scientific reports

Check for updates

OPEN

Global epidemiology of *Mycobacterium tuberculosis* lineage 4 insights from Ecuadorian genomic data

Gabriel Morey-León[1,3,4,10]✉, Paulina M. Mejía-Ponce[2,10], Juan Carlos Fernández-Cadena[4,9], Evelyn García-Moreira[5], Derly Andrade-Molina[1,4], Cuauhtémoc Licona-Cassani[2], Pablo Fresia[6] & Luisa Berná[7,8]✉

Tuberculosis is a global public health concern, and understanding *Mycobacterium tuberculosis* transmission routes and genetic diversity of *M. tuberculosis* is crucial for outbreak control. This study aimed to explore the genomic epidemiology and genetic diversity of *M. tuberculosis* in Ecuador by analyzing 88 local isolates and 415 public genomes from 19 countries within the Euro-American lineage (L4). Our results revealed significant genomic diversity among the isolates, particularly in the genes related to protein processing, carbohydrate metabolism, lipid metabolism, and xenobiotic biodegradation and metabolism. The population structure analysis showed that sub-lineages 4.3.2/3 (35.4%), 4.1.2.1 (22.7%), 4.4.1 (12.7%), and 4.1.1. (10.7%) were the most prevalent. Phylogenetic and transmission network analyses suggest that these isolates circulating within Ecuador share genetic ties with isolates from other continents, implying historical and ongoing intercontinental transmission events. Our findings underscore the importance of integrating genomic data into public health strategies for tuberculosis control and suggest that enhanced genomic surveillance is essential for understanding and mitigating the global spread of *M. tuberculosis*. This study provides a comprehensive genomic framework for future epidemiological investigations and control measures targeting *M. tuberculosis* L4 in Ecuador.

Keywords Ecuador, Genomic clusters, Genomic epidemiology, Surveillance, TMRCA, Tuberculosis

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (*Mtb*) is a significant global health concern, with an estimated 10.6 million infections and 1.6 million casualties by 2021[1]. The COVID-19 pandemic has disrupted access to medical healthcare, including TB diagnosis and treatment programs, aggravating the TB burden, and compromising the progress in TB control achieved in recent decades[2,3]. Furthermore, the incidence of TB has been increasing in several countries, including Ecuador, where the rate reached 48 cases per 100,000 inhabitants in 2021[4] due, among other factors, to COVID-19 containment measures[5,6]. The increase in global migration has contributed to the spread of TB, mainly in high-income countries where migrants seek better economic, educational, or living opportunities. In the Ecuadorian context, migration is primarily directed toward the USA, Spain, Italy, Canada, and Chile[7–9].

Identifying the transmission routes of TB cases is essential for reducing potential transmission networks. Different public health organizations have implemented screening programs among household contacts to control the spread of the disease, including the TB-Directly Observed Treatment Short Course (TB-DOTS)[10]. Genomic approaches, such as whole-genome sequencing (WGS) of *Mtb* strains, have been a landmark in the

[1]Facultad de Ciencias de la Salud, Universidad Espíritu Santo, Samborondón, Ecuador. [2]Centro de Biotecnología FEMSA, Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Monterrey, Nuevo León, México. [3]Universidad de la República, Montevideo, Uruguay. [4]Laboratorio de Ciencias Ómicas, Universidad Espíritu Santo, Samborondón, Ecuador. [5]Instituto Superior Tecnológico Argos, Guayaquil, Ecuador. [6]Unidad Mixta Pasteur + INIA (UMPI), Institut Pasteur de Montevideo, Montevideo, Uruguay. [7]Laboratorio de Interacciones Hospedero-Patógeno, Unidad de Biología Molecular, Institut Pasteur de Montevideo, Montevideo, Uruguay. [8]Unidad de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. [9]African Genome Center, University Mohammed VI Polytechnic (UM6P), Ben Guerir, Morocco. [10]Gabriel Morey-León and Paulina M. Mejía-Ponce contributed equally to this work. ✉email: garielmorey@uees.edu.ec; lberna@pasteur.edu.uy

traceability of transmitted TB cases and have provided invaluable information on drug resistance and sub-lineage patterns[11–16]. WGS and Bayesian phylogenetic approaches have reconstructed the historical patterns of TB spread in Central and South America, dating back to the introduction of *Mtb* strains in these regions. The apparent ancestral emergence and marked diversification of the L2 (ancestral and modern Beijing sub-lineages) and L4 lineages have also been emphasized, reflecting their close correlation with restricted geographic distribution, leading to the independent emergence of multiple sub-lineages and local adaptation to distinct human populations[17–20]. Studies conducted in Brazil, Paraguay, Mexico, and the United States have applied genomic approaches to identify frequent transmission clusters between prison inmates, drug users, migrants, and mixed groups[18–21], highlighting the need to prioritize contact tracing to groups with a higher likelihood of retrospective clustering to improve TB control.

