frontiers | Frontiers in Cellular and Infection Microbiology

TYPE Original Research
PUBLISHED 18 June 2025
DOI 10.3389/fcimb.2025.1573643

OPEN ACCESS

EDITED BY
Nasreen Zafar Ehtesham,
National Institute of Pathology, India

REVIEWED BY
Shahbaz Ahmed,
St. Jude Children's Research Hospital, United States
Anaximandro Gómez-Velasco,
National Polytechnic Institute of Mexico (CINVESTAV), Mexico

*CORRESPONDENCE
Xian-Jin Xie
doctorxxj@163.com
Ting-Ting Wang
1184961771@qq.com

RECEIVED 18 February 2025
ACCEPTED 14 May 2025
PUBLISHED 18 June 2025

CITATION
Yang J-J, Hu Y-l, Sun P-y, Wang L, Xie X-J and Wang T-T (2025) Type VII secretion system gene mutations driving global mycobacterium tuberculosis transmission revealed by whole genomic sequence. *Front. Cell. Infect. Microbiol.* 15:1573643. doi: 10.3389/fcimb.2025.1573643

Type VII secretion system gene mutations driving global mycobacterium tuberculosis transmission revealed by whole genomic sequence

Jian-Jun Yang [1,2], Yuan-long Hu [1], Ping-yi Sun [3], Ling Wang [4], Xian-Jin Xie [2*] and Ting-Ting Wang [2*]

[1]Shandong University of Traditional Chinese Medicine, Jinan, China, [2]Shandong Provincial Third Hospital, Shandong University, Jinan, Shandong, China, [3]College of Integrated Chinese and Western Medicine, Jining Medical University, Jining, China, [4]Intensive Care Unit, People's Hospital of Huaiyin Jinan, Jinan, China

Introduction: Pathogenic mycobacteria are able to transfer virulence factors across their complex cell wall using a type VII secretion system (T7SS)/early secreted antigenic target-6 of the kDa secretion system (ESX). Since the discovery of ESX loci during the Mycobacterium tuberculosis H37Rv genome project, extensive research in areas such as structural biology, cell biology, and evolutionary analysis has improved our understanding of the role of these systems. However, regulatory mechanisms for ESX in Mycobacterium tuberculosis remain elusive. Despite extensive research, the effects of ESX gene mutations on the dynamics of Mycobacterium tuberculosis transmission are not well understood. In this study, we investigated the role of ESX mutations in TB transmission, assessing their risk and characteristics. We analyzed 13582 whole genome sequences of Mycobacterium tuberculosis isolates, of which 6130 (45.13%) were clustered strains. Initially, Boruta algorithm was used to pinpoint SNPs that were significant for TB transmission. These SNPs were then subjected to univariate and multivariate logistic regression analysis to determine the significance of each SNP. The intersection of these two independent methods was recognized as the optimal set of risk mutations for TB transmission. Specifically, we identified one risk mutation (espA(Rv3616c, 4055801)) in L1, four risk mutations (espK(Rv3879c, 4357597), esxU(Rv3445c, 3863138), esxO(Rv2346c, 2626018), and esxW(Rv3620c, 4060588)) in L2, and four risk mutations (eccE1(Rv3882c, 4362807), espE(Rv3864, 4340330), espA (Rv3616c, 4055993), and eccC5(Rv1783, 2019942)) in L4. These risk mutations were significantly associated with clustering, potentially increasing TB transmission. Our findings suggest that mutations in ESX genes play a crucial role in Mycobacterium tuberculosis transmission. These results can be applied to the development of novel strategies for the treatment and prevention of disease.

KEYWORDS

Mycobacterium tuberculosis, mutation, ESX, transmission, phylogenetic analysis

1 Introduction

Tuberculosis (TB) is a widespread infectious disease caused by a pathogen known as Mycobacterium tuberculosis. In recent years, TB cases have decreased worldwide, but the number of people infected with TB each year is still high. In 2022, 10 million people will be infected with TB and 1.3 million will die from TB, causing TB, along with HIV/AIDS. Mycobacteria are adept at evading the host's immune system and establishing infection by engaging with host factors and secreting a range of protein families, including Esx, Esp, and PE/PPE, to exploit the host's nutrients and evade destruction by the immune system (Gröschel et al., 2016; Tufariello et al., 2016; Ates, 2020; Rivera-Calzada et al., 2021). The Mycobacterium tuberculosis genome encodes five specialized secretion systems, referred to as ESX or type VII systems, termed ESX-1 to ESX-5 (Cole et al., 1998). Mycobacterial virulence factors are typically defined as bacterial genes or cellular components that enable their overall survival in the host (Ly and Liu, 2020). ESX systems as well-established virulence factors (Ly and Liu, 2020) are multisubunit apparatuses that have similar structures and secrete related proteins which play a key role in mycobacterial proliferation, pathogenesis, cytosolic escape within macrophages, regulation of macrophage apoptosis, metal ion homeostasis, etc (Majlessi et al., 2015; van Winden et al., 2019; Bunduc et al., 2020a; Famelis et al., 2023). TB transmission is influenced by various factors such as human behavior, virulence of the Mycobacterium tuberculosis pathogen, and host immune responses. Numerous animal and immunological experiments have been carried out using ESX systems to investigate their significance in mycobacterial virulence. These studies indicate that ESX systems play a crucial role in mycobacterial pathogenesis (de Jonge et al., 2007; MacGurn and Cox, 2007; van der Wel et al., 2007; Smith et al., 2008; Sani et al., 2010; Abdallah et al., 2011; Houben et al., 2012; Simeone et al., 2012; Watson and Cox, 2012; Champion et al., 2014; Pajuelo et al., 2021). Thus, variations in ESX systems gene across Mycobacterium tuberculosis lineages may account for differences in TB transmissibility. Here we compared SNPs in the ESX gene region between "clustered" and "non-clustered" isolates in different lineages to identify the role of ESX mutations in TB transmission combining whole genome sequence (WGS) data from 13582 global Mycobacterium tuberculosis isolates collected from 1984-2018.

In recent years, whole genome sequencing (WGS) studies of Mycobacterium tuberculosis have significantly expanded our understanding of this notorious pathogen. The first genome of Mycobacterium tuberculosis was published in 1998, and WGS has since provided a more comprehensive overview of the genomic features of Mycobacterium tuberculosis, identifying specific mutations that help Mycobacterium tuberculosis reduce immune surveillance and drug treatment capabilities. This study used WGS to evaluate the influence of ESX-related gene mutations on the transmission of Mycobacterium tuberculosis and clustering was used to represent the transmission chain of Mycobacterium tuberculosis (Rodríguez et al., 2010).

2 Materials and methods

2.1 Clinical isolates and whole-genome sequencing

We extracted genomic DNA using Cetyltrimethylammonium Bromide (CTAB) from 1468 Mycobacterium tuberculosis samples from Shandong Province over a 5-year period, and 1445 samples passed Quality control(QC). QC was performed using FastQC software to ensure the quality of sequenced reads. The genomes were sequenced using the Illumina HiSeq 4000 system. We also used public databases to compile a global collection of clinical isolates of Mycobacterium tuberculosis, ensuring a diverse and representative collection of genomes with the broadest geographic coverage possible (Luo et al., 2015; Yang et al., 2017; Coll et al., 2018; Hicks et al., 2018; Koster et al., 2018; Liu et al., 2018; Chen et al., 2019; Huang et al., 2019; Jiang et al., 2020). The isolate metadata were downloaded using SRAtools v2.9.1 (https://github.com/ncbi/sra-tools). Only the genomes annotated with sampling date and country of origin were included in the present study. The dataset includes newly sequenced dataset of 1445 Mycobacterium tuberculosis strains and the 12413 Mycobacterium tuberculosis strains collected from 50 countries, a total of 13858 Mycobacterium tuberculosis of isolates. Among the 13858 Mycobacterium tuberculosis isolates, China contributed the most isolates (3408), while Gambia and Moldova contributed the least (1 each). Botswana, Guinea-Bissau, and Sierra-Leone contributed 2 isolates each, Ireland contributed 4 isolates, Switzerland and Malaysia contributed 5 isolates, Mexico, Ghana, and Estonia contributed 6 isolates each, South Africa, Nepal, and Kenya contributed 8 isolates each, Romania and Lebanon contributed 9 isolates each, and several other countries or regions contributed between 11 and 1650 isolates. (Figure 1) We utilized BWA-MEM (version 0.7.17-r1188) to accurately map the reference genome of the standard isolate Mycobacterium tuberculosis H37Rv. Our analysis only included samples exhibiting a coverage rate of 98% or higher and a minimum depth of at least 20% (Jung and Han, 2022). Finally, a total of 13582 genomes were analyzed, please refer to Additional File 1: Supplementary Tables S1, S2 for the specific sample numbers.

2.2 SNP identification

Variant calling was performed using Freebayes (version 1.3.2) and bcftools (version 1.15.1) with a filter parameter 'FMT/GT="1/1" && QUAL>=100 && FMT/DP>=10 && (FMT/AO)/(FMT/DP)>=0'. Single nucleotide polymorphisms in previously defined repetitive regions were excluded, including PPE and PE-PGRS genes, and mobile elements or repeat regions and repeat bases generated by TRF (version 4.09) and Repeatmask (version 4.1.2-p1) (Benson, 1999; Saha et al., 2008; Garrison and Marth, 2012; Danecek et al., 2021). Finally, SNP annotation was conducted via SnpEff v 4.1 l, with the resulting output obtained utilizing the Python programming language (Cingolani et al., 2012).

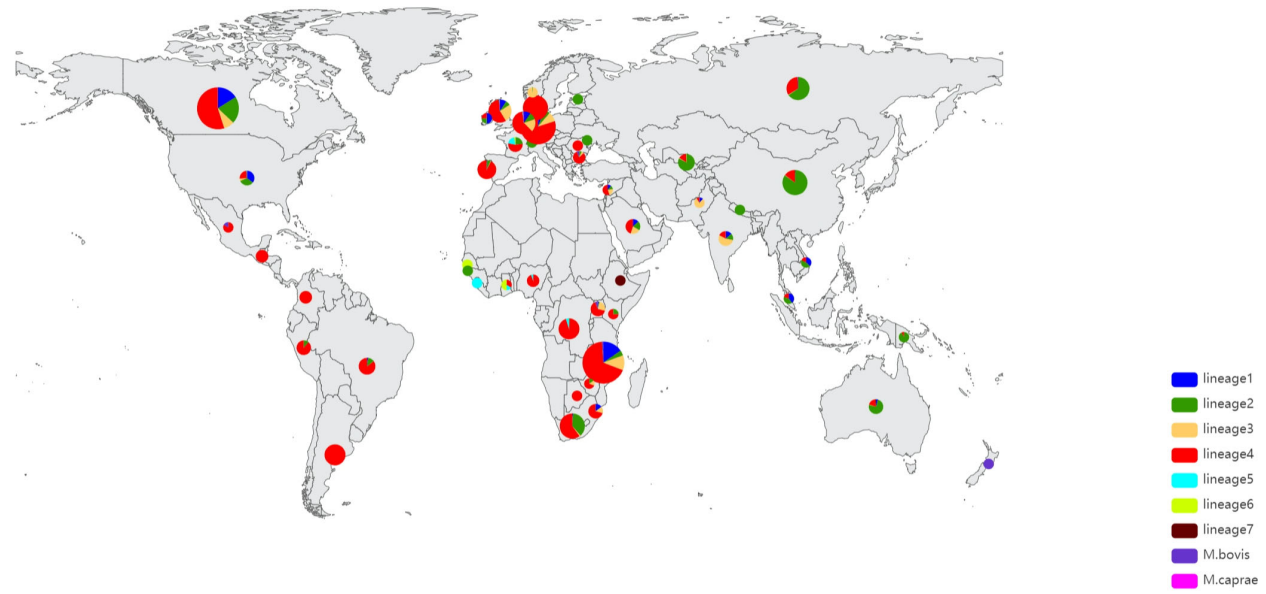

FIGURE 1
Sample size and lineages proportion in different countries of the 13858 isolates.

Genotypic drug resistance of each isolate was predicted in TBProfiler using an established library of mutations (https://github.com/jodyphelan/tbdb) (Coll et al., 2015). The virulence factor database (http://www.mgc.ac.cn/VFs/) contains various medically important bacterial pathogen virulence factors, which include 86 experimentally confirmed and 171 putative genes related to the virulence of Mycobacterium tuberculosis (Cole et al., 1998; Liu et al., 2022). Python was utilized to detect mutations in genes associated with ESXs (Additional File 1: Supplementary Table S3). FASTA sequences of these genes were used to search for corresponding genes in the 12 genomes using BLAST.

