Characterisation of drug-resistant Mycobacterium tuberculosis mutations and transmission in Pakistan

|  |  |
| --- | --- |
| Gary Napier1,\* | [gary.napier@lshtm.ac.uk](mailto:gary.napier@lshtm.ac.uk) |
| Anwar Sheed Khan2,3,\* | [anwar786kp@hotmail.com](mailto:anwar786kp@hotmail.com) |
| Muhammad Tahir Khan4 | [muhammad.tahir8@imbb.uol.edu.pk](mailto:muhammad.tahir8@imbb.uol.edu.pk) |
| Sajid Ali2 | [sajidjan@live.com](mailto:sajidjan@live.com) |
| Muhammad Qasim2 | [qasim@kust.edu.pk](mailto:qasim@kust.edu.pk) |
| Susana Campino1 | [susana.campino@lshtm.ac.uk](mailto:susana.campino@lshtm.ac.uk) |
| Sajjad Ahmad5 | [sajjadahmad793@yahoo.com](mailto:sajjadahmad793@yahoo.com) |
| Hazir Rahman6 | [hazirrahman@hotmail.com](mailto:hazirrahman@hotmail.com) |
| Taj Ali Khan5,\*\* | [tajalikhan.ibms@kmu.edu.pk](mailto:tajalikhan.ibms@kmu.edu.pk) |
| Jody E. Phelan1,\*\* | [jody.phelan@lshtm.ac.uk](mailto:jody.phelan@lshtm.ac.uk) |
| Taane G. Clark1,7,\*\* | [taane.clark@lshtm.ac.uk](mailto:taane.clark@lshtm.ac.uk) |

1 Faculty of Infectious & Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK  
2 Department of Microbiology, Kohat University of Science and Technology, Kohat, Pakistan;  
3 Provincial Tuberculosis Reference, Laboratory Hayatabad Medical Complex, Peshawar, Pakistan;  
4 Institute of Molecular Biology and Biotechnology (IMBB), The University of Lahore. KM, Defense Road, Lahore 58810, Pakistan  
5 Institute of Basic medical Science Khyber Medical University, Peshawar KP, Pakistan  
6 Department of Microbiology, Abdul Wali Khan University, Mardan, Pakistan  
7 Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK

\* Joint authors  
\*\* Joint Correspondence:  
Prof. Taane G. Clark, London School of Hygiene & Tropical Medicine, UK; [taane.clark@lshtm.ac.uk](mailto:taane.clark@lshtm.ac.uk) Dr. Jody Phelan, London School of Hygiene & Tropical Medicine, UK; [jody.phelan@lshtm.ac.uk](mailto:jody.phelan@lshtm.ac.uk) Dr. Taj ali Khan, Institute of Basic medical Science Khyber Medical University Peshawar KP Pakistan; [tajalikhan.ibms@kmu.edu.pk](mailto:tajalikhan.ibms@kmu.edu.pk)

## ABSTRACT

## INTRODUCTION

Tuberculosis disease (TB), caused by bacteria in the Mycobacterium tuberculosis (*Mtb*) complex, is a major global public health problem. Pakistan is a high-burden TB country, being one of eight countries accounting for two-thirds of the estimated 10 million people globally that fell ill with the disease [1]. In 2019, Pakistan had a total TB incidence of 570,000 and 43,900 deaths [1], but disease control is being compromised by increasing HIV prevalence and drug resistance. The country has a high burden for rifampacin resistant (RR-TB) and multidrug-resistance (MDR-TB), which is the additional resistance to isoniazid treatments. There were 25,000 cases of MDR-/RR-TB in 2019 [1]. The National TB control program aims to reduce by half the prevalence of TB in the general population by 2025, but to achieve this will require the scaling-up of TB detection and clinical care, as well as improved systems for inferring disease transmission, thereby facilitating further targeted interventions.

Whole genome sequencing (WGS) is revolutionizing our understanding of drug resistance and clinical management, as well as transmission patterns, thereby assisting disease control [2]. *M. tuberculosis* drug resistance is linked to genomic variants in drug targets or pro-drug activators, including single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels), some occurring in gene-gene interactions. It is therefore possible to predict resistance genotypically for 19 anti-TB drugs and their groups (e.g. floroquinolines) using curated libraries of >1000 mutations across >30 loci [3],[4], thereby personalizing treatment. Genotypic predictions are an alternative to bacterial culture-based phenotypic drug susceptibility testing (DST), which can be time-consuming and resource intensive, with reproducibility and inhibitory concentration cut-off challenges for particular drugs [3]. Further, WGS data infers the population structure with the M. tuberculosis complex, which is phylo-geographical in nature, with strains falling within distinct (sub-)lineages [5], and potential transmission chains identified through isolates with (near-)identical genomic variation [6]. The identification of highly virulent strain-types or lineages, drug resistance, and transmission clusters will assist the targeting of limited resources for TB control.

There have been recent studies using WGS to characterize *M. tuberculosis* genetic diversity in isolates sourced from Pakistan, where the predominant strains are from the Central Asian (CAS) family, set within lineage 3 [7][8][9][10][11][12]. A recent study of TB endemic province of Khyber Pakhtunkhwa (North West Pakistan) found that known mutations in *rpoB* (e.g. S405L), *katG* (e.g. S315T), or *inhA* promoter loci explain the majority of MDR-TB, but there was evidence of complex mixed infections and heteroresistance, which may reflect the high transmission nature of the setting [12]. An earlier study in the same province found similar MDR-TB mutations, but also others in genes conferring resistance to other first and second-line drugs, including in *pncA* (pyrazinamide), *embB* (ethambutol), *gyrA* (fluoroquinolones), *rrs* (aminoglycosides), *rpsL*, *rrs* and *giB* (streptomycin) loci. Further, acquisition of rifampicin resistance often preceded isoniazid in our isolates, but worringly a high proportion (~18%) of pre-MDR isolates had fluoroquinolone resistance markers, potentially due to unregulated anti-TB drug use [7]. Eighteen *M. tuberculosis* isolates clustered within eight networks, thereby providing evidence of drug-resistant TB transmission in the Khyber Pakhtunkhwa province [7]. An investigation of XDR-TB isolates sourced across four provinces found similar genes linked to drug resistance as in Khyber Pakhtunkhwa [10], and follow-up work found an increased frequency and expression of novel SNP mutations in efflux pump genes, potentially explaining some drug resistance mechanisms.

Here, we analyse 535 *M. tuberculosis* samples with WGS data, collected between 2003 and 2020, with phenotypic testing of resistance across 12 drugs (rifampicin, isoniazid, ethambutol, pyrazinamide, streptomycin, ofloxacin, moxifloxacin, amikacin, kanamycin, capreomycin, ciprofloxacin, ethionamide). By identifying ~38k SNPs, and inferring genotypic drug resistance across 19 anti-TB drugs (as well as fluoroquinolones and aminoglycosides classes), we sought to understand the phylogeny of *M. tuberculosis* in the largest Pakistan dataset, identify transmission events, and infer commonly circulating mutations linked to drug resistance. The genetic insights were validated in a large *M. tuberculosis* collection (n=34k) with WGS and drug susceptibility test data [5].

## RESULTS

**Isolates and whole genome sequencing data**

A total of 535 M. tuberculosis isolates sourced between 2003 and 2005, and 2013 and 2020 from Pakistan with publically available WGS and phenotypic susceptibility testing [7][8][9][10][11][12]. These isolates covered four provinces (Balochistan, Khyber Pakhtunkhwa, Punjab, Sindh), but a high proportion of locations were missing (69.5%), all from one study [11]. (**Table 1**). The majority of samples were from lineage 3 (L3 397, 74.2%; CAS strains), but the other main lineages were represented (L4, 80, 15.0%, including LAM, T and X strains); L2 36, 6.7%, including Beijing; L1 22, 4.1%) (**Table 1**; **S1 Table**). Phenotypic drug susceptibility testing (DST) was most complete for rifampicin (n=487, 91.03%), isoniazid (n=487, 91.03%), ethambutol (n=479, 89.53%), and pyrazinamide (n=444, 82.99%) drugs (**S2 Table**), but incomplete / not present for 12 drugs. A total of 432 total samples (80.7%) were phenotypically resistant to at least one drug (median 3, maximum 10). The number of potential errors on the phenotypic testing appeared low (218/2,430 tests), where established genotypic resistance markers were present in isolates with DST results that implied sensitive. The discordance appeared in nine drugs (capreomycin, ethambutol, ethionamide, isoniazid, moxifloxacin, ofloxacin, pyrazinamide, rifampicin, streptomycin), but predominantly in ethambutol (96) (**S2 Table**).

There were 100/2,430 tests in the case where established genotypic resistance markers were absent in isolates with DST results that implied resistance. The discordance appeared in eight drugs (amikacin, ciprofloxacin, ethambutol, isoniazid, kanamycin, pyrazinamide, rifampicin, streptomycin), mostly in isonizid (25) and pyrazinamide (24) (**S2 Table**).