The *Mycobacterium tuberculosis* complex (MTBC) comprises distinct phylogenetic lineages that have evolved over centuries[21–23]. The Euro-American lineage (L4) exhibits extensive sub-lineage diversity both within and across countries. In Central and South America, the Caribbean, Europe, and Middle Africa, the most prevalent L4 sub-lineages include Latin American (LAM), Haarlem, X-type, and T families[24–29]. Previous reports have revealed that the LAM and Haarlem families are the most prevalent in Ecuador, with few cases in the Beijing family[30–33]. Transmission clustering in these cases has often been linked to isolates from neighboring countries[34,35]. A pivotal WGS study on a limited dataset of Ecuadorian *Mtb* isolates identified the 4.3.2/3 (LAM) and 4.1.2 (Haarlem) sub-lineages, highlighting the significant genetic diversity present[36]. This study underscores the importance of understanding the local transmission dynamics of TB within Ecuador, particularly considering the potential influence of events occurring outside national boundaries, associated with migration patterns, trade, and regional cooperation. Although studies have investigated transnational TB transmission between Ecuador and countries such as Colombia[37], evidence remains sparse. We analyzed the transmission networks of 88 *Mtb* isolates of L4 lineage collected from various cities across Ecuador in comparison with 415 publicly available genomes in 19 Latin American countries. To accomplish this, we used a combination of phylogenetics and composition analysis. Additionally, we explored the role of genetic variation in virulence genes in the recent transmission of TB in Ecuador. Our findings provide crucial insights into the genetic diversity of *Mtb* in Ecuador, highlighting how migration influences the TB burden and transmission dynamics locally. This study contributes to the development of more effective targeted TB control strategies tailored to the unique characteristics of *Mtb* populations.

Results

Genomic and functional analysis of Ecuadorian Mtb isolates

This study focused on genomic analysis of raw reads of 88 M. *tuberculosis* isolates collected in Ecuador[38]. To extend our investigation of genomic variability and epidemiology, we combined these data with a selection of 415 M. *tuberculosis* isolates from different countries (see Methods). Initially, we used the raw sequences (average coverage of 61X) of the 88 Ecuadorian isolates to perform quality control and genome assembly for each isolate using Unicycler and Pilon while discarding contigs smaller than 300 bp. The quality of the genomes was assessed and an average N50 of 65,747. Annotation was performed using Prokka software. These analyses yielded approximately 4298 genes per isolate, including coding sequences and RNA gene transfer. Descriptive statistics for this analysis are presented in Table 1 and Supplementary Table 3.

Genomic analysis of the Ecuadorian *M. tuberculosis* isolates revealed that they belonged to the Euro-American Lineage (lineage 4). Among the sublineages, 4.3.3 (27.3%, 24/88) and 4.1.1 (23.9%, 21/88) were the most prevalent, followed by 4.4.1.1 and 4.1.2.1 (11.4%, 10/88 each). Clades 4.1.2, and 4.3.4.1/2 were present, but in smaller proportions.

Genomic features	Mean	Max	Min
Medium-coverage sequencing depth	61	132	10
Assembly size (base pairs)	4,320,543	4,337,587	4,213,530
% GC	65.53	65.58	65.39
N50	65,747	115,041	11,184
% sequenced genome	98.39	99.17	96.33
Coding sequences	4254	4346	4226
tRNA	44	44	44
Gene in subsystems	1961	2004	1891
Number of proteins	4011	4202	3921
Virulence factor	476	485	471
Pangenome	4397	5219	3812
Single nucleotide polymorphism	731	858	389
Intergenic	106	130	50
Nonsynonymous	372	441	205
Synonymous	233	282	122

Table 1. Genomics features of 88 *Mtb* Ecuadorian isolates.

Functional annotation and metabolic insights

We performed functional annotation of each of the 88 Ecuadorian genomes. A substantial proportion (77%) of the annotated genes was associated with specific functional roles (e.g., metabolic processes, cellular processes, energy, protein processing, and stress response, Defense, and Virulence). We enriched our dataset with functional annotations, assigning Enzyme Commission (EC) numbers to 1,061 proteins, Gene Ontology (GO) classifications to 918 proteins, linking 814 proteins to specific KEGG pathways, and identifying approximately 816 virulence factors distributed across three databases. This enrichment analysis provided deep insights into the metabolic functions and potential pathogenic mechanisms of the isolates (Supplementary Table 3).