2.3 Mtb lineage and genomic cluster

We used the web-based tool TBProfiler (version 4.3.0) to analyze 13858 Mycobacterium tuberculosis WGS data to assign lineages and predict drug resistance (Additional File 1: Supplementary Table S4) (Coll et al., 2014; Phelan et al., 2019). Genomic clusters were ascertained independently of the epidemiological data, and Genomic clusters were inferred based on how genetically similar two isolates were from each other. The upper thresholds of genomic relatedness or cluster is defined as 12 SNPs or alleles cut off or less and a recent transmission event is defined as 5 or less SNPs or alleles (Walker et al., 2013; Kohl et al., 2018). In this study Mycobacterium tuberculosis isolates with a genomic difference (s) ≤ 12 single nucleotide polymorphisms (SNPs) were defined as a genomic cluster (Yang et al., 2017) for further analysis of transmission cluster to avoid missing cases and incorporating recent and old transmission events, which is similar to definitions used in previous genomic studies of Mycobacterium tuberculosis transmission (Walker et al., 2013; Walker et al., 2014; Holt et al., 2018). As suggested by recent analysis of intra-patient variation, the estimate of 5 SNPs may be too low (Lieberman et al., 2016), we finally chose the cut-off of 12 SNPs to define transmission clusters for further analysis based on the previous study (Walker et al., 2013; Walker et al., 2014; Holt et al., 2018). Additionally, according to the classification of transmission clusters by scholars, we also divided transmission clusters into large, medium, or small (large, over 9 isolates; medium, between 3 and 9 isolates; and small, 2 isolates).

2.4 Phylogenetic analysis

Reference genome with only substitution variants instantiated was used as the sample's genome. Maximum-likelihood (ML) phylogenetic trees were constructed and dated by IQ-TREE (v1.6.12) model "JC+I+G4" with 1000 ultrafast bootstrap replicates and treetime (v0.9.0) [GitHub - neherlab/treetime: Maximum likelihood inference of time stamped phylogenies and ancestral reconstruction (Zelner et al., 2016) https://github.com/neherlab/treetime. The trees were constructed using the highest likelihood model selected by automatic model selection in IQ-TREE (v1.6.12), which utilized the JC model of nucleotide substitution and invariable site plus discrete Gamma model of rate heterogeneity to analyze the genome samples with only substitution variants replaced in reference sequence. The resultant phylogenetic tree was visualized through the utilization of iTOL (Letunic, 2021) (https://itol.embl.de/). A maximum likelihood phylogenetic tree was constructed for lineage 1 as shown in Figure 2. Additional tree analysis for lineages 2–7 is available in Additional File 2: Supplementary Figures S2–S7.

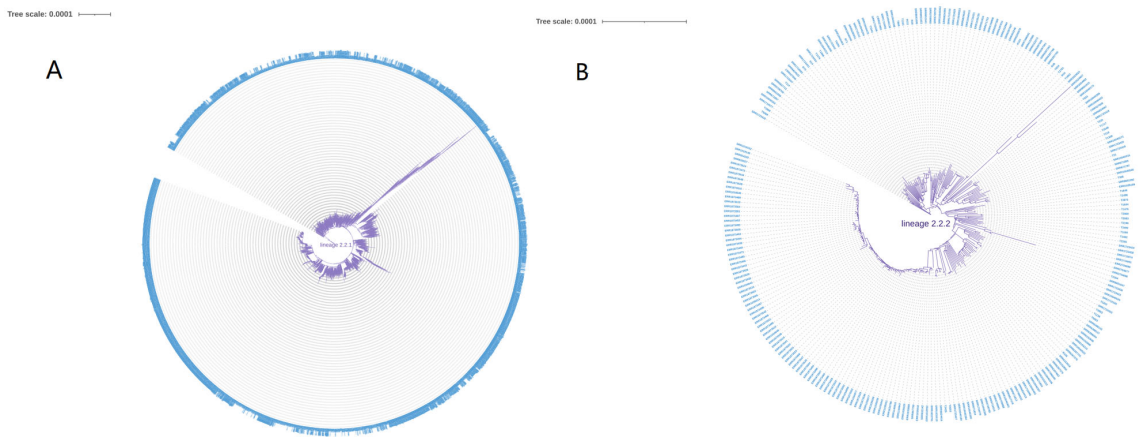

FIGURE 2
(A) The phylogenetic tree analysis of lineage1.1. **(B)** The phylogenetic tree analysis of lineage1.2.

2.5 Independent risk mutations selection

SNPs were found to be associated with clustering, which may potentially enhance TB transmission. These SNPs are referred to as risk mutations. A Boruta algorithm was used in R (version 4.3.0, Boruta package) to select independent risk mutations for TB transmission. The response variable is whether the bacterial strain is clustered or not. The features are these SNPs. The Boruta algorithm has proven to be effective in over 100 studies and is recognized as a premier tool for evaluating large datasets (Saulnier et al., 2011; Degenhardt and Szymczak, 2019). It is a feature selection algorithm based on a random forest classifier (Kursa, 2010; Kursa and Rudnicki, 2010; Degenhardt and Szymczak, 2019). Unlike a general feature selection algorithm, the Boruta feature selection algorithm aims to select the set of features that are most relevant to the dependent variable rather than to a particular model. The Boruta algorithm produces three outcomes for input features, which include confirmed features, tentative features, and rejected features. Thus, the independent risk mutations in this study were screened by the Boruta algorithm.

2.6 Univariate and multivariate/ordinal logistic regression analysis

To verify selected mutations with the Boruta algorithm, all mutations were estimated using univariate and multivariate/ordinal logistic regression analysis. We compared SNPs in the ESX gene region between "clustered" and "non-clustered" isolates using univariate and multivariate/ordinal logistic regression analysis in different lineages.

2.7 Statistical analysis

All statistics were performed with SPSS (version 26) and R software (version 4.2.0). Factors with a P-value less than 0.05 in the final model were considered to be independently associated with genomic clusters. The odds ratios (OR) and 95% confidence intervals (95% CI) were calculated. Due to the limited sample size, we analyzed only the isolates of lineages 1, 2, 3, and 4, excluding the remaining lineages for this study. In addition, in a thorough examination of 481 strains from Denmark (Additional File 1: Supplementary Table S5), we discovered that only twenty SNP variations were present in the ESX gene region, a stark contrast to the plethora of SNPs found in other strains. Therefore, we excluded these 481 Danish strains from our subsequent analysis. We just analyzed nonsynonymous sites and those sites with a mutation frequency higher than 0.02 in lineage1,2,3 and 4. There are 38,16,24 and 26 mutation sites in lineage1,2,3 and 4 respectively, totally 104 mutation sites. The mutation frequency was calculated as the percentage of mutation isolates among total isolates in different lineages (number of mutation isolates/number of total isolates in different lineages). In terms of SNPs, isolates that possess the SNP in the ESX gene region are referred to as mutation isolates.

2.8 Predicted impact of mutations on proteins

Protein prediction algorithm, I-Mutant v2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html), was used to predict the functional impact of noteworthy SNPs on protein structure and function.

2.9 Genomic data availability

The newly sequenced whole genome dataset of 1,445 M. tuberculosis strains was deposited in the National Center for Biotechnology Information (NCBI) under BioProject PRJNA1002108 and 12137 other isolates were downloaded from the public databases. For more details about the 13582 genomes,

please consult Additional File 1: Supplementary Tables S1, S2 containing the specific sample numbers. Additional data can be obtained by contacting the corresponding authors upon request.

3 Results

3.1 Sample description

All seven global lineages (L1-7) of Mycobacterium tuberculosis were detected. Amonge 13582 strains, 851 strains were classified as L1 (6.27%), 5136 strains were assigned to L2 (37.81%), 970 strains belonged to L3 (7.14%), 6489 strains to L4 (47.78%), and only 38 (0.28%), 10 (0.07%), 29 (0.21%), 55(0.40%), 1(0.00%) and 3(0.00%) isolates belonged to L5, 6, 7, M.bovis, M.orygis and M.caprae respectively. Most strains were sublineages 2.2.1 (n=4832, 35.58%), while the remaining sublineages contained less than 1919 strains. For further details, see Table 1.

3.2 Risk mutations associated with genomic clusters

In this study, we detected risk mutations associated with genomic clusters in the four lineages and sublineages of L2.2.1, L4.1, L4.3 and L4.8. The Boruta algorithm was used initially to identify mutations associated with genomic clusters (Figure 3; Additional File 2: Supplementary Figures S8–S11). Next, univariate and multivariate logistic regression analysis were performed to evaluate the significance of each selected mutations (Additional File 1: Supplementary Tables S6-S15). Finally, the intersection results of the two independent methods were identified as the optimal feature variables which were risk mutations associated with genomic clusters, and are summarized

TABLE 1 Characteristics of the 13582 M. tuberculosis.

Characteristic		Classification	No. (%)
Lineage	Lineage1		851(6.27)
	Lineage2		5136(37.81)
	Lineage3		970(7.14)
	Lineage4		6489(47.78)
	Lineage5		38(0.28)
	Lineage6		10(0.07)
	Lineage7		29(0.21)
	M.bovis		55(0.40)
	M.orygis		1(0.00)
	M.caprae		3(0.00)
Sub-lineage	Lineage2.1		46(0.34)
	Lineage2.2.1		4832(35.58)
	Lineage2.2.2		258 (1.90)
	Lineage4.1		1614(11.88)
	Lineage4.2		427(3.14)
	Lineage4.3		1919(14.13)
	Lineage4.4		626(4.61)
	Lineage4.8		1086(7.80)
	Other sub-L4		817(6.02)
Clustered Strains	Lineage1	Clustered strains	148(17.39)
		No-clustered strains	703(82.61)
	Lineage2	Clustered strains	2131(41.50)
		No-clustered strains	3004(58.50)
	Lineage3	Clustered strains	280(28.87)
		No-clustered strains	690(71.13)
	Lineage4	Clustered strains	3124(48.14)
		No-clustered strains	3365(51.86)
Clustered strains _size	Lineage1	Large clustered strains	0(0.00)
		Medium clustered strains	36(4.23)
		Small clustered strains	112(13.16)
		No-clustered strains	704(82.73)
	Lineage2	Large clustered strains	317(6.17)
		Medium clustered strains	797(15.52)
		Small clustered strains	1018(19.82)
		No-clustered strains	3004(58.50)
	Lineage3	Large clustered strains	10(1.03)
		Medium clustered strains	118(12.16)
		Small clustered strains	152(15.67)
		No-clustered strains	690(71.13)
	Lineage4	Large clustered strains	647(9.97)
		Medium clustered strains	1330(20.50)
		Small clustered strains	1148(17.69)
		No-clustered strains	2884(44.44)