The majority of isolates were genotypically assessed as MDR-TB (328, 61.3%), with high proportions of (pre-) XDR (113, 21.1) and pan-sensitive (60, 11.2%) (**Table 1**). There were 31 pre-MDR, and overall there was a high prevalence of rifampicin (460, 86.0%) and isoniazid (435, 81.3%) resistance strains. Resistance to drugs used in combination with isoniazid and rifampicin were also common, including ethambutol (385, 72.0%), pyrazinamide (258, 48.2%), streptomycin (238, 44.5%), ethionamide (102, 19.1%), any fluoroquinolones (277, 51.8%) or aminoglycoside (75, 14.0%). As expected, very few isolates appeared resistant to bedaquiliine, clofazimine and cycloserine (n<3; **Table 1**). Across all lineages, the majority of isolates (>75%) were at least MDR-TB resistance (**S3 Table**).

After sequence data alignment, high average coverage was observed across the samples (median 76-fold, range 30-2,027). A total of 37,970 SNPs were identified genome-wide, with 23,741 (62.5%) found in single isolates.

A phylogenetic tree constructed using the genome-wide SNPs, and has the expected lineage-based clustering (**Figure 1**). The tree also picked out clusters of samples separated by the 10-SNP distance threshold used to establish “transmission” and “non-transmission” samples. Clusters of Pre-XDR and XDR samples can also be observed in **Figure 1**.

### Evidence of transmission

The median (range) pairwise SNP differences across the 535 isolates was 390 (minimum 0, maximum 1811), with a multi-modal distribution, where modes representing differences within and between lineages (**S2 Figure**). At a threshold of 10 SNPs, 55 clusters formed consisting of a total of 169 isolates, where the median number of isolates in each cluster was 2 (range: 2 - 22) (**S2 Figure**). The effects of changing the SNP threshold on clusters number and sizes (**S4 Table**) indicated the presence of only six additional clusters (33 more isolates) when increasing from a stringent threshold (5 SNPs) to our pragmatic choice (10 SNPs).

Comparing the 169 "transmitted" isolates in clusters to the others ("non-transmitted"; n=366), there was an increased risk of transmission in lineages 2 and 4 (overall chi-sq=38.853, p=1.9e-08; post-hoc tests: lineage 2 standardised residual=3.57, p=0.003; lineage 4 standardised residual=4.62, p=<0.0001), and Pre-XDR and XDR samples (overall chi-sq=98, p=0.0005; post-hoc tests: Pre-XDR standardised residual=5.31, p=<0.0001; XDR standardised residual=6.55, p=<0.0001).

"Transmitted" isolates were found in three of the four provinces - Khyber Pakhtunkhwa (71), Punjab (9) and Sindh (9). At the city and regional level, only the city of Peshawar was significant for increased risk of transmission (overall chi-sq=70, p=0.0005; post-hoc tests: standardised residual=5.74, p=<0.0001) (**S5 Table**; **S3 Figure**; **S4 Figure**).

A genome-wide association study (GWAS) revealed genes *Rv2102*, *Rv0914c*, *nusG*, *Rv2184c* and *Rv1896c* to be the top-five most closely associated with the "transmitted" samples (**Table XXX**, **Figure XXX**). Two of these genes are uncharacterised (*Rv2102* and *Rv2102*). Two genes have putative functions: *Rv0914c* is thought to be a ‘lipid carrier protein’ or ‘keto acyl-CoA thiolase’, and *Rv1896c* is thought to code for ‘S-adenosyl-L-methionine-dependent methyltransferase’. Only *nusG* is definitively characterised, coding for transcription termination/antitermination protein NusG [13].

**Drug resistance mutations** The common mutations underlying genotypic drug resistance were in known mutations in *rpoB* (rifampicin; 460), *katG* (isoniazid; 416), *embA* / *embB* (ethambutol; 385), *gyrA* (fluoroquinolones; 277) and *pncA* (pyrazinamide; 258) (**S6 Table**).

**Potential novel resistance markers** We investigated isolates that had a DST implying resistance, but no established genetic mutations to explain this phenotype. There were 82 isolates (100 tests) with discordance in 8 drugs (amikacin, ciprofloxacin, ethambutol, isoniazid, kanamycin, pyrazinamide, rifampicin, streptomycin), and identified 93 distinct potential genetic markers in candidate genes to explain the discordance (**S2 Table**; **Table 3**).

For rifampicin resistance, we identified three inframe indels in *rpoB* (1291\_1292insGCC, 1294\_1296del and 1309\_1311del) in three isolates. For isoniazid, several nonsense mutations in *katG* were found, with 3 mutations leading to premature stop codons (W438\*, W2048\*, Q36\*) and a frameshift mutation (587\_588insGGT). For ethambutol resistance, variants in the *embA* promoter region (-42CAT>C, -27TA>T-16C>A, -8C>A) and *embB* were observed. For pyrazinamide resistance, several potentially new mutations we found in pncA, including three inframe indels (511\_512insTCGCCG, 392\_393insGGT and 451\_462del) a premature stop codon (S18\*), and SNPs in both the coding region (Val180Ala) and the promoter (-7T>G). For streptomycin resistance, several mutations were found in *gid* including a premature stop codon (G71\*), frameshift (102\_102del), and SNPs (A119D, A82P and D67G). These SNPs were found in the 34k global dataset, and likely acquired as the result of homoplasy. The *gid* A119D mutation was present in 15 isolates (ten different sublineages), of which two had DSTs and were resistant. The *gid* A82P mutation was present in three isolates from two different sublineages, but no DST was available for these samples. The *gid* D67G was present in 38 global isolates from five different sublineages. Of these, seven had DST data available with four presenting resistance.

For second line injectables, the *rrs* 878g>a mutation (seen previously[7]) was observed in four lineage 3 strains with three independent homoplastic acquisitions, indicating it is unlikely to be strain-specifc. Mutations in *rrs* are generally clustered in two regions with the most common mutations involved with streptomycin resistance being located around position 514 and those involved with resistance to amikacin, kanamycin and capreomycin located around 1401. The *rrs* 878g>a falls between the two mutation hotspots, and of the three strains which had DST data (amikacin and kanamycin) in this study, two were resistant to both amikacin and kanamycin and the other was sensitive to both. For fluoroquinolones, the *gyrA* A288D mutation was found in three lineage 3 isolates and was acquired in each sample independently. One isolate was tested resistant to ciprofloxacin with no known resistance mutation found in the *gyrA* and *gyrB* genes. The mutation has been previously observed to increase MIC to both ofloxacin and moxifloxacin [14].

### DISCUSSION XXXXXXX NOT LOOKED AT, NEEDS EDITING XXXXX

This study reinforces the value of whole genome sequencing in the context of low resource and high TB burden settings. Our findings on potential transmission among lineages 2 and 4, Pre-XDR and XDR TB, and in Peshawar, as well as drug resistance mutations can inform better epidemiological, clinical, and control decisions in Pakistan and, more generally, provides insight into mutations relevant to drug design. WGS can potentially be particularly useful in countries similar to Pakistan where effective public health surveillance is inadequate due to socio-economic problems - genetic data can fill gaps in data collection, revealing salient connections and variation.

Evidence of increased transmission among lineages 2 and 4 is consistent with previous characterisations of these clades as more transmissible [15][16], and therefore ought to be monitored more closely despite greater prevalence of lineage 3. It is surprising that Pre-XDR and XDR samples were found to be clustered more than expected given the usual fitness cost of drug resistance [17][18]. This however suggests unknown compensatory mutations and ought to be investigated in future work.

Inframe deletions have not been widely reported as a major mechanism of resistance to rifampicin and it is surprising to see the relative high number of these mutations in our dataset. Nonsense mutations are presumed to lead loss of function in the *katG* gene since it codes for the activator of isoniazid (catalase-peroxidase enzyme), so it is unsurprising to see potentially drug-resistance conferring mutations here. Mutations in the -16 and -8 position of the *embA* promoter have been reported to enhance binding of the *embR* transcription regulator to the promoter and increase expression of *embCAB* operon. The deletions are also presumed to have a similar effect to enhance binding of *embR*, though further experimental evidence would be required in future. The two strains which were resistant to both amikacin and kanamycin also contained a mutation in the promoter region to *eis* (-14C>T). This mutation could potentially explain the resistance to both these drugs and may point to the 878g>a being either increasing the level of resistance to these drugs or potentially conferring resistance to streptomycin (which was not tested for in these samples). The *pncA* gene codes for the activator of pyrazinamide and loss of function leads to resistance. The functional consequence of the premature stop codons and indels are usually severe to protein function and these mutations in our isolates represent highly likely candidates as conferring resistance. Mutations in the promoter region of *pncA* leads to changes in the expression of PncA and resistance [19]. The c.-7T>G mutation is thus also likely to cause resistance. The functional effect of SNPs in the coding region of *pncA* can be harder to predict, however the Val180Ala mutation was reported previously to be associated to pyrazinamide resistance in the CARD database [20].

The incomplete penetrance of the streptomycin-associated D67Gly mutation could be explained by the relative low-level resistance conferred by mutation in gid which could be under the critical concentration but still elevated with respect to wild-type.

