source researcher/Laboratory
PatientAge	Age
PatientSex	Male/female
HIVStatus	negative or positive or NA

Spfamily_parentstrain[3]	Lineage family name
RFLPfamily	Lineage family name
PatientID	Integer
Spoligo_octal[4]	7.78E+14

[1]*This ID is the same ID as in the "Genetic Variation" dataset.*
[2]*S: pan-sensitive; LDR: 'low level' resistant, i.e. resistant to any number of drugs but not to both isonaizid or rifampicin; MDR: multidrug resistant i.e. resistant to both isonaizid and rifampicin; XDR: extensively drug resistant i.e. MDR with additional resistance to any fluoroquinolone (ofloxacin, ciprofloxacin, levofloxacin, moxifloxacin etc) and resistance to a second line injectable (amikacin, kanamycin or capreomycin)*
[3]*spfamily_parentstrain is the spoligotype based strain family and provides information on the lineage of the strain.*
[4]*Spoligotyping is a TB fingerprinting technique where PCR is used to amplify 43 unique spacers between base pair repeats. The products are hybridized in a membrane with oligonucleotides. Different strains of TB vary in terms of their spacers, so each strain produces a unique spot pattern that translates into a 15-digit code (called octal codes).*

Acronyms: DST: drug susceptibility test; s: sensitive, r: resistant; *RFLP restriction fragment length polymorphism analysis.*

DST description:

Isolate Source	**Culture**	**Drug Sensitivity Testing***
Stellenbosch University, South Africa	BACTEC MGIT 960 system (BD Diagnostics Systems, Sparks, MD)	Indirect proportion method on Middlebrook 7H11 agar slants supplemented individually with: RIF (1.0 µg/ml), INH (0.2 µg/ml), EMB (7.5 µg/ml), OFLX (2.0 µg/ml), KAN (5.0 µg/ml), STR (2.0 µg/ml), AMK (5.0 µg/ml), CAP (10 µg/ml). PZA sensitivity was tested using the MGIT 960 system (100 µg/ml).
Center for Disease Control, Atlanta, USA	Middlebrook 7H9 broth supplemented with 10% (vol/vol) albumin-dextrose-catalase enrichment (Difco Laboratories) and 0.05% (vol/vol) Tween 80 (Sigma-Aldrich) at 37°C until they reached an approximate optical density at 600 nm of 1.0 (corresponding to 5 x 10^8 CFU/ml)	Indirect proportion method on Middlebrook 7H10 agar plates supplemented individually with: RIF (1 µg/ml), INH (0.2, 1, and 5µg/ml), EMB (5 µg/ml), OFL (2 µg/ml), CIP (2 µg/ml), KAN (5 µg/ml), CAP (10 µg/ml), and AMI (4 µg/ml). PZA was tested using the BACTEC 460 (100 µg/ml), MGIT (100 µg/ml), or agar proportion (25 µg/ml) method.

Massachusetts State Laboratory (source country Peru, Russia)	Radiometric BACTEC 460 TB system (Becton-Dickinson)	Indirect proportion method on Middlebrook 7H10 agar plates supplemented with : INH (0.2, 1, and 5 µg/mL), RIF (1 µg/mL), EMB (5 µg/mL), STR (2 and 10 µg/mL), KAN (5 µg/mL), CAP (10 µg/mL), ETH (5 µg/mL), CYS (30 µg/mL), PAS (1 µg/mL), AMK (6 µg/mL), LEVO (1 µg/mL), OFLX (2 µg/mL), and CIP (2 µg/mL). PZA was tested using the BACTEC (100 µg/mL).
Public Health Research Institute, Rutgers University, Newark, NJ	Lowenstein-Jensen slant culture	Indirect agar proportion method using Middlebrook 7H10 agar plates containing the following drugs: RIF (1 µg/ml), INH (0.2, 1, and 5µg/ml), EMB (5 µg/ml), CIP (2 µg/ml), KAN (5 µg/ml), CAP (10µg/ml)
RVIM	MGIT and Middlebrook 7H10 solid	MGIT INH (0.2), RIF (1 µg/mL), RFB (2 µg/mL), EMB (5 µg/mL), STR (5 µg/mL), ETH (5 µg/mL), CYS (50 µg/mL), PRO (5 µg/mL), AMK (5 µg/mL), CLO (2 µg/mL), and CIP (2 µg/mL). PZA (50 µg/mL), CLARI (>16 µg/mL), PAS (1 µg/mL).
WHO/TDR	Dubos Agar then Lowenstein-Jensen slant culture	The indirect proportions method on two different media: On LJ for INH (0.2 µg/ml), RIF (40 µg/ml), EMB (2µg/ml),STR (4 µg/ml),PAS(0.5 µg/ml); on Middlebrook 7H11 agar for OFLX (2µg/ml), KAN (6µg/ml), CAP (10µg/ml).

Dataset: Genetic Variation

Description: Pooled data on genetic variation. Each row represents an individual Mycobacterium tuberculosis patient isolate. ID's starting with w are isolates that were whole genome sequenced. ID's not starting with 'w' belong to isolates that underwent targeted sequencing focusing on 27 loci putatively associated with drug resistance. In the targeted sequencing case the entire open reading frame and flanking 100bp (as annotated in the MTB H37Rv genome) were sequenced. (see reference Farhat et. al Manuscript submitted for more details)

Subject: Medicine, Health and Life Sciences

Keyword: Mycobacterium tuberculosis, nucleotide sequence

Labels are ID (same as column 1 of previous dataset) and SNP.