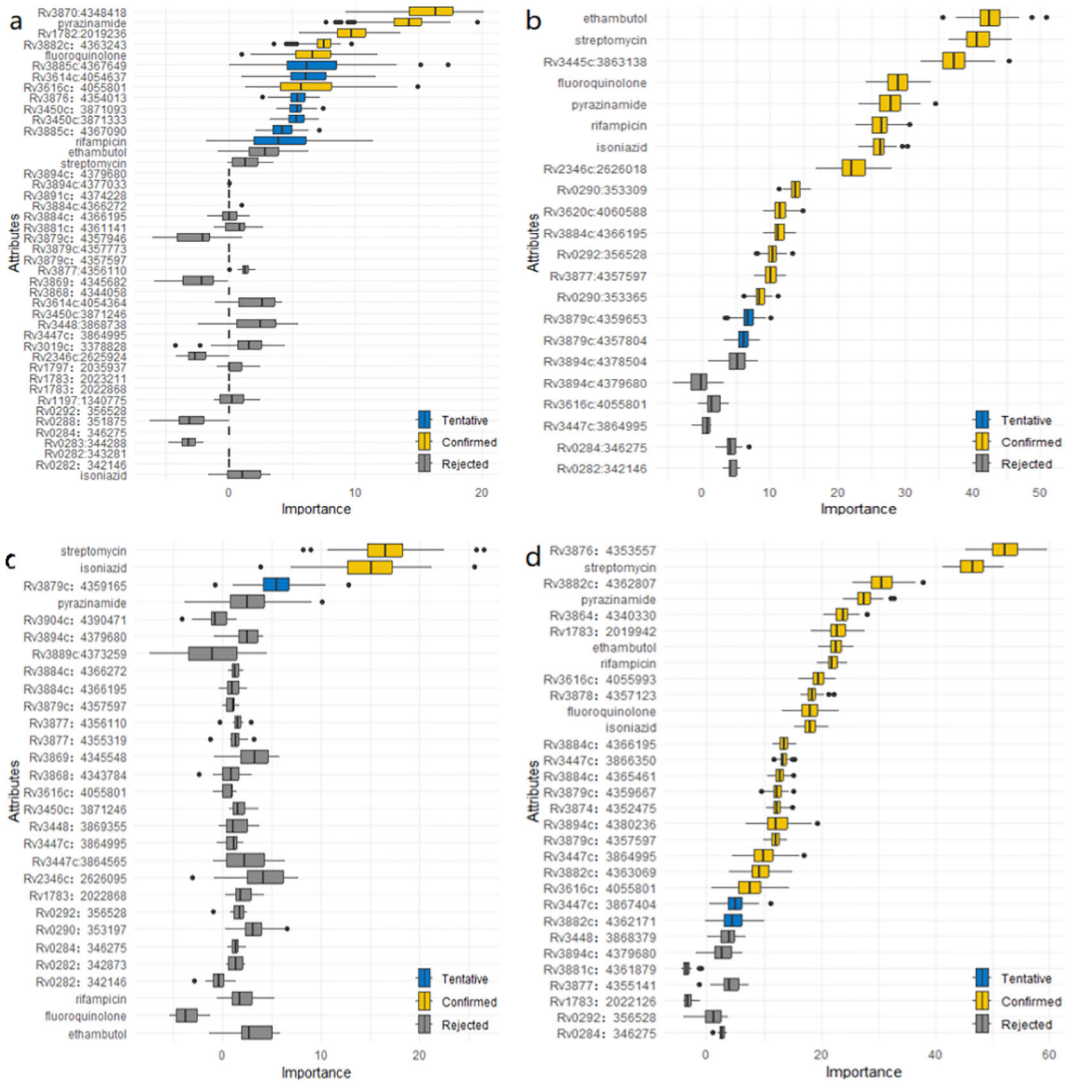

FIGURE 3
(a) Mutations associated with genomic clusters in ESX gene region of lineage 1 identified by the Boruta algorithm. **(b)** Mutations associated with genomic clusters in ESX gene region of lineage 2 identified by the Boruta algorithm. **(c)** Mutations associated with genomic clusters in ESX gene region of lineage 3 identified by the Boruta algorithm. **(d)** Mutations associated with genomic clusters in ESX gene region of lineage 4 identified by the Boruta algorithm.

in Figure 4; Additional File 2: Supplementary Figures S12-S15. Specifically, there was one risk mutation[espA(Rv3616c, 4055801)] in L1, four risk mutations [esxU(Rv3445c, 3863138), esxO (Rv2346c, 2626018), esxW(Rv3620c, 4060588) and espK(Rv3879c, 4357597)] in L2, four risk mutations[espE(Rv3864, 4340330), espl (Rv3876, 4353557), eccC5(Rv1783, 2019942) and eccE1(Rv3882c, 4362807)] in L4. Three risk mutations [esxU(Rv3445c, 3863138), esxO(Rv2346c, 2626018) and eccA2(Rv3884c, 4366195)] in L2.2.1, one risk mutation [espK(Rv3879c, 4359667)]in L4.1 and one risk mutation [eccE1 (Rv3882c, 4362171)]in L4.8.

In addition, the intersection results in antimicrobial resistance mutations of the two independent methods were also identified in different lineages (Figure 4; Additional File 2: Supplementary Figures S12–S15). Antimicrobial resistance mutation at L1, L2, and L4 was determined to be associated with an elevated clustering risk. Specifically, a single mutation was noted for fluoroquinolones at L1, two mutations were found for streptomycin and ethambutol at L2, and two mutations were identified for pyrazinamide and streptomycin at L4. Three mutations (streptomycin, ethambutol and pyrazinamide) in L2.2.1, two mutations (pyrazinamide and streptomycin) in L4.1 and three mutations (fluoroquinolones, pyrazinamide and streptomycin) in L4.3 were associated with genomic clusters. Mutations occurred mainly in drug resistance genes such as katG, rpoB, rpsL, embB, pncA, gyrA, and ethA. Drug resistance is an important factor of TB transmission. In our study, we mainly used the Drug resistance mutations as exposure factors in multivariate logistic regression analysis to improve the sensitivity of analysis results.

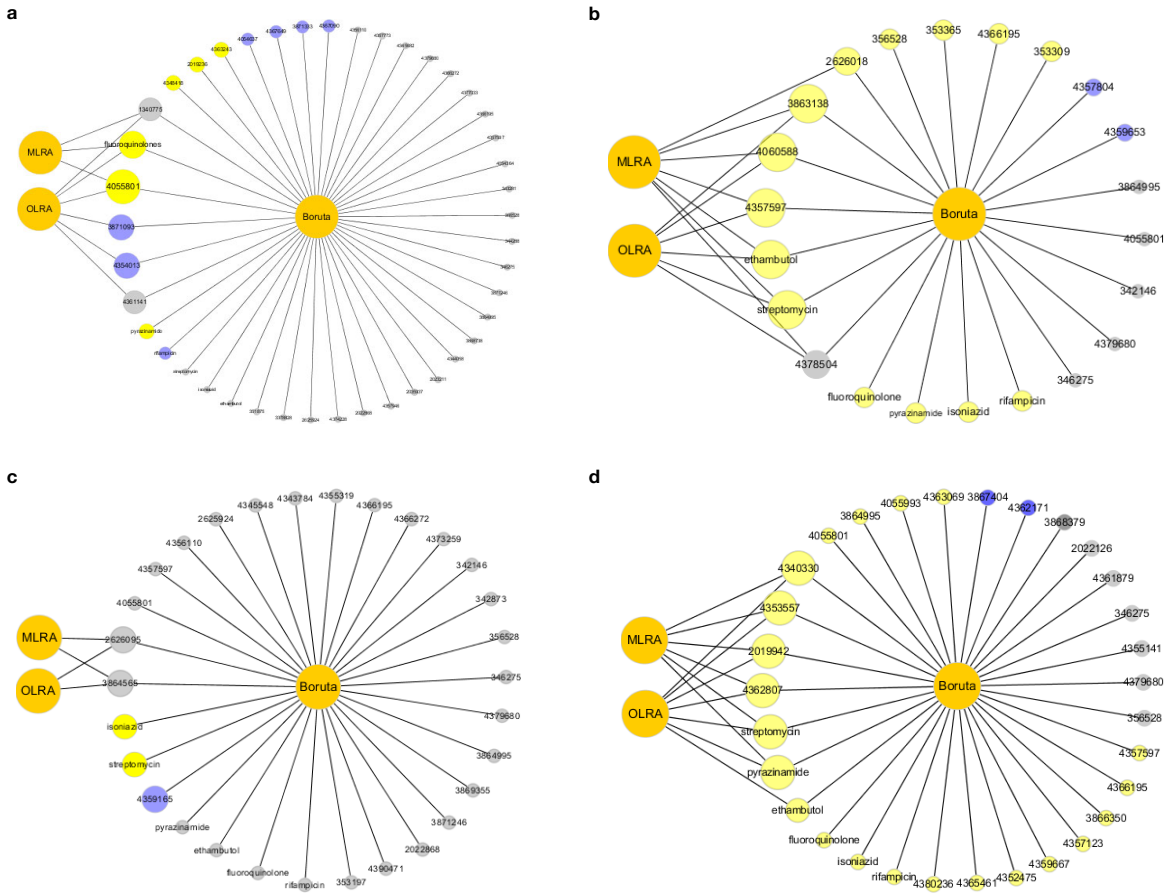

FIGURE 4
(a) The intersection results of Lineage 1. The grey color means reject in the Boruta algorithm. The blue color means tentative in the Boruta algorithm. The yellow color means confirm in the Boruta algorithm. MLRA was the abbreviation of Multivariate Logistic Regression Analysis. OLRA was the abbreviation of Ordinal Logistic Regression Analysis. If the circle connected with MLRA or OLRA, it means the SNPs were risk mutations in MLRA or OLRA. **(b)** The intersection results of Lineage 2. **(c)** The intersection results of Lineage 3. **(d)** The intersection results of Lineage 4.

3.2.1 Risk mutations associated with genomic clusters selected by Boruta algorithm

By comparing original mutations (attributes) importance with importance achievable at random, the results of Boruta algorithm for L1, L2, L3 and L4 were shown in Figure 3.

In Figure 3a, six mutations (Rv3870:4348418, Rv1782:2019236, Rv3882c:4363243, Rv3616c:4055801, fluoroquinolone and pyrazinamide) were confirmed as important feature in lineage 1. An additional seven mutations (Rv3885c:4367649, Rv3614c:4054637, Rv3450c:3871333, Rv3876:4354013, Rv3450c:3871093, Rv3885c:4367090 and rifampicin) were tentative features, while the remaining mutations were rejected features.

Similarly, In Figure 3b, indicates that 14 mutations (Rv3445c:3863138, Rv2346c:2626018, Rv0290:353309, Rv3884c:4366195, Rv3620c:4060588, Rv0292:356528, Rv3879c:4357597, Rv0290:353365, ethambutol, streptomycin, fluoroquinolone, pyrazinamide, rifampicin, and isoniazid) were confirmed as important feature in lineage 2. Two additional mutations (Rv3879c:4359653, Rv3879c:4357804) were tentative features, while the remainder were rejected features.

In Figure 3c, two mutations (streptomycin and isoniazid) were confirmed as important features in lineage 3. One additional mutation (Rv3879c:4359165) was tentative feature, while the remainder were rejected features.

Finally, Figure 3d shows that 22 mutations (Rv3876:4353557, Rv3882c:4362807, Rv3864:4340330, Rv1783:2019942, Rv3616c:4055993, Rv3878:4357123, Rv3884c:4366195, Rv3879c:4357597, Rv3447c:3866350, Rv3884c:4365461, Rv3894c:4380236, Rv3874:4352475, Rv3879c:4359667, Rv3447c:3864995, Rv3882c:4363069, Rv3616c:4055801, streptomycin, pyrazinamide, ethambutol, rifampicin, fluoroquinolone and isoniazid) were confirmed as important features in lineage 4. Two additional mutations (Rv3882c:4362171 and Rv3447c:3867404) were tentative features, while the remainder were rejected features.

In addition, we conducted a detailed analysis of the L2 and L4 sublineages. As demonstrated in Additional File 2: Supplementary Figures S8–S11, mutations in L2.2.1, L4.1, L4.3 and L4.8 are a total of 22, 18, 14 and 10, respectively, which have been verified as follows:

L2.2.1: twelve mutations (Rv3445c:3863138, Rv2346c:2626018, Rv0290:353309, Rv3620c:4060588, Rv3879c:4357597, Rv3884c:4366195, ethambutol, streptomycin, pyrazinamide, rifampicin, fluoroquinolone and isoniazid) were confirmed. Three

(Rv0290:353365, Rv0292:356528 and Rv3894c:4378504) mutations were tentative and the rest were rejected.

L4.1: 11 mutations (Rv3882c:4362807, Rv3879c:4359667, Rv3874:4352475, Rv3878:4357123, Rv3884c:4366195, streptomycin, pyrazinamide, rifampicin, ethambutol, isoniazid and fluoroquinolone) were confirmed and the rest were rejected.

L4.3: nine mutations (Rv1783:2019942, Rv3864:4340330, Rv3882c:4363069, streptomycin, fluoroquinolone, pyrazinamide, ethambutol, rifampicin and isoniazid) were confirmed. One (Rv3894c:4379680) was tentative and four were rejected.

L4.8: Two mutation (Rv3882c:4362171 and streptomycin) were confirmed and all other mutations were rejected.

These confirmed mutations have been identified as factors associated with genomic clusters.

3.2.2 Risk mutations associated with genomic clusters selected by logistic regression analysis

To verify selected mutations with the Boruta algorithm, all 104 mutations were analyzed in univariate logistic regression analysis and any variable with a P value <0.2 in the univariate logistic regression analysis was included in the subsequent multivariate logistic regression analysis.