## METHODS

**Sequence data and processing**

WGS were sourced across 6 studies [7][8][9][10][11][12] (ENA accessions: PRJEB7798, PRJEB10385, PRJEB25972, PRJEB32684, PRJEB43284). Raw reads were trimmed to remove low-quality sequences in Trimmomatic (v0.39) [21], and aligned to the H37Rv reference genome (AL123456) with BWA mem (v0.7.17) [22]. SNPs and indels were called using samtools software [23] and was processed in gatk GenotypeGVCFs (v4.1.3.0) (gatk.broadinstitute.org). Monomorphic SNPs and variants in non-unique regions of the genome (e.g. pe/ppe genes) were excluded. A multi-FASTA format file was created from the filtered SNP file and H37Rv reference fasta using bedtools makewindows (v2.28.0) [24]. This multiple alignment was used to construct a phylogenetic tree with IQ-TREE (v1.6.12), involving a general time reversible model with rate heterogeneity set to a discrete Gamma model and an ascertainment bias correction (parameters -m GTR+G+ASC), with 1000 bootstrap samples [25]. Pairwise distance matrices were calculated in Plink software (v1.90b4) [26]. Drug resistance and lineages were predicted in silico from raw sequence data using TB-Profiler (v2.4) [3]. The Pakistan analysis results were compared to a global collection of 34k M. tuberculosis with WGS and DST data [5].

Linear mixed models have been previously applied to adjust for the confounds of Mtb (sub-)lineage and outbreak-based population structure [27]. Therefore a linear mixed regression model was conducted to estimate the strength of association between the binary "transmission"/"non-transmission" outcome and SNPs, conducted in GEMMA (v.1.1.2) [28]. For sublineages in the Pakistan dataset for which there were > 10 samples (1.1.2, 2.2.1, 3, 3.1.2, 4.5, 4.9), other samples from a global dataset were found which also belonged to these lineages. Phylogenetic trees of the concatenated Pakistan and global samples were then constructed, for each sublineage separately, with FastTree [29]. Likelihoods of ancestral locations were conducted with the ape 5.0 [30] and phytools [31] packages in R (**Figure 2**).

### Tables

Table 1: Mycobacterium tuberculosis samples (N = 535)

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Characteristic | N | % |
| Lineage | 1 | 22 | 4.11 |
| Lineage | 2 | 36 | 6.73 |
| Lineage | 3 | 397 | 74.21 |
| Lineage | 4 | 80 | 14.95 |
| DR status | Sensitive | 60 | 11.21 |
| DR status | Pre-MDR | 31 | 5.79 |
| DR status | MDR | 328 | 61.31 |
| DR status | Pre-XDR | 47 | 8.79 |
| DR status | XDR | 66 | 12.34 |
| DR status | Other | 3 | 0.56 |
| Predicted drug resistance | rifampicin | 460 | 85.98 |
| Predicted drug resistance | isoniazid | 435 | 81.31 |
| Predicted drug resistance | ethambutol | 385 | 71.96 |
| Predicted drug resistance | pyrazinamide | 258 | 48.22 |
| Predicted drug resistance | streptomycin | 238 | 44.49 |
| Predicted drug resistance | ofloxacin | 277 | 51.78 |
| Predicted drug resistance | moxifloxacin | 277 | 51.78 |
| Predicted drug resistance | levofloxacin | 277 | 51.78 |
| Predicted drug resistance | amikacin | 75 | 14.02 |
| Predicted drug resistance | kanamycin | 79 | 14.77 |
| Predicted drug resistance | capreomycin | 78 | 14.58 |
| Predicted drug resistance | ciprofloxacin | 277 | 51.78 |
| Predicted drug resistance | ethionamide | 102 | 19.07 |
| Predicted drug resistance | para\_aminosalicylic\_acid | 10 | 1.87 |
| Predicted drug resistance | cycloserine | 2 | 0.37 |
| Predicted drug resistance | clofazimine | 1 | 0.19 |
| Predicted drug resistance | bedaquiline | 1 | 0.19 |
| Predicted drug resistance | linezolid | 0 | 0.00 |
| Predicted drug resistance | delamanid | 0 | 0.00 |
| Predicted drug resistance | clarithromycin | NA | NA |
| Predicted drug resistance | rifabutin | NA | NA |
| Predicted drug resistance | fluoroquinolones | 277 | 51.78 |
| Predicted drug resistance | prothionamide | NA | NA |
| Predicted drug resistance | aminoglycosides | 75 | 14.02 |
| Collection year | 2013 | 49 | 9.16 |
| Collection year | 2015 | 364 | 68.04 |
| Collection year | 2016 | 38 | 7.10 |
| Collection year | 2017 | 36 | 6.73 |
| Collection year | 2018 | 21 | 3.93 |
| Collection year | 2019 | 22 | 4.11 |
| Collection year | 2020 | 5 | 0.93 |
| Region | Missing | 372 | 69.53 |
| Region | Peshawar | 77 | 14.39 |
| Region | Dera Ismail Khan | 25 | 4.67 |
| Region | Abbottabad | 13 | 2.43 |
| Region | Other | 13 | 2.43 |
| Region | Swat | 13 | 2.43 |
| Region | Rawalpindi | 7 | 1.31 |
| Region | Hyderabad | 5 | 0.93 |
| Region | Karachi | 5 | 0.93 |
| Region | Lahore | 5 | 0.93 |

**S1 Table: Strain-types**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Lineage | LSP lineage | Spoligotype family | Region of difference no. | N | % |
| 3 | East-African-Indian | CAS | RD750 | 363 | 67.850 |
| 2.2.1 | East-Asian (Beijing) | Beijing-RD181 | RD105;RD207;RD181 | 29 | 5.421 |
| 4.5 | Euro-American | H;T | RD122 | 23 | 4.299 |
| 4.9 | Euro-American (H37Rv-like) | T1 | None | 22 | 4.112 |
| 1.1.2 | Indo-Oceanic | EAI3;EAI5 | RD239 | 14 | 2.617 |
| 3.1.2 | East-African-Indian | CAS;CAS2 | RD750 | 11 | 2.056 |
| 3.1.2.1 | East-African-Indian | CAS2 | RD750 | 10 | 1.869 |
| 3.1.3 | East-African-Indian | CAS | RD750 | 9 | 1.682 |
| 4.8 | Euro-American (mainly T) | T1;T2;T3;T5 | RD219 | 9 | 1.682 |
| 1.2.2.2 | Indo-Oceanic | NA | RD239 | 7 | 1.308 |
| 4.2.2.2 | Euro-American (Ural) | T;LAM7-TUR | None | 7 | 1.308 |
| 2.2.1.1 | East-Asian (Beijing) | Beijing-RD150 | RD105;RD207;RD181;RD150 | 5 | 0.935 |
| 3.1 | East-African-Indian | Non-CAS1-Delhi | RD750 | 4 | 0.748 |
| 2.2.1.2 | East-Asian (Beijing) | Beijing-RD142 | RD105;RD207;RD181;RD142 | 2 | 0.374 |
| 4 | Euro-American | LAM;T;S;X;H | None | 2 | 0.374 |
| 4.1.1.1 | Euro-American (X-type) | X2 | RD183 | 2 | 0.374 |
| 4.1.1.3 | Euro-American (X-type) | X1;X3 | RD193 | 2 | 0.374 |
| 4.2.1 | Euro-American (TUR) | H3;H4 | None | 2 | 0.374 |
| 4.2.2 | Euro-American (Ural) | T;LAM7-TUR | None | 2 | 0.374 |
| 4.6 | Euro-American | T;LAM | None | 2 | 0.374 |
| 4.6.5 | Euro-American | T;LAM | None | 2 | 0.374 |
| 1.1.3.3 | Indo-Oceanic | NA | RD239 | 1 | 0.187 |
| 4.1.1.2 | Euro-American (X-type) | X1 | None | 1 | 0.187 |
| 4.1.2.1 | Euro-American (Haarlem) | T1;H1 | RD182 | 1 | 0.187 |
| 4.6.2 | Euro-American | T;LAM | RD726 | 1 | 0.187 |
| 4.6.2.1 | Euro-American | T3 | RD726 | 1 | 0.187 |
| 4.6.2.2 | Euro-American (Cameroon) | LAM10-CAM | RD726 | 1 | 0.187 |

**S2 Table: Drug-resistant samples according to drug susceptibility tests (DST) and genotypic predictions**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Drug | DST N | DST N resistant | DST % resistant | DST % resistant of total | N genotypic resistant | % genotypic resistant | DST - Geno + | DST + Geno - |
| rifampicin | 487 | 417 | 85.63 | 77.94 | 460 | 85.98 | 6 | 6 |
| isoniazid | 487 | 411 | 84.39 | 76.82 | 435 | 81.31 | 7 | 25 |
| ethambutol | 479 | 265 | 55.32 | 49.53 | 385 | 71.96 | 96 | 17 |
| pyrazinamide | 444 | 189 | 42.57 | 35.33 | 258 | 48.22 | 42 | 24 |
| streptomycin | 43 | 24 | 55.81 | 4.49 | 238 | 44.49 | 4 | 6 |
| ofloxacin | 85 | 46 | 54.12 | 8.60 | 277 | 51.78 | 5 | 0 |
| moxifloxacin | 52 | 4 | 7.69 | 0.75 | 277 | 51.78 | 29 | 0 |
| levofloxacin | 0 | 0 | NaN | 0.00 | 277 | 51.78 | 0 | 0 |
| amikacin | 110 | 42 | 38.18 | 7.85 | 75 | 14.02 | 0 | 9 |
| kanamycin | 112 | 44 | 39.29 | 8.22 | 79 | 14.77 | 0 | 7 |
| capreomycin | 57 | 15 | 26.32 | 2.80 | 78 | 14.58 | 18 | 2 |
| ciprofloxacin | 37 | 37 | 100.00 | 6.92 | 277 | 51.78 | 0 | 4 |
| ethionamide | 37 | 6 | 16.22 | 1.12 | 102 | 19.07 | 11 | 0 |
| para\_aminosalicylic\_acid | 0 | 0 | NaN | 0.00 | 10 | 1.87 | 0 | 0 |
| cycloserine | 0 | 0 | NaN | 0.00 | 2 | 0.37 | 0 | 0 |
| clofazimine | 0 | 0 | NaN | 0.00 | 1 | 0.19 | 0 | 0 |
| bedaquiline | 0 | 0 | NaN | 0.00 | 1 | 0.19 | 0 | 0 |
| linezolid | 0 | 0 | NaN | 0.00 | 0 | 0.00 | 0 | 0 |
| delamanid | 0 | 0 | NaN | 0.00 | 0 | 0.00 | 0 | 0 |
| clarithromycin | 0 | 0 | NaN | 0.00 | NA | NA | NA | NA |
| rifabutin | 0 | 0 | NaN | 0.00 | NA | NA | NA | NA |
| fluoroquinolones | NA | NA | NA | NA | 277 | 51.78 | NA | NA |
| prothionamide | 0 | 0 | NaN | 0.00 | NA | NA | NA | NA |
| aminoglycosides | NA | NA | NA | NA | 75 | 14.02 | NA | NA |

**S3 Table: Drug resistance categories by Lineage (L)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DR status | L1 | L2 | L3 | L4 | Total |
| Sensitive | 1 (0.19%) | 0 (0.00%) | 53 (9.91%) | 6 (1.12%) | 60 (11.21%) |
| Pre-MDR | 2 (0.37%) | 0 (0.00%) | 28 (5.23%) | 1 (0.19%) | 31 (5.79%) |
| MDR | 12 (2.24%) | 25 (4.67%) | 242 (45.23%) | 49 (9.16%) | 328 (61.31%) |
| Pre-XDR | 0 (0.00%) | 7 (1.31%) | 25 (4.67%) | 15 (2.80%) | 47 (8.79%) |
| XDR | 5 (0.93%) | 4 (0.75%) | 48 (8.97%) | 9 (1.68%) | 66 (12.34%) |
| Other | 2 (0.37%) | 0 (0.00%) | 1 (0.19%) | 0 (0.00%) | 3 (0.56%) |
| Total | 22 (4.11%) | 36 (6.73%) | 397 (74.21%) | 80 (14.95%) | 535 (100.00%) |

**S4 Table: Sensitivity analysis of clustering by SNP distance, summary statistics**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SNP\_dist | N\_clusters | N | Median | Min | Max | Lin\_1 | Lin\_2 | Lin\_3 | Lin\_4 | Sensitive | Pre.MDR | MDR | Pre.XDR | XDR | Other |
| 0 | 28 | 60 | 2 | 2 | 3 | 2 | 6 | 35 | 17 | 2 | 2 | 17 | 19 | 20 | NA |
| 1 | 28 | 60 | 2 | 2 | 3 | 2 | 6 | 35 | 17 | 2 | 2 | 17 | 19 | 20 | NA |
| 5 | 49 | 136 | 2 | 2 | 17 | 7 | 16 | 77 | 36 | 2 | 3 | 60 | 29 | 40 | 2 |
| 10 | 55 | 169 | 2 | 2 | 22 | 7 | 21 | 98 | 43 | 2 | 3 | 87 | 31 | 44 | 2 |
| 15 | 54 | 176 | 2 | 2 | 22 | 8 | 21 | 103 | 44 | 2 | 4 | 90 | 32 | 46 | 2 |
| 20 | 63 | 200 | 2 | 2 | 22 | 8 | 24 | 121 | 47 | 2 | 5 | 106 | 35 | 49 | 3 |
| 25 | 68 | 213 | 2 | 2 | 22 | 8 | 24 | 131 | 50 | 2 | 5 | 118 | 35 | 50 | 3 |
| 30 | 71 | 220 | 2 | 2 | 22 | 8 | 25 | 137 | 50 | 2 | 5 | 124 | 35 | 51 | 3 |
| 35 | 74 | 227 | 2 | 2 | 22 | 8 | 27 | 142 | 50 | 2 | 6 | 130 | 35 | 51 | 3 |
| 40 | 75 | 232 | 2 | 2 | 22 | 8 | 27 | 147 | 50 | 3 | 6 | 134 | 35 | 51 | 3 |
| 45 | 80 | 242 | 2 | 2 | 22 | 8 | 27 | 157 | 50 | 6 | 7 | 140 | 35 | 51 | 3 |
| 50 | 80 | 254 | 2 | 2 | 22 | 8 | 27 | 169 | 50 | 7 | 8 | 149 | 35 | 52 | 3 |

**S5 Table: Characteristics of 169 M. tuberculosis isolates in 55 clusters with a SNP distance of 10 compared to others**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | in\_trans | in\_trans\_pc | not\_in\_trans | not\_in\_trans\_pc |
| 1 | Lineage | 1 | 7 | 4.1 | 15 | 4.1 |
| 2 | Lineage | 2 | 21 | 12.4 | 15 | 4.1 |
| 3 | Lineage | 3 | 98 | 58.0 | 299 | 81.7 |
| 4 | Lineage | 4 | 43 | 25.4 | 37 | 10.1 |
| 5 | DR status | Sensitive | 2 | 1.2 | 58 | 15.8 |
| 31 | DR status | Pre-MDR | 3 | 1.8 | 28 | 7.7 |
| 11 | DR status | MDR | 87 | 51.5 | 241 | 65.8 |
| 41 | DR status | Pre-XDR | 31 | 18.3 | 16 | 4.4 |
| 6 | DR status | XDR | 44 | 26.0 | 22 | 6.0 |
| 21 | DR status | Other | 2 | 1.2 | 1 | 0.3 |
| 12 | Location | Abbottabad | 6 | 3.6 | 7 | 1.9 |
| 22 | Location | Dera Ismail Khan | 13 | 7.7 | 12 | 3.3 |
| 32 | Location | Hyderabad | 3 | 1.8 | 2 | 0.5 |
| 42 | Location | Karachi | 2 | 1.2 | 3 | 0.8 |
| 51 | Location | Lahore | 4 | 2.4 | 1 | 0.3 |
| 61 | Location | Missing | 80 | 47.3 | 292 | 79.8 |
| 7 | Location | Other | 6 | 3.6 | 7 | 1.9 |
| 8 | Location | Peshawar | 46 | 27.2 | 31 | 8.5 |
| 9 | Location | Rawalpindi | 3 | 1.8 | 4 | 1.1 |
| 10 | Location | Swat | 6 | 3.6 | 7 | 1.9 |

**S6 Table: Number of samples with known drug resistance-associated mutations**

**TOO LARGE - ATTACHED**

**S7 Table: Phenotypically resistance samples (n=82) with variants previously unknown to be associated with drug resistance**

**TOO LARGE - ATTACHED**

**S8 Genome-wide association analysis of transmission – top five genes.**

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | Function | beta | -log10 Wald p-value |
| nusG | Transcription termination/antitermination protein NusG | 0.791 | 9.23 |
| Rv0914c | Possible lipid carrier protein or keto acyl-CoA thiolase | -0.781 | 9.64 |
| Rv1896c | Putative S-adenosyl-L-methionine-dependent methyltransferase | -0.643 | 8.05 |
| Rv2102 | Uncharacterised | -0.775 | 9.71 |
| Rv2184c | Uncharacterized | -0.699 | 8.56 |