The analysis comprised 38 mutation sites in ESX genes in L1. Among the clustered and non-clustered strains of L1, 12 ESX gene mutation sites showed statistically significant differences (P<0.05), which showed these ESX gene sites were associated with the clustering of L1 when compared with non-clustered strains of L1 (Additional File 1: Supplementary Table S6). Then, 14 mutation sites of ESX genes and three drug resistance genes with P<0.2 in univariate analysis were analyzed by multivariate regression. The results revealed that four mutations in ESX gene sites and two antimicrobial resistance mutations had a significant impact on the clustering of L1 (P<0.05), see Additional File 1: Supplementary Table S12, including two ESX mutation sites [espA (Rv3616c:4055801, OR, 5.053; 95% CI, 1.965-12.998), esxK (Rv1197:1340775, OR, 2.303; 95% CI, 1.134,4.680)] and one antimicrobial resistance mutation [fluoroquinolone (OR, 11.616; 95% CI, 2.420-55.769)] that were identified as risk factors for clustering.

The analysis focused on 16 ESX gene mutation sites in L2. In the comparison between clustered and non-clustered strains of lineage2, the difference in the mutation of nine ESX gene sites was statistically significant (P<0.05). The specific results can be found in (Additional File 1: Supplementary Table S7). Subsequently, all the 9 ESX gene mutation sites and drug resistant sites with P<0.2 in univariate analysis were included in a multivariate regression analysis, which showed that five mutation sites and two antimicrobial resistance mutations that were identified as risk factors for clustering, see Additional File 1: Supplementary Table S13, including esxU (Rv3445c:3863138, OR, 1.566; 95% CI, 1.327-1.848), esxO (Rv2346c:2626018, OR, 1.224; 95% CI, 1.083-1.383), esxW (Rv3620c:4060588, OR, 6.170; 95% CI, 1.375, 27.686), espK (Rv3879c:4357597, OR, 9.249; 95% CI, 1.093, 78.251), eccC2 (Rv3894c:4378504, OR, 3.669; 95% CI, 1.360, 9.900), ethambutol (OR, 2.310; 95% CI, 1.884, 2.832) and streptomycin (OR, 1.576; 95% CI, 1.310, 1.897).

24 mutation sites of ESX genes in L3 were analyzed. In the comparison between clustered and non-clustered strains, four ESX gene mutation sites showed significant differences (P<0.05), as detailed in Additional File 1: Supplementary Table S8. Subsequently, five mutation sites of ESX genes and one antimicrobial resistance mutation with P<0.2 were included in multiple regression analysis, and the results showed that 3 mutation sites of ESX genes that were identified as risk factors for clustering, see Additional File 1: Supplementary Table S14, including esxO (Rv2346c:2626095, OR, 14.519; 95% CI, 1.966-107.210) and eccC4 (Rv3447c:3864565, OR, 2.089; 95% CI, 1.377-3.170) (Additional File 1: Supplementary Table S13).

We analyzed 26 mutation sites in L4. Comparing clustered and non-clustered strains, there were 15 mutations in ESX gene sites with statistical significance (P<0.05), as detailed in Additional File 1: Supplementary Table S9. Subsequently, 18 ESX mutation sites and 6 antimicrobial resistance mutations with P<0.2 were analyzed by multivariate regression, and finally seven ESX gene sites and two antimicrobial resistance mutations with significant influence on clustering were determined(P<0.05), as shown in Additional File 1: Supplementary Table S15. Four ESX gene sites and two antimicrobial resistance mutations were risk factors for clustering, which were espE (Rv3864:4340330; OR, 2.203; 95% CI, 1.749-2.775), espA (Rv3876:4353557; OR, 21.020; 95% CI, 11.903-37.120), eccC5 (Rv1783:2019942; OR, 1.630; 95% CI, 1.209-2.198), eccE1 (Rv3882c:4362807; OR, 1.579; 95% CI, 1.008-2.473), pyrazinamide (OR, 1.760; 95% CI, 1.374, 2.256) and streptomycin (OR, 1.450; 95% CI, 1.199, 1.753).

The same analysis of L2.2.1, L4.1, L4.3 and L4.8, as detailed in Additional File 1: Supplementary Tables S10, S16. There were six risk mutations in L2.2.1. They were esxU (Rv3445c:3863138, OR, 1.722; 95% CI, 1.413-2.098), esxO (Rv2346c:2626018, OR, 1.252; 95% CI, 1.105-1.419), eccA2 (Rv3884c:4366195, OR, 10.571; 95% CI, 1.309-85.350), eccC2 (Rv3894c:4378504, OR, 2.928; 95% CI, 1.144-7.493), ethambutol (OR, 1.918; 95% CI, 1.575-2.336) and streptomycin (OR, 1.363; 95% CI, 1.183-2.837). Significantly, the SNP at Rv2346c:2626018 was found to be P<0.05 in the multivariate logistic regression analysis for L2, while the P value was 0.055 for L2.2.1. Similarly, the SNP at Rv3884c:4366195 was shown to be statistically significant (P<0.05) in L2.2.1, although its P value was 0.08 in L2. Therefore, it is clear that these two SNPs (Rv3884c:4366195 and Rv2346c:2626018) are potential significant mutations, and further investigation should be conducted. Two risk mutations in L4.1. They were streptomycin (OR, 1.832; 95% CI, 1.575-2.336) and pyrazinamide (OR, 2.469; 95% CI, 1.383-4.408). Three risk mutations in L4.3. They were streptomycin (OR, 1.997; 95% CI, 1.448-2.754), fluoroquinolone (OR, 2.172; 95% CI, 1.430-3.299) and pyrazinamide (OR, 1.684; 95% CI, 1.119-2.535). One risk mutation [eccA2 (Rv3882c:4362171, OR, 1.725; 95% CI, 1.232-2.415)] in L4.8.

3.3 Sensitivity analysis

In the sensitivity analysis, the lineage 1, lineage 2, lineage 3 and lineage 4 data were divided into four groups and then reanalyzed using Boruta algorithm and ordinal regression analysis. As shown

in Table 1, the first, second, third, and fourth group included non-clustered isolates, small clusters containing two isolates, medium clusters containing 3 to 9 isolates, and large clusters containing >9 isolates, respectively. Only the mutations with a P value <0.2 in the univariate logistic regression analysis was included in the ordinal regression analysis.

The results of the Boruta algorithm were shown in Figure 5. As shown in Figure 5a, ten mutations (Rv3870:4348418, Rv1782:2019236, Rv3616c:4055801, Rv3882c:4363243, Rv3614c:4054637, Rv3450c:3871333, Rv3876:4354013, pyrazinamide, fluoroquinolone and streptomycin) were confirmed in lineage 1. Five mutations (Rv3450c:3871093, Rv3885c:4367649, Rv3885c:4367090, Rv3614c:4054364 and isoniazid) were tentative. The rest were rejected. As shown in Figure 5b, fourteen mutations (Rv3445c:3863138, Rv2346c:2626018, Rv0290:353309, Rv3620c:4060588, Rv3884c:4366195, Rv3879c:4357597, Rv0292:356528, Rv0290:353365, ethambutol, streptomycin, fluoroquinolone, pyrazinamide, rifampicin and isoniazid) were confirmed in lineage 2. Two mutations (Rv3879c:4359653 and Rv3879c:4357804) were tentative. The rest were rejected. As shown in Figure 5c, three mutations (streptomycin, isoniazid and ethambutol) were confirmed in lineage 3. One mutation (pyrazinamide) was tentative. The rest were rejected. As shown in Figure 5d, twenty-three mutations (Rv3882c:4362807, Rv3876:4353557, Rv1783:2019942, Rv3878:4357123, Rv3884c:4365461, Rv3879c:4359667, Rv3879c:4357597, Rv3874:4352475, Rv3447c:3866350, Rv3864:4340330, Rv3882c:4363069, Rv3881c:4361879, Rv0292:356528, Rv1783:2022126, Rv3616c:4055993, Rv3884c:4366195, Rv3894c:4379680, streptomycin, pyrazinamide, isoniazid, ethambutol, rifampicin and fluoroquinolone) were confirmed in lineage 4. One mutation (Rv3616c:4055801) was tentative. The rest

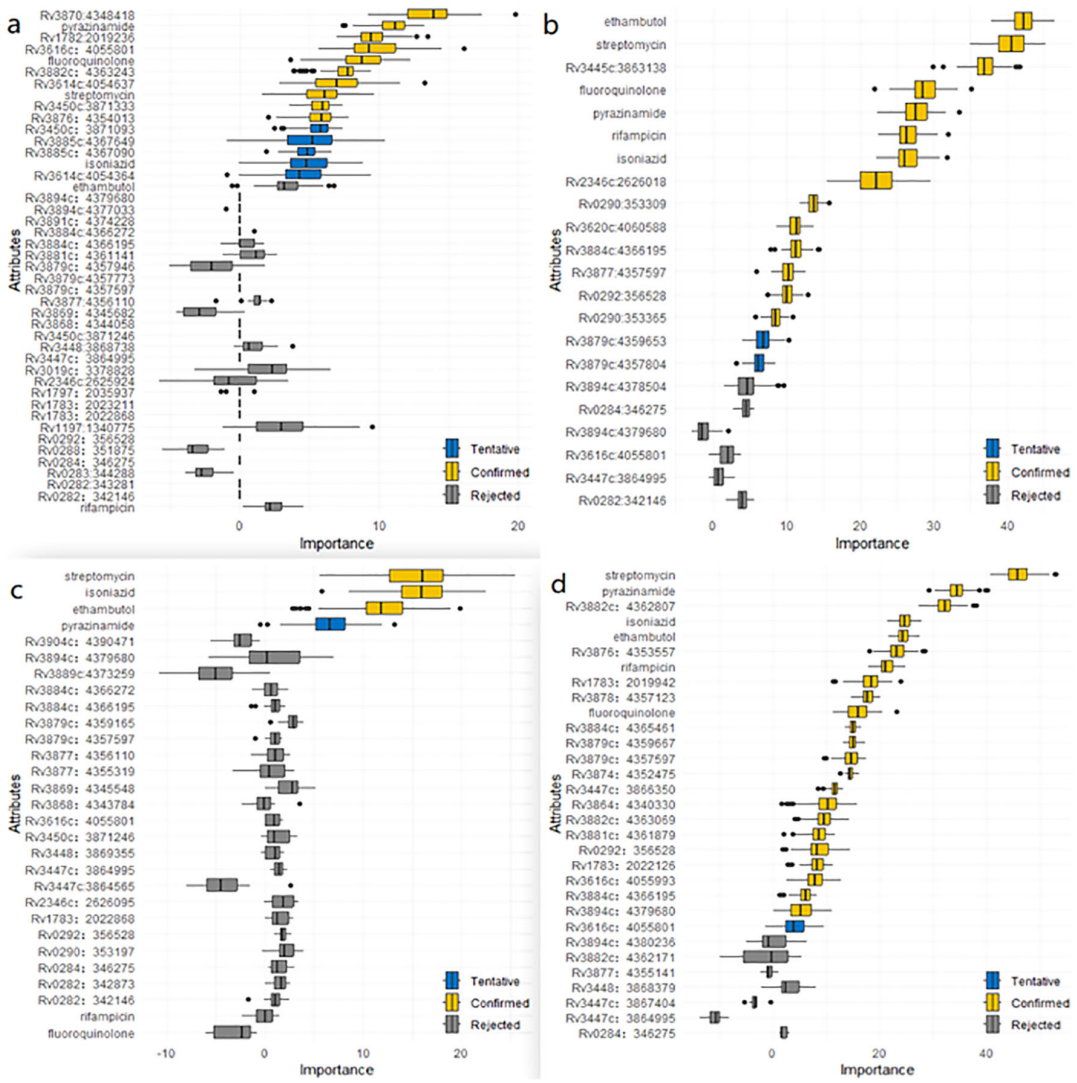

FIGURE 5
(a) Mutations associated with layered genomic clusters in ESX gene region of lineage 1 identified by the Boruta algorithm. **(b)** Mutations associated with layered genomic clusters in ESX gene region of lineage 2 identified by the Boruta algorithm. **(c)** Mutations associated with layered genomic clusters in ESX gene region of lineage 3 identified by the Boruta algorithm. **(d)** Mutations associated with layered genomic clusters in ESX gene region of lineage 4 identified by the Boruta algorithm.