### Figures

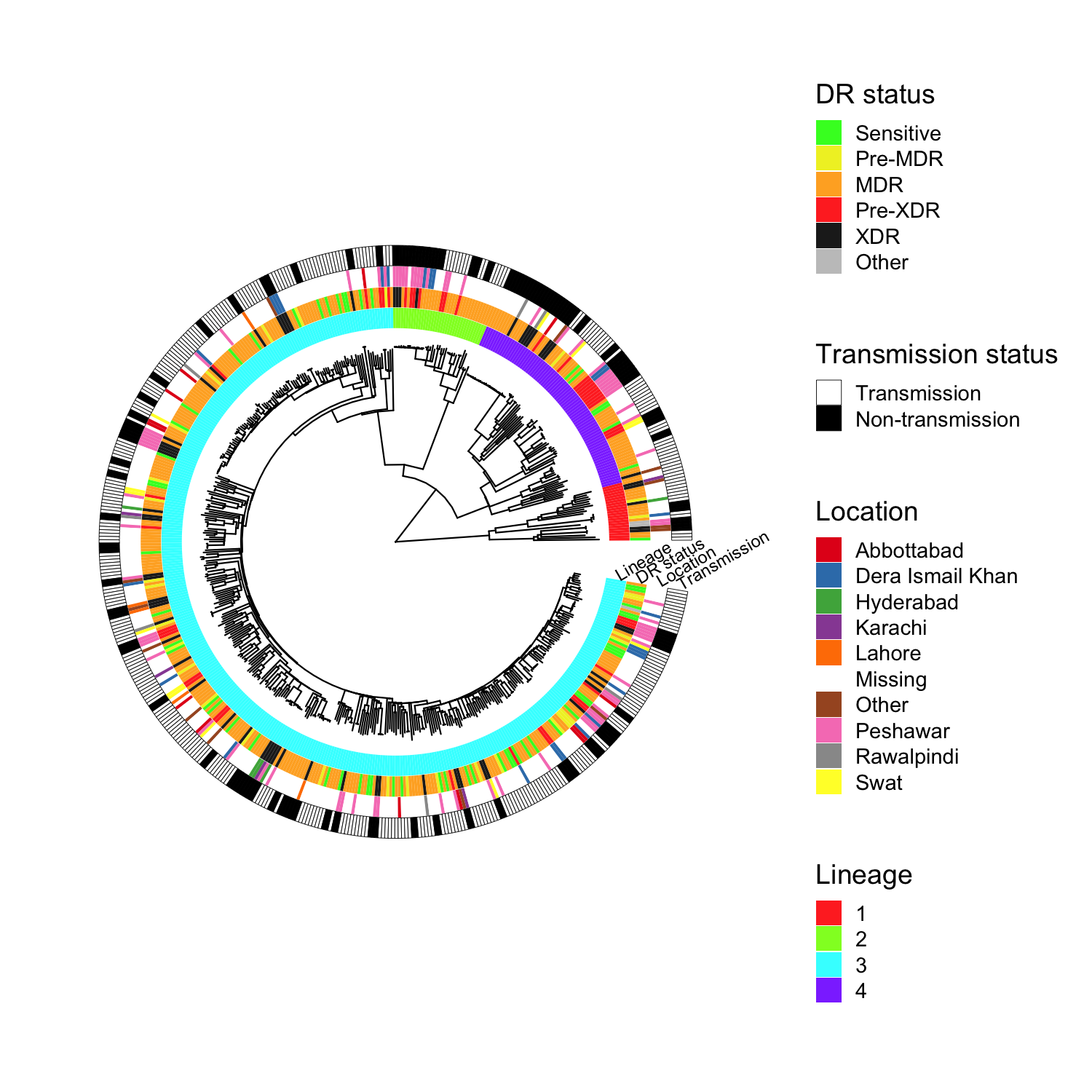
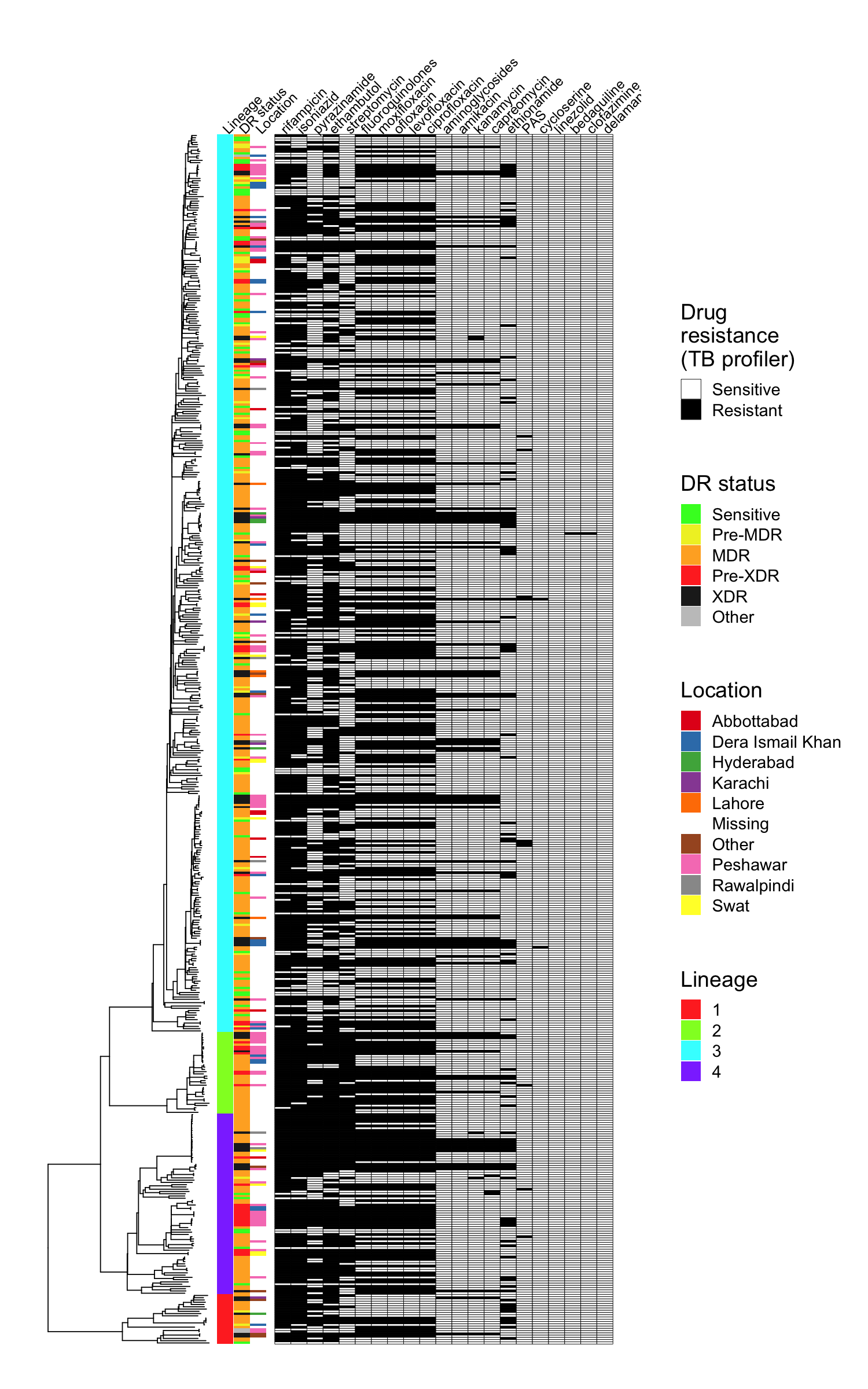


Figure 1: The 535 M. tuberculosis isolates: A phylogenetic tree constructed using 37,970 SNPs

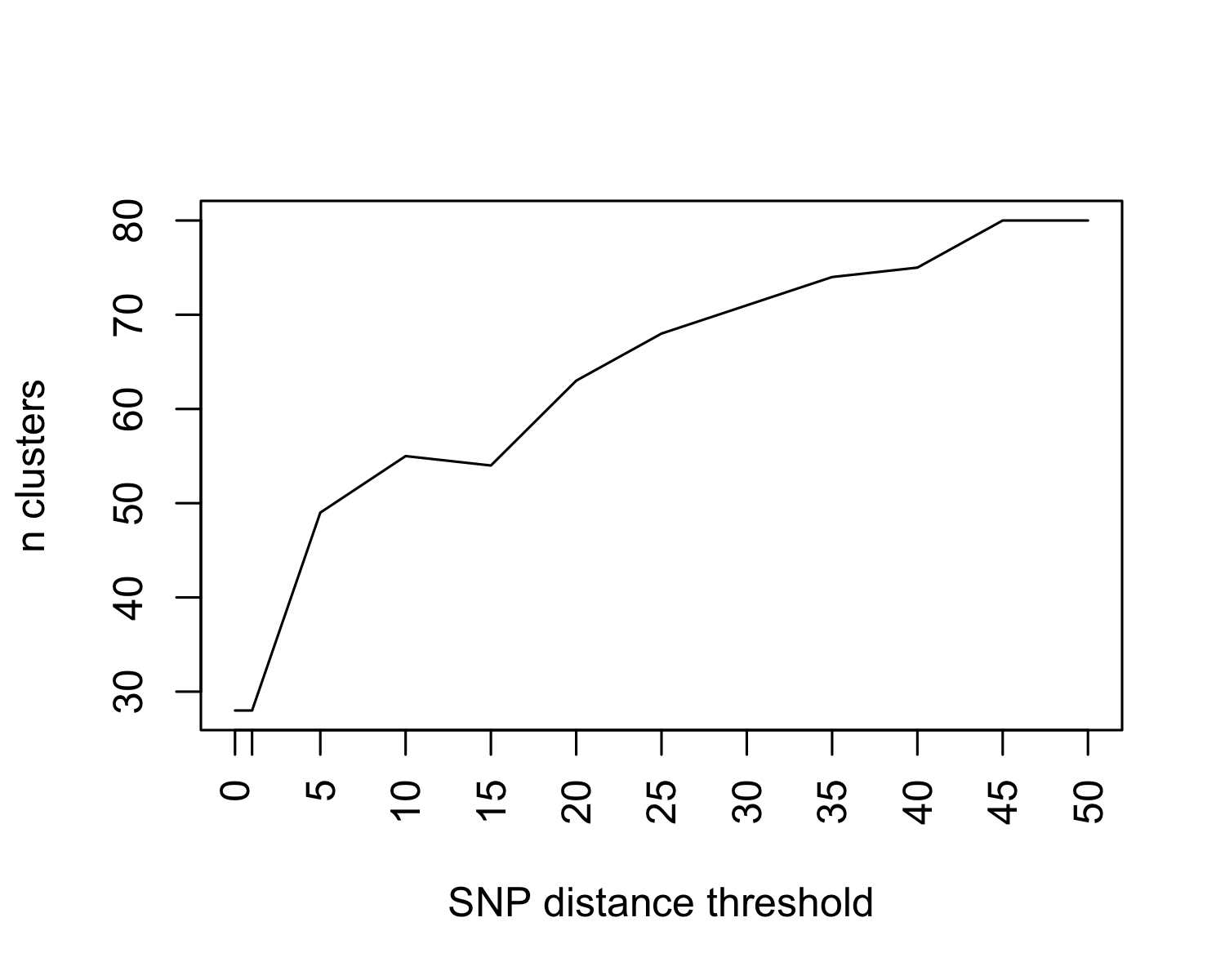
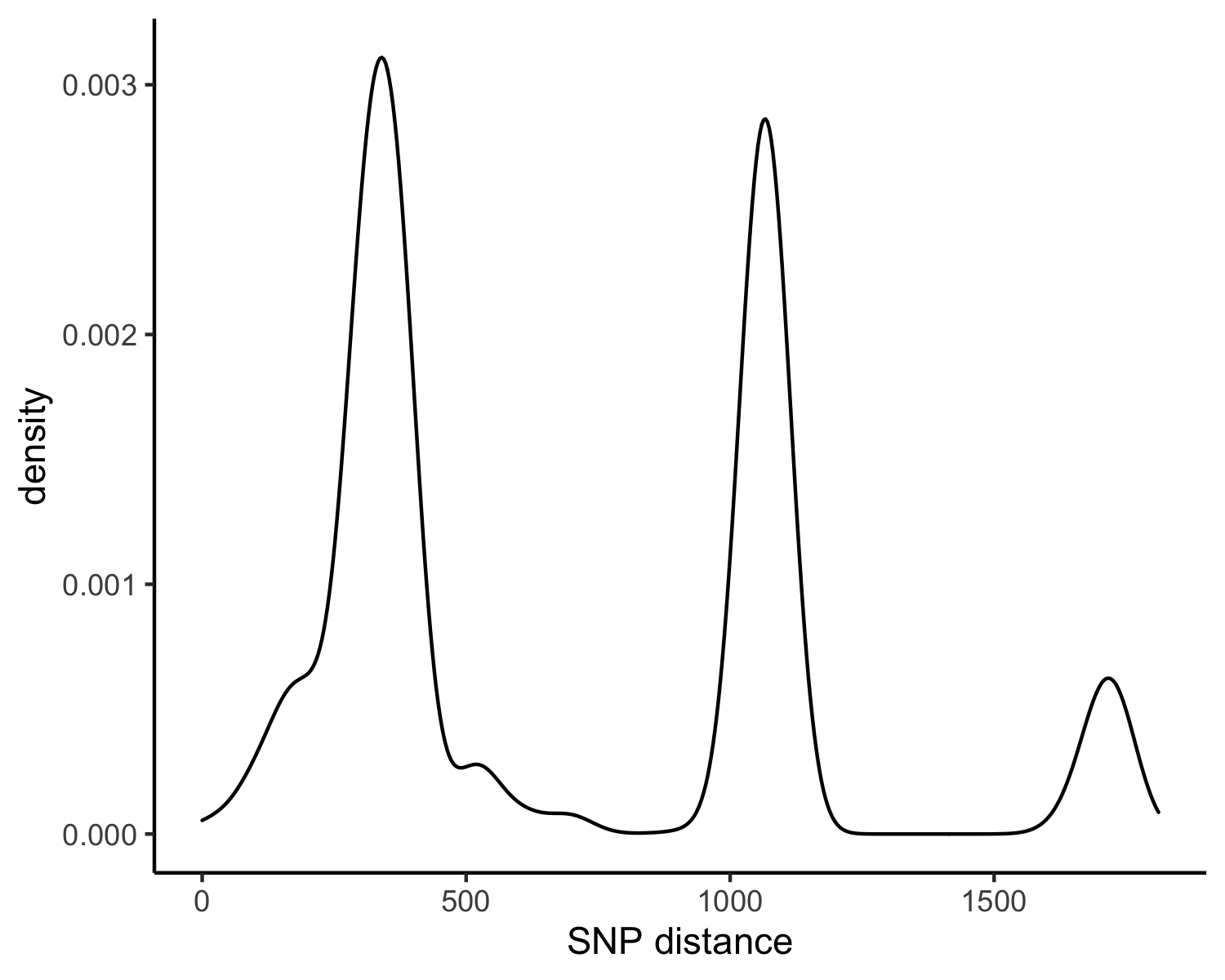
Figure 2: Phylogenetic tree with the Pakistan samples and closely-related global isolates from previously published datasets (n= ) XXXX need to describe in methods and results XXXX. The posterior probability for the ancestral state reconstruction of subregion of sample collection is indicated by the pie charts on the internal nodes. Sublineages 1.1.2, 2.2.1, 3, 3.1.2, 4.5 and 4.9

**TOO LARGE – ATTACHED**

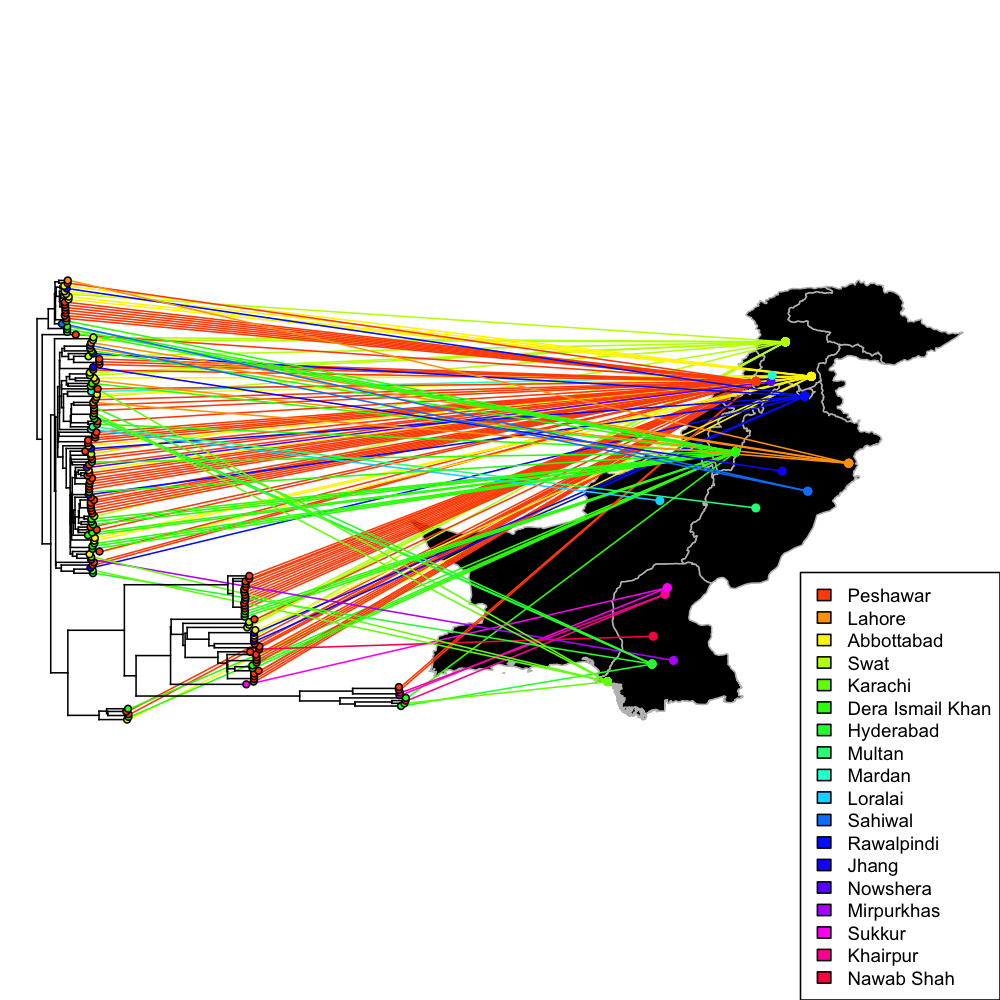
**S1 Figure: Phylogenetic tree for the 535 M. tuberculosis isolates with individual genomic drug resistance predictions.**