were rejected. The results of ordinal regression analysis were shown in Additional File 1: Supplementary Tables S17-S20. There were six risk mutations in L1. They were eccB4 (Rv3450c:3871093, OR, 204944.568; 95% CI, 92661.384-453287.808), espA (Rv3616c:4055801, OR, 6.22; 95% CI, 2.327-16.623), espl (Rv3876:4354013, OR, 88974.512; 95% CI, 40227.955-196790.114), esxK (Rv1197:1340775, OR, 2.028; 95% CI, 1.005-4.093), espB (Rv3881c:4361141, OR, 514221.196; 95% CI, 232493.784-1137335.516) and fluoroquinolone (OR, 17.686; 95% CI, 4.299-72.765). There were six risk mutations in L2. They were esxU (Rv3445c:3863138, OR, 1.667; 95% CI, 1.42-1.958), esxW (Rv3620c:4060588, OR, 6.569; 95% CI, 1.457-29.607), eccC2 (Rv3894c:4378504, OR, 3.493; 95% CI, 1.279-9.538), espK (Rv3879c:4357597, OR, 9.102; 95% CI, 1.081-76.626) ethambutol (OR, 2.618; 95% CI, 2.158-3.175) and streptomycin (OR, 1.604; 95% CI, 1.346-1.911). The results showed that 2 mutation sites of ESX genes in L3 that were identified as risk factors for clustering, including esxO (Rv2346c:2626095, OR, 14.309; 95% CI, 1.938-105.672) and eccC4 (Rv3447c:3864565, OR, 2.29; 95% CI, 1.512-3.467). There were seven risk mutations in L4. These were espE (Rv3864:4340330; OR, 2.11; 95% CI, 1.692-2.631), eccE1 (Rv3882c:4362807; OR, 2.024; 95% CI, 1.307-3.137), eccC5 (Rv1783:2019942; OR, 1.404; 95% CI, 1.092-1.804), espA (Rv3876:4353557; OR, 35.653; 95% CI, 21.81-58.284), pyrazinamide (OR, 2.244; 95% CI, 1.801-2.797), streptomycin (OR, 1.412; 95% CI, 1.188-1.677) and ethambutol (OR, 1.491; 95% CI, 1.167-1.905).

As shown in Figure 4, the intersection results of the two independent methods were identified. Specifically, there was two risk mutations[espA(Rv3616c:4055801) and fluoroquinolone] in L1, five risk mutations [esxU(Rv3445c:3863138), esxW (Rv3620c:4060588), espK(Rv3879c:4357597), streptomycin and ethambutol], in L2, six risk mutations[espE(Rv3864:4340330), espl (Rv3876:4353557), eccE1(Rv3882c:4362807), eccC5 (Rv1783:2019942), streptomycin and pyrazinamide] in L4. The sensitivity analysis results did not change significantly compared to those of the Boruta algorithm and multivariate regression analysis. The results of ordinal regression analysis based on the size of clustered isolates were like the main findings: one risk mutation[espA(Rv3616c:4055801)] in L1, four risk mutations [esxU(Rv3445c:3863138), esxO(Rv2346c:2626018), esxW (Rv3620c:4060588) and espK(Rv3879c:4357597)] in L2, four risk mutations[espE(Rv3864:4340330), espl (Rv3876:4353557), eccC5 (Rv1783:2019942) and eccE1(Rv3882c:4362807)] in L4.

3.4 Deleterious effect of risk mutations on proteins

Nine SNPs[espA(Rv3616c:4055801), espK(Rv3879c:4357597), espE(Rv3864:4340330), espl(Rv3876:4353557), eccE1 (Rv3882c:4362807), eccC5(Rv1783:2019942), esxU (Rv3445c:3863138), esxO(Rv2346c:2626018) and esxW (Rv3620c:4060588)] in the ESX gene region were predicted to negatively affect the respective proteins that affect the protein instability in nearby structural areas (Table 2).

4 Discussion

From the analysis of genetic diversity, it was determined that the majority of these isolates belonged to lineage 4. While lineage 2 contributed a significant proportion, lineage 3 and lineage 1 were less prevalent. As illustrated in Figure 1, Lineage 2, also known as the East Asian lineage, including the Beijing family of strains, is primarily located in East Asia. However, it is also found in Central Asia, Russia, and South Africa. Lineage 4, also known as the Euro-American lineage, is frequently detected in individuals from Asia, Europe, Africa, and America. Lineages 1 and 3 were identified in regions in East Africa, South and Southeast Asia, Europe, and North America. On the other hand, lineages 5 and 7 are geographically more restricted and are generally confined to specific regions of Africa. These data are congruent with previous observations (Gagneux, 2007; Comas et al., 2009; Coscolla, 2014; Bañuls et al., 2015; Zheng et al., 2017; Gagneux, 2018; Koster et al., 2018).

Those risk mutations [espA(Rv3616c:4055801), espK (Rv3879c:4357597), espE(Rv3864:4340330), espl(Rv3876:4353557), eccE1(Rv3882c:4362807), eccC5(Rv1783:2019942), esxU (Rv3445c:3863138), esxO(Rv2346c:2626018) and esxW (Rv3620c:4060588)] were also predicted to negatively affect the respective proteins that affect the protein instability in nearby structural areas, supporting the hypothesis that they may affect TB transmission. Further research conducted through a range of biological and biochemical approaches has highlighted the critical role these genes play in Mycobacterium tuberculosis virulence. These findings have been confirmed in various animal and cellular models, demonstrating the fundamental role of these genes in Mycobacterium tuberculosis virulence (de Jonge et al., 2007; MacGurn and Cox, 2007; van der Wel et al., 2007; Smith et al., 2008; Sani et al., 2010; Abdallah et al., 2011; Houben et al., 2012; Simeone et al., 2012; Watson and Cox, 2012; Champion et al., 2014; Pajuelo et al., 2021). In addition, the findings of this study concur with previous genomic epidemiological articles (Mahairas et al., 1996; Hsu et al., 2003; Guinn et al., 2004; Wirth et al., 2012; Pang et al., 2013; Pajuelo et al., 2021).

The ESX protein complex plays a crucial role in the physiology, cell envelope integrity, conjugation, and host-pathogen interactions of mycobacteria. In order to evaluate the impact of mutations in the ESX gene on the global transmission of TB, a comprehensive analysis of 13582 strains of Mycobacterium tuberculosis, including 62 ESX genes, was performed. This study discovered ten risk mutations in the ESX gene regions that have the potential to boost TB transmission. These mutations are occurred in espA(Rv3616c:4055801), espK (Rv3879c:4357597), espE(Rv3864:4340330), espl(Rv3876:4353557), eccE1(Rv3882c:4362807), eccC5(Rv1783:2019942), esxU (Rv3445c:3863138), esxO(Rv2346c:2626018) and esxW (Rv3620c:4060588) gene region. EspA, espK, espE and espl are ESX-1 substrates. EccE1 is an essential component of the ESX-1 secretion system. esxU, esxO and esxW are ESAT-6 family members. EccC5 is predicted to be component of the ESX-5-membrane-associated complex.

ESX-1 is the prototype of type VII secretion systems (Hsu et al., 2003; Lewis et al., 2003; Pym et al., 2003), which is considered a major virulence factor of Mycobacterium tuberculosis through its

essential role in phagosomal rupture and subsequent translocation of the pathogen to the host cytosol (Stamm et al., 2003; van der Wel et al., 2007; Houben et al., 2012; Simeone et al., 2012). Our study showed that five SNPs [espA(Rv3616c:4055801), espK (Rv3879c:4357597), espE(Rv3864:4340330), espl(Rv3876:4353557) and eccE1(Rv3882c:4362807)] in ESX-1 gene region were associated with clustering which could improve the TB transmission. EspA, espK, espE and espl are ESX-1 substrates which are virulence factors of Mycobacterium tuberculosis.

Research studies have indicated that disrupting disulfide bond formation within EspA could lead to a strain retaining ESX-1 function, while exhibiting significantly reduced virulence (Garces et al., 2010). Moreover, the secretion of EspA and EsxA is mutually dependent, with the deletion of espA resulting in the loss of EsxA secretion and subsequent attenuation (Fortune et al., 2005). A study also demonstrated that the espK gene is an essential player in preventing the formation of mature phagolysosomes and antigen presentation by host macrophages, and inhibiting espK expression could lead to a synergistic reduction in virulence and pathogenesis of Mycobacterium tuberculosis (Mishra et al., 2019).

The EspE substrate is found tethered to the mycobacterial cell surface and secreted into the extracellular environment *in vitro* (Carlsson et al., 2009; van der Woude et al., 2012). Recently, it has been reported that EspE is essential for the secretion of EsxA and plays a critical role in virulence within a macrophage infection model. Chirakos, Alexandra E et al. demonstrate that EspE is required for the lytic activity of the ESX-1 system and functions within the mycobacterial cytoplasm to negatively regulate the transcriptional activity of the WhiB6 protein, thereby modulating the production levels of ESX-1 substrates (Chirakos et al., 2020; Lienard et al., 2020).

EspI was not the subject of a comprehensive study. However, a limited body of research suggests that EspI may be involved in restraining the ESX-1-regulated secretion of bacteria when cellular ATP levels are low (Zhang et al., 2014). In addition, it is thought that the down-regulation of EspI may be critical for Mycobacterium tuberculosis's ability to persist in chronic infections, as it maintains ATP levels and promotes virulence (Kruh et al., 2010; Zhang et al., 2014). The protein EccE1 is an indispensable component of the mycobacterial ESX-1 secretion system, which is crucial for the process of virulence factor secretion. EccE1 was initially believed to function as the inner membrane pore unit of a membrane complex, facilitating the transport of various substrates (Abdallah et al., 2007). In Mycobacterium tuberculosis, scientists discovered that the deletion of the eccE1 gene reduced the levels of EccB1, EccCa1 and EccD1, thereby abolishing ESX-1 secretion and reducing Mycobacterium tuberculosis ex vivo virulence (Soler-Arnedo et al., 2020). Additionally, in Mycobacterium smegmatis, the homolog of EccE1 was found to be required for the secretion of EsxA and EsxB (Converse, 2005). This indicates that EccE1 is critical for ex vivo virulence, stabilization of ESX-1 membrane proteins, and the secretion of EsxA, EsxB, EspA, and EspC. From the function of ESX-1substrates, a notable feature of the ESX-1 secretion system is the mutually dependent nature of protein export; i.e., the secretion of each substrate relies on the secretion of the other. Therefore, these SNPs could potentially modify the functionality of these proteins, subsequently impacting the secretion of one another and the overall virulence of Mycobacterium tuberculosis.

The Early Secreted Antigenic Target 6 (ESAT-6) family proteins are a group of small, spiral-structured proteins. These molecules are exported out of the cell via the ESX secretion system (Pallen, 2002). This family comprises a total of 23 members, labeled EsxA-W (Cole et al., 1998). These proteins have a pivotal role in host-pathogen interactions, serve as immunodominant antigens in the recognition of the human immune system, with the majority being immunodominant T cell antigens, playing a critical role in Mycobacterium tuberculosis pathogenesis and individual immune protection mechanism (Skjøt et al., 2000; Uplekar et al., 2011).

We found three SNPs [esxU(Rv3445c:3863138), esxO (Rv2346c:2626018) and esxW(Rv3620c:4060588)] in ESAT-6 family proteins have the potential to enhance the TB transmission. Previous research indicates that esxU can form a stable helical complex with esxT and has notable immunogenicities, characterized by significant lymphocyte proliferation and the induction of tumor necrosis factor-α (TNF-α) and interleukin-6 (IL-6) (Pandey et al., 2018). Currently esxW is an emerging vaccine candidate under investigation for inclusion in several TB vaccines (Baldwin et al., 2009; Bertholet et al., 2010; Knudsen et al., 2014; Baldwin et al., 2015), as it has demonstrated the ability to stimulate an immune response in mice, demonstrated safety and efficacy in non-human primates (Bertholet et al., 2010), and specifically targets T cells in humans (Bertholet et al., 2008). In one study, researchers identified an alteration in EsxW that could potentially contribute to enhanced transmission of Mycobacterium tuberculosis from the Beijing lineage (Holt et al.,

TABLE 2 Deleterious effect of risk mutations on ESX proteins†.