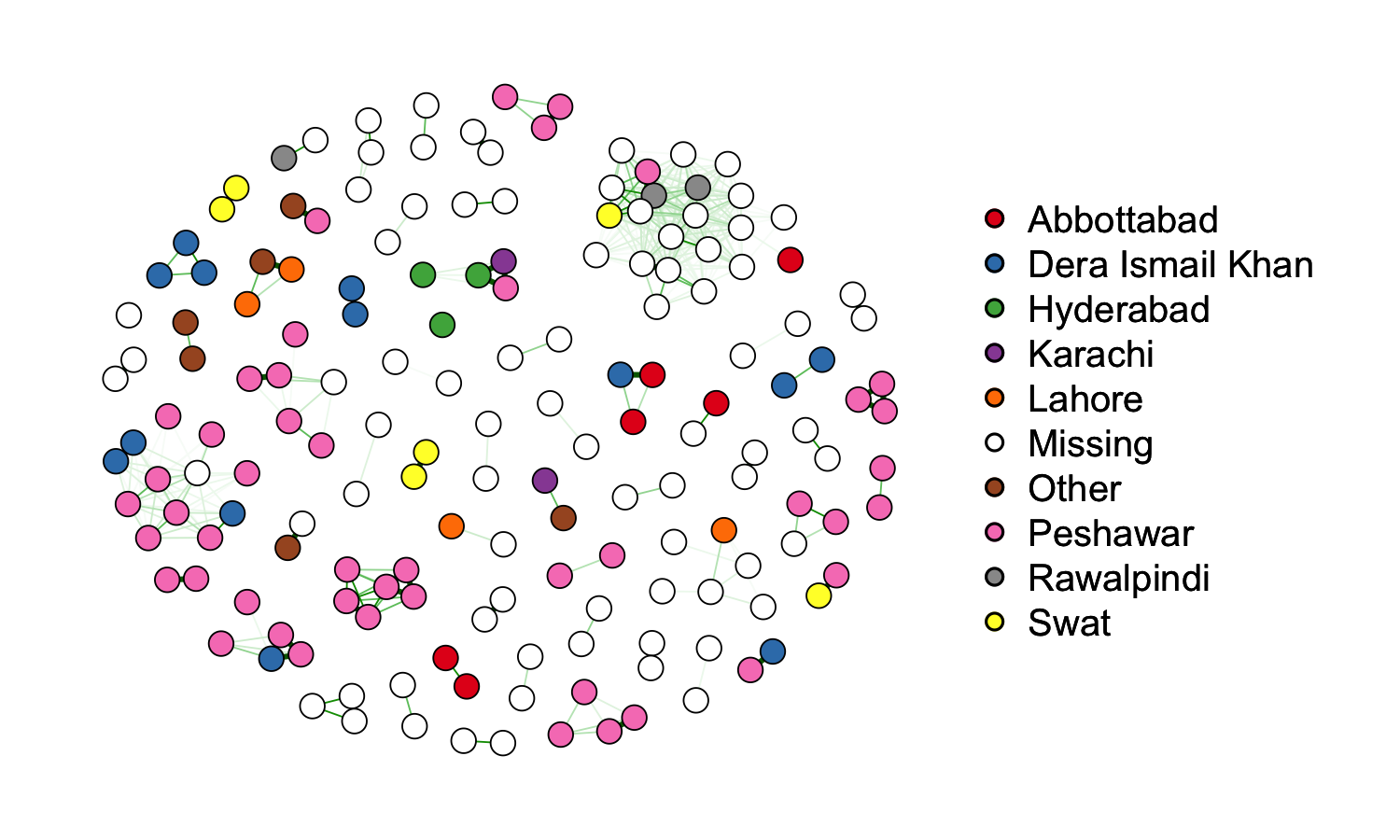
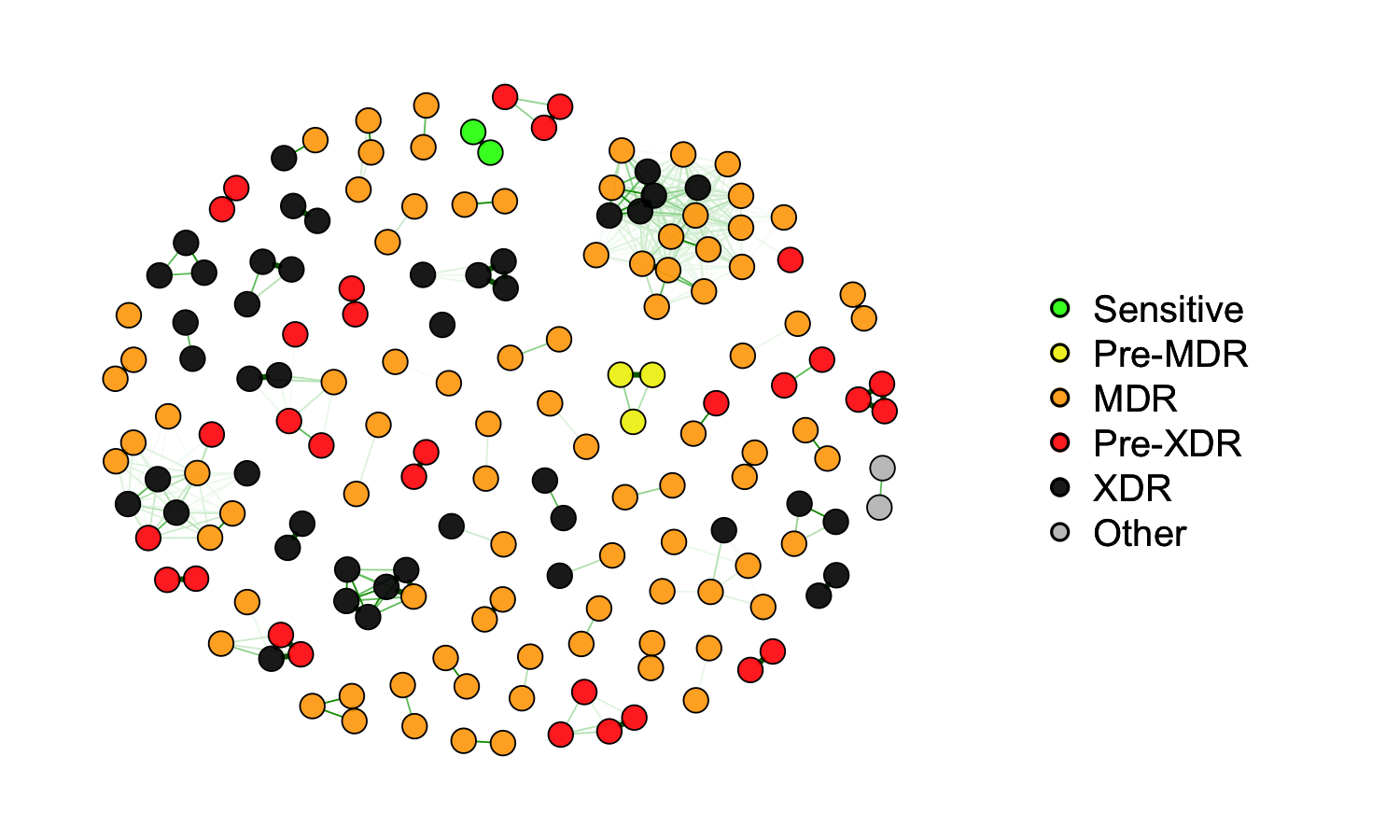
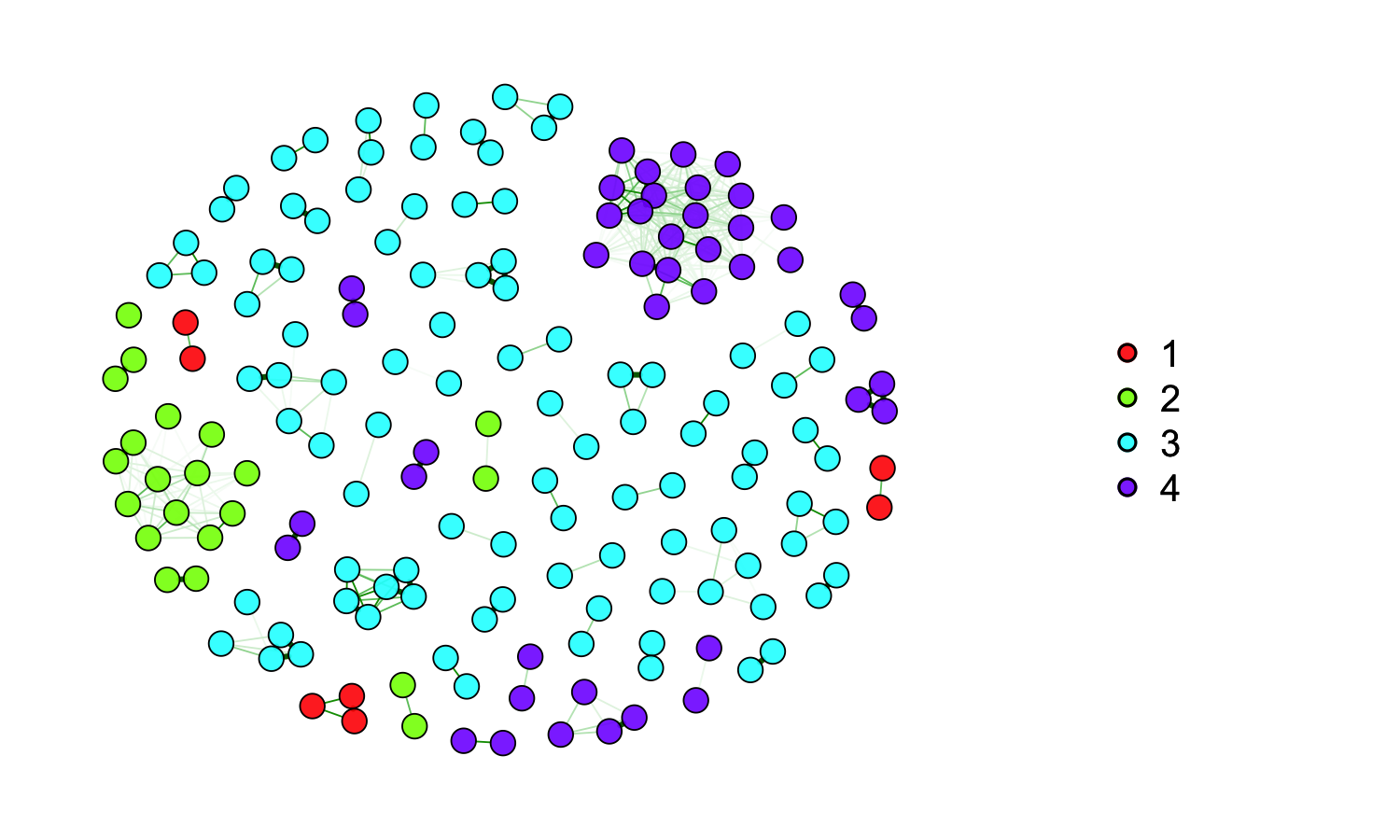
**S2 Figure: SNP distance analyses and clusters (n=535). (top) Density of pairwise SNP differences for all samples; (bottom) number of clustering samples at minimum pairwised SNP difference thresholds**