Genomic position*	Nucleotide change	Amino acid change	stability
Rv3616c_4055801	G=>A	T192I	Large decrease of stability
Rv3879c_4357597	C=>G	C729S	Large decrease of stability
Rv3864_4340330	T=>G	L21V	Large decrease of stability
Rv3876_4353557	C=>T	P183L	Large decrease of stability
Rv3882c_4362807	A=>G	V205A	Large decrease of stability
Rv3445c_3863138	G=>A	P43S	Large decrease of stability
Rv2346c_2626018	T=>C	E52G	Large decrease of stability
Rv3620c_4060588	T=>C	T2A	Large decrease of stability
Rv1783_2019942	A=>G	Q229R	Large decrease of stability

†Functional impact of the risk mutations on protein structure and function was predicted on one protein prediction algorithms, I-Mutant v2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html).

2018). Lastly, the esxO gene in Mycobacterium tuberculosis is involved in virulence by enhancing survival within macrophages, by inhibiting the production of cytokines such as TNF-α and IL-6 (Mohanty et al., 2016; Yao et al., 2018).

The ESX-5 secretion system is critical for PPE protein secretion, cell wall stability and virulence; it is also critical for the uptake of nutrients across the outer membrane (Bottai et al., 2012; Ates et al., 2015). In this study, we discovered a single nucleotide polymorphism (SNP) in the EccC5 gene that might significantly elevate the odds of contracting Mycobacterium tuberculosis (Mtb). EccC5, a membrane-bound ATPase protein, is hypothesized to play a crucial role in the formation of the ESX-5 membrane-associated Mycobacterium tuberculosis complex. Ex vivo experiments have demonstrated that EccB5 and EccC5 encoding genes are parts of an operon and are indispensable for the survival and progression of Mycobacterium tuberculosis (Di Luca et al., 2012). The creation of a conditional mutant strain (MtbPptreccC5), in which the expression of the eccC5 gene was regulated by an anhydrotetracycline-repressible promoter, confirmed that eccC5 gene suppression is detrimental to the growth of Mycobacterium tuberculosis both *in vitro* and within human THP-1 macrophage cells (Di Luca et al., 2012). Further analysis of the secretome of Mycobacterium tuberculosis PptreccC5 strains revealed that EccC5 is required for the secretion of ESX-5-specific substrates, thus confirming that EccC5 is indeed a component of the ESX-5 secretion machinery (Di Luca et al., 2012). Moreover, a recent study has generated an M. marinum- Mycobacterium tuberculosis EccC5 chimera, demonstrating that the secretion specificity of PE_PGRS proteins in both M. marinum and Mycobacterium tuberculosis is reliant on the presence of EccC5 cognate linker 2 domain (Bunduc et al., 2020b).

In conclusion, this study provides statistical evidence that ten SNPs in Mycobacterium tuberculosis ESX genes may be associated with disease progression and increased transmission of certain strains. These SNPs may alter the protein's function, further impacting adaptive responses by influencing the structure of nearby domains and triggering gene expression, ultimately influencing Mycobacterium tuberculosis transmission. It is important to note that while we have established the impact of these SNPs in ESX on the transmission of Mycobacterium tuberculosis, animal and immunological experiments should be conducted to gain further biological evidence and help us better comprehend the molecular biology characteristics of Mycobacterium tuberculosis. Research into the virulence of mycobacteria has demonstrated that the process of mutation is not primarily a result of the accumulation of nonsynonymous mutations, but more so, several crucial mutations that affect the activity of specific gene products (Hershberg et al., 2008; Mikheecheva et al., 2017). These proteins are produced by ten specific genes and are critical to the ability of the pathogen to survive in the host's hostile environment. When these proteins are eliminated or deficient in biological experiments, Mycobacterium tuberculosis loses its virulence within the host (Fortune et al., 2005; Garces et al., 2010; Kruh et al., 2010; Di Luca et al., 2012; Zhang et al., 2014; Mohanty et al., 2016; Yao et al., 2018; Soler-Arnedo et al., 2020), highlighting the significant role that these proteins play in the pathogen's life cycle. The findings from this study suggest that the ESX protein is a vital component of Mycobacterium tuberculosis ' life activities. Furthermore, this study offers several potential methods to intervene in the production or function of the ESX proteins, offering a wealth of potential targets for the design of novel antimycobacterial drugs. As such, the ten specific mutations identified in the ESX protein could potentially be considered as promising targets for future anti-tuberculosis therapies. Moreover, Due to the limitations of strain collection, the results of this study are only related to the whole set of strains used in the study.

5 Strength and limitations

This study presents several limitations. First, we did not perform animal and immunological experiments to provide biological validation for the risk mutations identified herein. Second, we lack critical host factors that may influence disease transmissibility, such as age, host immune status, and pulmonary cavitation. This absence hinders our ability to account for confounding variables that could elucidate the independent effects of risk mutations on transmissibility. Finally, contribution of analyzed strains is biased to one region (China). This has an impact in the biological inferences from a globally distributed human pathogen. In our subsequent research, we should endeavor to conduct more pertinent work. Of course, the sample size is large enough. The risk mutations we found were more reliable, which could provide credible data for TB prevention and treatment.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

The manuscript presents research on animals that do not require ethical approval for their study. No potentially identifiable images or data are presented in this study.

Author contributions

J-JY: Conceptualization, Supervision, Writing – original draft. Y-LH: Data curation, Formal analysis, Methodology, Software, Writing – original draft. P-YS: Formal analysis, Investigation, Validation, Writing – original draft. LW: Investigation, Software, Writing – original draft. X-JX: Conceptualization, Data curation, Formal analysis, Writing – review & editing. T-TW: Data curation, Methodology, Writing – original draft, Writing – review & editing, Formal analysis.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Department of Science & Technology of Shandong Province (CN) (Nos. 2007GG30002033 and 2017GSF218052), Natural Science Foundation of Shandong Province (CN) (No. ZR2020KH013 and ZR2021MH006), and Jinan Science and Technology Bureau (CN) (No. 201704100).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2025.1573643/full#supplementary-material

SUPPLEMENTARY TABLE 1
Information of 1445 strains of mycobacterium tuberculosis.

SUPPLEMENTARY TABLE 2
Information of 12137 strains of mycobacterium tuberculosis.

SUPPLEMENTARY TABLE 3
Information of gene mutations in ESX.

SUPPLEMENTARY TABLE 4
Six antimicrobial resistance mutations information of 13582 strains of mycobacterium tuberculosis.1 means resistant. 0 means sensitive.

SUPPLEMENTARY TABLE 5
SNP information of 13582 strains of mycobacterium tuberculosis in different countries.

SUPPLEMENTARY TABLE 6
Univariate regression analysis on SNPs associated with clustering in ESX gene region of lineage 1. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962. Mutation isolates: In terms of SNPs, isolates that possess the mutation in the esx gene region are referred to as mutation isolates. The clustering rate was calculated as the percentage of cluster mutation isolates among total mutation isolates in L1 (number of cluster mutation isolates/number of total mutation isolates in L1). The mutation frequency was calculated as the percentage of mutation isolates among the number of total isolates in L1.

SUPPLEMENTARY TABLE 7
Univariate regression analysis on SNPs associated with clustering in ESX gene region of lineage 2. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962. Mutation isolates: In terms of SNPs, isolates that possess the mutation in the esx gene region are referred to as mutation isolates. The clustering rate was calculated as the percentage of cluster mutation isolates among total mutation isolates in L2(number of cluster mutation isolates/number of total mutation isolates in L2). The mutation frequency was calculated as the percentage of mutation isolates among the number of total isolates in L2.

SUPPLEMENTARY TABLE 8
Univariate regression analysis on SNPs associated with clustering in ESX gene region of lineage 3. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962. Mutation isolates: In terms of SNPs, isolates that possess the mutation in the esx gene region are referred to as mutation isolates. The clustering rate was calculated as the percentage of cluster mutation isolates among total mutation isolates in L3 (number of cluster mutation isolates/number of total mutation isolates in L3). The mutation frequency was calculated as the percentage of mutation isolates among the number of total isolates in L3.

SUPPLEMENTARY TABLE 9
Univariate regression analysis on SNPs associated with clustering in ESX g.

SUPPLEMENTARY TABLE 12
Multivariable regression analysis on SNPs associated with clustering in ESX gene region of lineage 1. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962. * means that there is no result in statistical software or the result was too large and nonsense. OR, odds ratio.

SUPPLEMENTARY TABLE 13
Multivariable regression analysis on SNPs associated with clustering in ESX gene region of lineage 2. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 14
Multivariable regression analysis on SNPs associated with clustering in ESX gene region of lineage 3. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 15
Multivariable regression analysis on SNPs associated with clustering in ESX gene region of lineage 4. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 16
Multivariable regression analysis on SNPs associated with clustering in ESX gene region of sublineages. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 17
Ordinal regression analysis on SNPs associated with clustering in ESX gene region of lineage 1. Small clusters containing two isolates, medium clusters containing 3 to 9 isolates, and large clusters containing >9 isolates, respectively. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 18
Ordinal regression analysis on SNPs associated with clustering in ESX gene region of lineage 2. Small clusters containing two isolates, medium clusters containing 3 to 9 isolates, and large clusters containing >9 isolates, respectively. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 19
Ordinal regression analysis on SNPs associated with clustering in ESX gene region of lineage 3. Small clusters containing two isolates, medium clusters containing 3 to 9 isolates, and large clusters containing >9 isolates, respectively. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY TABLE 20
Ordinal regression analysis on SNPs associated with clustering in ESX gene region of lineage 4. Small clusters containing two isolates, medium clusters containing 3 to 9 isolates, and large clusters containing >9 isolates, respectively. Genomic position are genomic nucleotide positions in Mtb H37Rv genome NC_000962.

SUPPLEMENTARY FIGURE 1
(a) The phylogenetic tree analysis of lineage1.1. **(b)** The phylogenetic tree analysis of lineage1.2.

SUPPLEMENTARY FIGURE 2
(a) The phylogenetic tree analysis of lineage2.2.1. **(b)** The phylogenetic tree analysis of lineage2.2.2.

SUPPLEMENTARY FIGURE 3
(a) The phylogenetic tree analysis of lineage3. **(b)** The phylogenetic tree analysis of lineage3.1.

SUPPLEMENTARY FIGURE 4
(a) The phylogenetic tree analysis of lineage4.1. **(b)** The phylogenetic tree analysis of lineage4.2. **(c)** The phylogenetic tree analysis of lineage4.3. **(d)** The phylogenetic tree analysis of lineage4.4. **(e)** The phylogenetic tree analysis of lineage4.8.

SUPPLEMENTARY FIGURE 5
The phylogenetic tree analysis of lineage5.

SUPPLEMENTARY FIGURE 6
The phylogenetic tree analysis of lineage6.

SUPPLEMENTARY FIGURE 7
The phylogenetic tree analysis of lineage7.

SUPPLEMENTARY FIGURE 8
Mutations associated with genomic clusters in ESX gene region of lineage 2.2.1 identified by the Boruta algorithm. The yellow color represents confirmed feature. The blue color represents tentative feature and the grey color represents rejected feature.

SUPPLEMENTARY FIGURE 9
Mutations associated with genomic clusters in ESX gene region of lineage 4.1 identified by the Boruta algorithm. The yellow color represents confirmed feature. The blue color represents tentative feature and the grey color represents rejected feature.

SUPPLEMENTARY FIGURE 10
Mutations associated with genomic clusters in ESX gene region of lineage 4.3 identified by the Boruta algorithm. The yellow color represents confirmed feature. The blue color represents tentative feature and the grey color represents rejected feature.

SUPPLEMENTARY FIGURE 11
Mutations associated with genomic clusters in ESX gene region of lineage 4.8 identified by the Boruta algorithm. The yellow color represents confirmed feature. The blue color represents tentative feature and the grey color represents rejected feature.

SUPPLEMENTARY FIGURE 12
The intersection results of Lineage 2.2.1. The grey color means reject in the Boruta algorithm. The blue color means tentative in the Boruta algorithm. The yellow color means confirm in the Boruta algorithm. MLRA was the abbreviation of Multivariate Logistic Regression Analysis. OLRA was the abbreviation of Ordinal Logistic Regression Analysis. If the circle connected with MLRA or OLRA, it means the SNPs were risk mutations in MLRA or OLRA.

SUPPLEMENTARY FIGURE 13
The intersection results of Lineage 4.1. The grey color means reject in the Boruta algorithm. The blue color means tentative in the Boruta algorithm. The yellow color means confirm in the Boruta algorithm. MLRA was the abbreviation of Multivariate Logistic Regression Analysis. OLRA was the abbreviation of Ordinal Logistic Regression Analysis. If the circle connected with MLRA or OLRA, it means the SNPs were risk mutations in MLRA or OLRA.

SUPPLEMENTARY FIGURE 14
The intersection results of Lineage 4.3. The grey color means reject in the Boruta algorithm. The blue color means tentative in the Boruta algorithm. The yellow color means confirm in the Boruta algorithm. MLRA was the abbreviation of Multivariate Logistic Regression Analysis. OLRA was the abbreviation of Ordinal Logistic Regression Analysis. If the circle connected with MLRA or OLRA, it means the SNPs were risk mutations in MLRA or OLRA.

SUPPLEMENTARY FIGURE 15
The intersection results of Lineage 4.8. The grey color means reject in the Boruta algorithm. The blue color means tentative in the Boruta algorithm. The yellow color means confirm in the Boruta algorithm. MLRA was the abbreviation of Multivariate Logistic Regression Analysis. OLRA was the abbreviation of Ordinal Logistic Regression Analysis. If the circle connected with MLRA or OLRA, it means the SNPs were risk mutations in MLRA or OLRA.

References

Abdallah, A. M., Champion, P. A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C. M., Appelmelk, B. J., et al. (2007). Type VII secretion–mycobacteria show the way. *Nat. Rev. Microbiol.* 5, 883–891. doi: 10.1038/nrmicro1773

Abdallah, A. M., Savage, N. D., de Punder, K., van Zon, M., Wilson, L., Korbee, C. J., et al. (2011). Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation. *J. Immunol.* 187, 4744–4753. doi: 10.4049/jimmunol.1101457

Ates, L. S. (2020). New insights into the mycobacterial PE and PPE proteins provide a framework for future research. *Mol. Microbiol.* 113, 4–21. doi: 10.1111/mmi.v113.1

Ates, L. S., Commandeur, S., van de Weerd, R., Sparrius, M., Weerdenburg, E., Alber, M., et al. (2015). Essential role of the ESX-5 secretion system in outer membrane permeability of pathogenic mycobacteria. *PloS Genet.* 11, e1005190. doi: 10.1371/journal.pgen.1005190

Baldwin, S. L., Huang, P. W., Beebe, E. A., Podell, B. K., Reed, S. G., and Coler, R. N. (2015). Protection and Long-Lived Immunity Induced by the ID93/GLA-SE Vaccine Candidate against a Clinical Mycobacterium tuberculosis Isolate. *Clin. Vacc. Immunol.* 23, 137–147. doi: 10.1128/CVI.00458-15

Baldwin, S. L., Kahn, M., Zharkikh, I., Ireton, G. C., Vedvick, T. S., Reed, S. G., et al. (2009). Intradermal immunization improves protective efficacy of a novel TB vaccine candidate. *Vaccine* 27, 3063–3071. doi: 10.1016/j.vaccine.2009.03.018

Bañuls, A. L., Van Anh, N. T., and Godreuil, S. (2015). Mycobacterium tuberculosis: ecology and evolution of a human bacterium. *J. Med. Microbiol.* 64, 1261–1269. doi: 10.1099/jmm.0.000171

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bertholet, S., Kahn, M., Guderian, J., Mohamath, R., Stride, N., Laughlin, E. M., et al. (2008). Identification of human T cell antigens for the development of vaccines against Mycobacterium tuberculosis. *J. Immunol.* 181, 7948–7957. doi: 10.4049/jimmunol.181.11.7948

Bertholet, S., Ordway, D. J., Windish, H. P., Pine, S. O., Kahn, M., Phan, T., et al. (2010). A defined tuberculosis vaccine candidate boosts BCG and protects against multidrug-resistant Mycobacterium tuberculosis. *Sci. Transl. Med.* 2, 53ra74. doi: 10.1126/scitranslmed.3001094

Bottai, D., Majlessi, L., Frigui, W., Simeone, R., Sayes, F., Bitter, W., et al. (2012). Disruption of the ESX-5 system of Mycobacterium tuberculosis causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. *Mol. Microbiol.* 83, 1195–1209. doi: 10.1111/j.1365-2958.2012.08001.x

Bunduc, C. M., Bitter, W., and Houben, E. N. G. (2020a). Structure and function of the mycobacterial type VII secretion systems. *Annu. Rev. Microbiol.* , 74:315–74:335. doi: 10.1146/annurev-micro-012420-081657

Bunduc, C. M., Bitter, W., and Houben, E. N. G. (2020b). Species-specific secretion of ESX-5 type VII substrates is determined by the linker 2 of EccC5. *Mol. Microbiol.* 114, 66–76. doi: 10.1111/mmi.v114.1

Carlsson, F., Rangell, L., and Brown, E. J. (2009). Polar localization of virulence-related Esx-1 secretion in mycobacteria. *PloS Pathog.* 5, e1000285. doi: 10.1371/journal.ppat.1000285

Champion, M. M., Pinapati, R. S., and Champion, P. A. (2014). Correlation of phenotypic profiles using targeted proteomics identifies mycobacterial esx-1 substrates. *J. Proteome Res.* 13, 5151–5164. doi: 10.1021/pr500484w

Chen, X., Wang, S., Lin, S., Chen, J., and Zhang, W. (2019). Corrigendum: evaluation of whole-genome sequence method to diagnose resistance of 13 anti-tuberculosis drugs and characterize resistance genes in clinical multi-drug resistance mycobacterium tuberculosis isolates from China. *Front. Microbiol.* , 10:2221. doi: 10.3389/fmicb.2019.02221

Chirakos, A. E., Huffman, A., and Champion, P. A. (2020). Conserved ESX-1 substrates espE and espF are virulence factors that regulate gene expression. *Infect. Immun.* 88, e00289–e00220. doi: 10.1128/IAI.00289-20

Cingolani, P., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695

Cole, S. T., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nat. Ecol. Evol.* 393, 537–544. doi: 10.1038/31159

Coll, F., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., Portugal, I., et al. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* 5, 4812. doi: 10.1038/ncomms5812

Coll, F., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., Abdallah, A. M., et al. (2018). Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* 50, 307–316. doi: 10.1038/s41588-017-0029-0

Coll, F., Preston, M. D., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorne, G., Mallard, K., et al. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* 7, 51. doi: 10.1186/s13073-015-0164-0

Comas, I., Niemann, S., and Gagneux, S. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PloS One* 4, e7815. doi: 10.1371/journal.pone.0007815

Converse, S. E. (2005). A protein secretion pathway critical for Mycobacterium tuberculosis virulence is conserved and functional in Mycobacterium smegmatis. *J. Bacteriol.* 187, 1238–1245. doi: 10.1128/JB.187.4.1238-1245.2005

Coscolla, M. (2014). Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin. Immunol.* 26, 431–444. doi: 10.1016/j.smim.2014.09.012

Danecek, P., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008. doi: 10.1093/gigascience/giab008

Degenhardt, F., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform.* 20, 492–503. doi: 10.1093/bib/bbx124

de Jonge, M. I., Fretz, M. M., Romain, F., Bottai, D., Brodin, P., Honoré, N., et al. (2007). ESAT-6 from Mycobacterium tuberculosis dissociates from its putative chaperone CFP-10 under acidic conditions and exhibits membrane-lysing activity. *J. Bacteriol.* 189, 6028–6034. doi: 10.1128/JB.00469-07

Di Luca, M., Batoni, G., Orgeur, M., Aulicino, A., Counoupas, C., Campa, M., et al. (2012). The ESX-5 associated eccB-EccC locus is essential for Mycobacterium tuberculosis viability. *PloS One* 7, e52059. doi: 10.1371/journal.pone.0052059

Famelis, N., Geibel, S., and van Tol, D. (2023). Mycobacterial type VII secretion systems. *Biol. Chem.* 404, 691–702. doi: 10.1515/hsz-2022-0350

Fortune, S. M., Sarracino, D. A., Chase, M. R., Sassetti, C. M., Sherman, D. R., Bloom, B. R., et al. (2005). Mutually dependent secretion of proteins required for mycobacterial virulence. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10676–10681. doi: 10.1073/pnas.0504922102

Gagneux, S. (2007). Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infect. Dis.* 7, 328–337. doi: 10.1016/S1473-3099(07)70108-1

Gagneux, S. (2018). Ecology and evolution of Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* 16, 202–213. doi: 10.1038/nrmicro.2018.8

Garces, A., Chase, M. R., Woodworth, J. S., Krastins, B., Rothchild, A. C., Ramsdell, T. L., et al. (2010). EspA acts as a critical mediator of ESX1-dependent virulence in Mycobacterium tuberculosis by affecting bacterial cell wall integrity. *PloS Pathog.* 6, e1000957. doi: 10.1371/journal.ppat.1000957

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *Quantit. Biol.*

Gröschel, M. I., Simeone, R., Majlessi, L., and Brosch, R. (2016). ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat. Rev. Microbiol.* 14, 677–691. doi: 10.1038/nrmicro.2016.131

Guinn, K. M., Mathur, S. K., Zakel, K. L., Grotzke, J. E., Lewinsohn, D. M., Smith, S., et al. (2004). Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of Mycobacterium tuberculosis. *Mol. Microbiol.* 51, 359–370. doi: 10.1046/j.1365-2958.2003.03844.x

Hershberg, R., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., Roach, J. C., et al. (2008). High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PloS Biol.* 6, e311. doi: 10.1371/journal.pbio.0060311

Hicks, N. D., Zhang, X., Zhao, B., Grad, Y. H., Liu, L., Ou, X., et al. (2018). Clinically prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and mediate multidrug tolerance. *Nat. Microbiol.* 3, 1032–1042. doi: 10.1038/s41564-018-0218-3

Holt, K. E., Thai, P. V. K., Thuong, N. T. T., Ha, D. T. M., Lan, N. N., Lan, N. H., et al. (2018). Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* 50, 849–856. doi: 10.1038/s41588-018-0117-9

Houben, D., van Ingen, J., Perez, J., Baldeón, L., Abdallah, A. M., Caleechurn, L., et al. (2012). ESX-1-mediated translocation to the cytosol controls virulence of mycobacteria. *Cell Microbiol.* 14, 1287–1298. doi: 10.1111/j.1462-5822.2012.01799.x

Hsu, T., Chen, B., Chen, M., Dai, A. Z., Morin, P. M., Marks, C. B., et al. (2003). The primary mechanism of attenuation of bacillus Calmette-Guerin is a loss of secreted lytic function required for invasion of lung interstitial tissue. *Proc. Natl. Acad. Sci. U.S.A.* 100, 12420–12425. doi: 10.1073/pnas.1635213100

Huang, H., Yang, T., Li, C., Jia, X., Wang, G., Zhong, J., et al. (2019). Cross-sectional whole-genome sequencing and epidemiological study of multidrug-resistant mycobacterium tuberculosis in China. *Clin. Infect. Dis.* 69, 405–413. doi: 10.1093/cid/ciy883

Jiang, Q., Ji, L., Li, J., Zeng, Y., Meng, L., Luo, G., et al. (2020). Citywide transmission of multidrug-resistant tuberculosis under China's rapid urbanization: A retrospective population-based genomic spatial epidemiological study. *Clin. Infect. Dis.* 71, 142–151. doi: 10.1093/cid/ciz790

Jung, Y., and Han, D. (2022). BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* 38, 2404–2413. doi: 10.1093/bioinformatics/btac137

Knudsen, N. P., Dolganov, G. M., Schoolnik, G. K., Lindenstrøm, T., Andersen, P., Agger, E. M., et al. (2014). Tuberculosis vaccine with high predicted population coverage and compatibility with modern diagnostics. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1096–1101. doi: 10.1073/pnas.1314973111

Kohl, T. A., Rothgänger, J., Walker, T., Diel, R., and Niemann, S. (2018). Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. *EBioMedicine* 34, 131–138. doi: 10.1016/j.ebiom.2018.07.030

Koster, K. J., Foster, J. T., Drees, K. P., Qian, L., Desmond, E., Wan, X., et al. (2018). Whole genome SNP analysis suggests unique virulence factor differences of the Beijing and Manila families of Mycobacterium tuberculosis found in Hawaii. *BMC Infect. Dis.* 18, 608. doi: 10.1186/s12879-018-3502-1