**S3 Figure: Locations of samples in the transmission chains (n = 163)**



**S4 Figure: The clusters with isolates at <= 10 SNP distance (n = 169), by lineage (top), drug resistance status (middle), and location (bottom).**



**S5 Figure**

**Genome-wide association analysis of transmission. The top-ten scoring genes are labelled.**

Chart

Description automatically generated

**S6 Figure: Phylogenetic location of mutations in nusG gene compared with location of transmission samples.**

**Chart, sunburst chart

Description automatically generated**

### References

1. W.H.O. Global Tuberculosis Report 2020. W.H.O.; 2020. <http://apps.who.int/bookorders.>

2. Phelan J, O’Sullivan DM, Machado D, Ramos J, Whale AS, O’Grady J, et al. The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. Genome Medicine. 2016;8. doi:[10.1186/s13073-016-0385-x](https://doi.org/10.1186/s13073-016-0385-x).

3. Phelan JE, O’Sullivan DM, Machado D, Ramos J, Oppong YE, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Medicine. 2019;11:41. doi:[10.1186/s13073-019-0650-x](https://doi.org/10.1186/s13073-019-0650-x).

4. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Medicine. 2015;7:51. doi:[10.1186/s13073-015-0164-0](https://doi.org/10.1186/s13073-015-0164-0).

5. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome Medicine. 2020;12:114. doi:[10.1186/s13073-020-00817-3](https://doi.org/10.1186/s13073-020-00817-3).

6. Glynn JR, Guerra-Assunção JA, Houben RMGJ, Sichali L, Mzembe T, Mwaungulu LK, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. PLOS ONE. 2015;10:e0132840. doi:[10.1371/JOURNAL.PONE.0132840](https://doi.org/10.1371/JOURNAL.PONE.0132840).

7. Jabbar A, Phelan JE, Sessions PF de, Khan TA, Rahman H, Khan SN, et al. Whole genome sequencing of drug resistant Mycobacterium tuberculosis isolates from a high burden tuberculosis region of North West Pakistan. Scientific Reports. 2019;9:1–9. doi:[10.1038/s41598-019-51562-6](https://doi.org/10.1038/s41598-019-51562-6).

8. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nature Genetics. 2018;50:307–16. doi:[10.1038/s41588-017-0029-0](https://doi.org/10.1038/s41588-017-0029-0).

9. A K, R H, A Z, A A, K I, M A, et al. Alternate efflux pump mechanism may contribute to drug resistance in extensively drug-resistant isolates of Mycobacterium tuberculosis. International journal of mycobacteriology. 2016;5 Suppl 1:S97–8. doi:[10.1016/J.IJMYCO.2016.09.064](https://doi.org/10.1016/J.IJMYCO.2016.09.064).

10. Ali A, Hasan Z, McNerney R, Mallard K, Hill-Cawthorne G, Coll F, et al. Whole Genome Sequencing Based Characterization of Extensively Drug-Resistant Mycobacterium tuberculosis Isolates from Pakistan. PLOS ONE. 2015;10:e0117771. doi:[10.1371/journal.pone.0117771](https://doi.org/10.1371/journal.pone.0117771).

11. The CRyPTIC Consortium and the 100000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. https://doiorg/101056/NEJMoa1800474. 2018;379:1403–15. doi:[10.1056/NEJMOA1800474](https://doi.org/10.1056/NEJMOA1800474).

12. Khan MT, Ali S, Khan A sheed, Ali A, Khan A, Kaushik AC, et al. Insight into the drug resistance whole genome of Mycobacterium tuberculosis isolates from Khyber Pakhtunkhwa, Pakistan. Infection, Genetics and Evolution. 2021;92:104861.

13. Wang B, Artsimovitch I. NusG, an Ancient Yet Rapidly Evolving Transcription Factor. Frontiers in Microbiology. 2021;0:3382.

14. Farhat MR, Jacobson KR, Franke MF, Kaur D, Sloutsky A, Mitnick CD, et al. Gyrase Mutations Are Associated with Variable Levels of Fluoroquinolone Resistance in Mycobacterium tuberculosis. Journal of Clinical Microbiology. 2016;54:727–33. doi:[10.1128/JCM.02775-15](https://doi.org/10.1128/JCM.02775-15).

15. Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. 2014;26:431–44.

16. JA G-A, AC C, RM H, T M, K M, F C, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. eLife. 2015;4. doi:[10.7554/ELIFE.05166](https://doi.org/10.7554/ELIFE.05166).

17. T C, B S, M M. The effect of drug resistance on the fitness of Mycobacterium tuberculosis. The Lancet Infectious diseases. 2003;3:13–21. doi:[10.1016/S1473-3099(03)00483-3](https://doi.org/10.1016/S1473-3099(03)00483-3).

18. Gygli SM, Borrell S, Trauner A, Gagneux S. Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives. FEMS Microbiology Reviews. 2017;41:354–73. doi:[10.1093/FEMSRE/FUX011](https://doi.org/10.1093/FEMSRE/FUX011).

19. Sheen P, Lozano K, Gilman RH, Valencia HJ, Loli S, Fuentes P, et al. pncA gene expression and prediction factors on pyrazinamide resistance in Mycobacterium tuberculosis. Tuberculosis (Edinburgh, Scotland). 2013;93:515. doi:[10.1016/J.TUBE.2013.03.005](https://doi.org/10.1016/J.TUBE.2013.03.005).

20. BP A, AR R, TTY L, KK T, M B, A E, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research. 2020;48:D517–25. doi:[10.1093/NAR/GKZ935](https://doi.org/10.1093/NAR/GKZ935).

21. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. <http://arxiv.org/abs/1303.3997>.

23. H L. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England). 2011;27:2987–93. doi:[10.1093/BIOINFORMATICS/BTR509](https://doi.org/10.1093/BIOINFORMATICS/BTR509).

24. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. doi:[10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).

25. Nguyen L-T, Schmidt HA, Haeseler A von, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32:268–74. doi:[10.1093/molbev/msu300](https://doi.org/10.1093/molbev/msu300).

26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81:559–75.

27. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Medicine. 2016;14:31. doi:[10.1186/s12916-016-0575-9](https://doi.org/10.1186/s12916-016-0575-9).

28. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 2012 44:7. 2012;44:821–4. doi:[10.1038/ng.2310](https://doi.org/10.1038/ng.2310).

29. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490. doi:[10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490).

30. E P, K S. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics (Oxford, England). 2019;35:526–8. doi:[10.1093/BIOINFORMATICS/BTY633](https://doi.org/10.1093/BIOINFORMATICS/BTY633).

31. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012;3:217–23. doi:[10.1111/J.2041-210X.2011.00169.X](https://doi.org/10.1111/J.2041-210X.2011.00169.X).