Kruh, N. A., Izzo, A., Prenni, J., and Dobos, K. M. (2010). Portrait of a pathogen: the Mycobacterium tuberculosis proteome *in vivo*. *PloS One* 5, e13938. doi: 10.1371/journal.pone.0013938

Kursa, M. B. (2010). Feature selection with boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

Kursa, M. B., and Rudnicki, W. R. (2010). Boruta - A system for feature selection. *Fundament Inform.* 101, 271–285. doi: 10.3233/FI-2010-288

Letunic, I. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Lewis, K. N., Guinn, K. M., Hickey, M. J., Smith, S., Behr, M. A., and Sherman, D. R. (2003). Deletion of RD1 from Mycobacterium tuberculosis mimics bacille Calmette-Guérin attenuation. *J. Infect. Dis.* 187, 117–123. doi: 10.1086/jid.2003.187.issue-1

Lieberman, T. D., Misra, R., Xiong, L. L., Moodley, P., Cohen, T., and Kishony, R. (2016). Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. *Nat. Med.* 22, 1470–1474. doi: 10.1038/nm.4205

Lienard, J., Lovins, V., Movert, E., Valfridsson, C., and Carlsson, F. (2020). The Mycobacterium marinum ESX-1 system mediates phagosomal permeabilization and type I interferon production via separable mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* 117, 1160–1166. doi: 10.1073/pnas.1911646117

Liu, Q., Wei, L., Pang, Y., Wu, B., Luo, T., Zhou, Y., et al. (2018). China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. *Nat. Ecol. Evol.* 2, 1982–1992. doi: 10.1038/s41559-018-0680-6

Liu, B., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* 50, D912–D917. doi: 10.1093/nar/gkab1107

Luo, T., Luo, D., Lu, B., Wu, J., Wei, L., Yang, C., et al. (2015). Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8136–8141. doi: 10.1073/pnas.1424063112

Ly, A., and Liu, J. (2020). Mycobacterial virulence factors: surface-exposed lipids and secreted proteins. *Int. J. Mol. Sci.* 21 (11). doi: 10.3390/ijms21113985

MacGurn, J. A., and Cox, J. (2007). A genetic screen for Mycobacterium tuberculosis mutants defective for phagosome maturation arrest identifies components of the ESX-1 secretion system. *Infect. Immun.* 75, 2668–2678. doi: 10.1128/IAI.01872-06

Mahairas, G. G., Hickey, M. J., Singh, D. C., and Stover, C. K. (1996). Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. *bovis*. *J. Bacteriol.* 178, 1274–1282. doi: 10.1128/jb.178.5.1274-1282.1996

Majlessi, L., Casadevall, A., and Brosch, R. (2015). Release of mycobacterial antigens. *Immunol. Rev.* 264, 25–45. doi: 10.1111/imr.2015.264.issue-1

Mikheecheva, N. E., Melerzanov, A. V., and Danilenko, V. N. (2017). A nonsynonymous SNP catalog of mycobacterium tuberculosis virulence genes and its use for detecting new potentially virulent sublineages. *Genome Biol. Evol.* 9, 887–899. doi: 10.1093/gbe/evx053

Mishra, S. K., Jain, N., Sikri, K., Tyagi, J. S., Sharma, T. K., Mergny, J. L., et al. (2019). Characterization of G-Quadruplex Motifs in espB, espK, and cyp51 Genes of Mycobacterium tuberculosis as Potential Drug Targets. *Mol. Ther. Nucleic Acids* 16, 698–706. doi: 10.1016/j.omtn.2019.04.022

Mohanty, S., Ganguli, G., Padhi, A., Jena, P., Selchow, P., Sengupta, S., et al. (2016). Mycobacterium tuberculosis EsxO (Rv2346c) promotes bacillary survival by inducing oxidative stress mediated genomic instability in macrophages. *Tubercul. (Edinb)* 96, 44–57. doi: 10.1016/j.tube.2015.11.006

Pajuelo, D., Zhang, L., Danilchanka, O., Tischler, A. D., and Niederweis, M. (2021). Toxin secretion and trafficking by Mycobacterium tuberculosis. *Nat. Commun.* 12, 6592. doi: 10.1038/s41467-021-26925-1

Pallen, M. J. (2002). The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system? *Trends Microbiol.* 10, 209–212. doi: 0.1016/s0966-842x(02)02345-4

Pandey, H., Yabaji, S. M., Kumari, M., Tripathi, S., Srivastava, K., Tripathi, D. K., et al. (2018). Biophysical and immunological characterization of the ESX-4 system ESAT-6 family proteins Rv3444c and Rv3445c from Mycobacterium tuberculosis H37Rv. *Tubercul. (Edinb)* 109, 85–96. doi: 10.1016/j.tube.2018.02.002

Pang, X., Cao, G., Wang, X., Tvinnereim, A. R., Chen, X. L., and Howard, S. T. (2013). MprAB regulates the espA operon in Mycobacterium tuberculosis and modulates ESX-1 function and host cytokine response. *J. Bacteriol.* 195, 66–75. doi: 10.1128/JB.01067-12

Phelan, J. E., MaChado, D., Ramos, J., Oppong, Y. E. A., Campino, S., O'Grady, J., et al. (2019). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* 11, 41. doi: 10.1186/s13073-019-0650-x

Pym, A. S., Majlessi, L., Brosch, R., Demangel, C., Williams, A., Griffiths, K. E., et al. (2003). Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. *Nat. Med.* 9, 533–539. doi: 10.1038/nm859

Rivera-Calzada, A., Llorca, O., and Geibel, S. (2021). Type VII secretion systems: structure, functions and transport models. *Nat. Rev. Microbiol.* 19, 567–584. doi: 10.1038/s41579-021-00560-5

Rodríguez, N. A., Chaves, F., Iñigo, J., Herranz, M., Ritacco, V., EpiMOLTB Madrid; INDAL-TB group, et al. (2010). Differences in the robustness of clusters involving the Mycobacterium tuberculosis strains most frequently isolated from immigrant cases in Madrid. *Clin. Microbiol. Infect.* 16, 1544–1554. doi: 10.1111/j.1469-0691.2010.03161.x

Saha, S., Magbanua, Z. V., and Peterson, D. G. (2008). Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.* 36, 2284–2294. doi: 10.1093/nar/gkn064

Sani, M., Geurtsen, J., Pierson, J., de Punder, K., van Zon, M., Wever, B., et al. (2010). Direct visualization by cryo-EM of the mycobacterial capsular layer: a labile structure containing ESX-1-secreted proteins. *PloS Pathog.* 6, e1000794. doi: 10.1371/journal.ppat.1000794

Saulnier, D. M., Mistretta, T. A., Diaz, M. A., Mandal, D., Raza, S., Weidler, E. M., et al. (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology* 141, 1782–1791. doi: 10.1053/j.gastro.2011.06.072

Simeone, R., Lippmann, J., Bitter, W., Majlessi, L., Brosch, R., and Enninga, J. (2012). Phagosomal rupture by Mycobacterium tuberculosis results in toxicity and host cell death. *PloS Pathog.* 8, e1002507. doi: 10.1371/journal.ppat.1002507

Skjøt, R. L., Rosenkrands, I., Ravn, P., Brock, I., Jacobsen, S., and Andersen, P. (2000). Comparative evaluation of low-molecular-mass proteins from Mycobacterium tuberculosis identifies members of the ESAT-6 family as immunodominant T-cell antigens. *Infect. Immun.* 68, 214–220. doi: 10.1128/IAI.68.1.214-220.2000

Smith, J., Pan, M., Bohsali, A., Xu, J., Liu, J., McDonald, K. L., et al. (2008). Evidence for pore formation in host cell membranes by ESX-1-secreted ESAT-6 and its role in Mycobacterium marinum escape from the vacuole. *Infect. Immun.* 76, 5478–5487. doi: 10.1128/IAI.00614-08

Soler-Arnedo, P., Zhang, M., Cole, S. T., and Piton, J. (2020). Polarly localized eccE1 is required for ESX-1 function and stabilization of ESX-1 membrane proteins in mycobacterium tuberculosis. *J. Bacteriol.* 202, e00662–e00619. doi: 10.1128/JB.00662-19

Stamm, L. M., Gao, L. Y., Jeng, R. L., McDonald, K. L., Roth, R., Takeshita, S., et al. (2003). Mycobacterium marinum escapes from phagosomes and is propelled by actin-based motility. *J. Exp. Med.* 198, 1361–1368. doi: 10.1084/jem.20031072

Tufariello, J. M., Kerantzas, C. A., Wong, K. W., Vilchèze, C., Jones, C. M., Cole, L. E., et al. (2016). Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. *Proc. Natl. Acad. Sci. U.S.A.* 113, E348–E357. doi: 10.1073/pnas.1523321113

Uplekar, S., Friocourt, V., Rougemont, J., and Cole, S. T. (2011). Comparative genomics of Esx genes from clinical isolates of Mycobacterium tuberculosis provides evidence for gene conversion and epitope variation. *Infect. Immun.* 79, 4042–4049. doi: 10.1128/IAI.05344-11

van der Wel, N., Houben, D., Fluitsma, D., van Zon, M., Pierson, J., Brenner, M., et al. (2007). M. tuberculosis and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells. *Cell* 29. 129, 1287–1298. doi: 10.1016/j.cell.2007.05.059

van der Woude, A. D., Bhatt, A., Sparrius, M., Raadsen, S. A., Boon, L., Geurtsen, J., et al. (2012). Unexpected link between lipooligosaccharide biosynthesis and surface protein release in Mycobacterium marinum. *J. Biol. Chem.* 287, 20417–20429. doi: 10.1074/jbc.M111.336461

van Winden, V. J. C., Houben, E., and Braunstein, M. (2019). Protein Export into and across the Atypical Diderm Cell Envelope of Mycobacteria. *Microbiol. Spectr.* 7. doi: 10.1128/microbiolspec.GPP3-0043-2018

Walker, T. M., Broda, A., Ortega, L. S., Morgan, M., Parker, L., Churchill, S., et al. (2014). Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. *Lancet Respir. Med.* 2, 285–292. doi: 10.1016/S2213-2600(14)70027-X

Walker, T. M., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., Eyre, D. W., et al. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* 13, 137–146. doi: 10.1016/S1473-3099(12)70277-3

Watson, R. O., and Cox, J. S. (2012). Extracellular M. tuberculosis DNA targets bacteria for autophagy by activating the host DNA-sensing pathway. *Cell* 150, 803–815. doi: 10.1016/j.cell.2012.06.040

Wirth, S. E., Aldridge, B. B., Fortune, S. M., Fernandez-Suarez, M., Gray, T. A., and Derbyshire, K. M. (2012). Polar assembly and scaffolding proteins of the virulence-associated ESX-1 secretory apparatus in mycobacteria. *Mol. Microbiol.* 83, 654–664. doi: 10.1111/j.1365-2958.2011.07958.x

Yang, C., Shen, X., Wu, J., Gan, M., Xu, P., Wu, Z., et al. (2017). Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect. Dis.* 17, 275–284. doi: 10.1016/S1473-3099(16)30418-2

Yao, J., Chen, S., Shao, Y., Deng, K., Jiang, M., Liu, J., et al. (2018). Rv2346c enhances mycobacterial survival within macrophages by inhibiting TNF-α and IL-6 production via the p38/miRNA/NF-κB pathway. *Emerg. Microbes Infect.* 7, 158. doi: 10.1038/s41426-018-0162-6

Zelner, J. L., Becerra, M. C., Galea, J., Lecca, L., Calderon, R., Yataco, R., et al. (2016). Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. *J. Infect. Dis.* 213, 287–294. doi: 10.1093/infdis/jiv387

Zhang, M., Sala, C., Rybniker, J., Dhar, N., and Cole, S. T. (2014). EspI regulates the ESX-1 secretion system in response to ATP levels in Mycobacterium tuberculosis. *Mol. Microbiol.* 93, 1057–1065. doi: 10.1111/mmi.2014.93.issue-5

Zheng, C., Zhao, C., Zozio, T., Li, S., Luo, D., Sun, Q., et al. (2017). New Mycobacterium tuberculosis Beijing clonal complexes in China revealed by phylogenetic and Bayesian population structure analyses of 24-loci MIRU-VNTRs. *Sci. Rep.* 7, 6065. doi: 10.1038/s41598-017-06346-1