Our subsystem analysis, categorized according to the KEGG database, showed that approximately 1,961 genes per isolate are involved in various biological processes and structural complexes (Supplementary Table 3). The analysis revealed that the majority of the annotated protein-coding genes were related to metabolism (40.3%), with 37.9% of these genes being associated with cofactors, vitamins, and prosthetic groups (300 genes). This was followed by genes related to the stress response, Defense, Virulence (10.8%), and energy (10.5%). In our detailed metabolic pathway analysis, we observed a predominance of genes involved in amino acid (369 genes), carbohydrate (312 genes), lipid (262 genes), and xenobiotic biodegradation and metabolism (240 genes). These pathways are crucial for the pathogenicity and survival of *Mtb*[39] (Supplementary Table 3–5). Notably, variability was observed among different sub-lineages. Sub-lineage 4.1.2.1 shows a higher proportion of proteins related to metabolism and a lower proportion of proteins related to protein processing. Conversely, sub-lineage 4.1.2 presents a higher proportion of proteins related to protein processing and fewer proteins related to metabolism (Supplementary Fig. 1A). Lineage 4.1.2.1 exhibits a higher number of gene variations related to metabolism, suggesting significant metabolic diversity, indicating a complex and adaptable metabolic network, whereas lineage 4.3.2/3 showed greater variations in genes associated with protein processing, suggesting a diverse set of genes involved in protein synthesis, folding, and degradation, reflecting the complexity of protein processing mechanisms (Table 2).

Pangenome analysis of the Ecuadorian Mtb isolates

We also performed pan-genome analysis to understand the genetic diversity and conservation of 88 Mtb isolates from Ecuador. Pangenome reconstruction of the 88 *Mtb* isolates revealed 4,397 gene families, including 3,104 classified as core (present in at least 99% of isolates), 666 as accessory, and 270 as cloud gene families (Supplementary Fig. 1B and Supplementary Table 3). Interestingly, 70.5% of the isolates belonged to core gene families, suggesting minimal within-genome variability and emphasizing the high level of genetic conservation among the isolates. According to pangenome calculations, a *b* value of 0.086 in the power-law regression model indicated a close pangenome for *Mtb*. Three isolates (S1454, S1477, and S1453) from sub-lineages 4.1.1 and 4.3.2/3 showed the highest number of cloud genes (261, 200, and 177, respectively). Most gene families within the core and accessory partitions belong to metabolic subsystems. Conversely, the majority of the unique gene families were classified within the Environmental Information Processing subsystem (Supplementary Fig. 1C,D).

Genotypic drug-resistance analysis

Furthermore, when analyzing the phylogenetic relationships and genetic diversity associated with resistance in Ecuadorian isolates (Fig. 1), our findings revealed that sub-lineages 4.3.2/3 and 4.1.1, exhibited a higher prevalence of resistant variants (46.6% and 26.1%, respectively). We identified 42 single-nucleotide variants (SNVs) in resistance-related genes, with *rpoB* showing the most mutations. Additionally, SNVs were more frequently identified in genes related to intermediary metabolism and respiration (21.71%), followed by cell wall processes (21.34%) and conserved hypotheticals (20.44%). Exhaustive details of the SNV are provided in Supplementary Table 5.

Subsystems	4.1.2, *n* = 4	4.1.2.1, *n* = 9	4.3.2/3, *n* = 39	4.8, *n* = 5	4.4.1.1, *n* = 10	4.1.1, *n* = 21
Cell envelope	80 (3)	80 (4)	80 (3)	78 (1)	79 (2)	81 (3)
Cellular processes	164 (1)	164 (3)	164 (4)	164 (3)	164 (4)	164 (5)
DNA processing	85 (2)	85 (8)	85 (11)	85 (3)	85 (2)	85 (2)
Energy	206 (7)	206 (8)	207 (7)	207 (7)	209 (1)	206 (6)
Membrane transport	132 (8)	132 (14)	131 (10)	132 (8)	131 (8)	131 (10)
Metabolism	792 (7)	787 (75)	790 (32)	787 (25)	788 (23)	792 (26)
Miscellaneous	48 (3)	47 (2)	48 (3)	47 (2)	48 (2)	48 (3)
Protein processing	188 (30)	184 (34)	185 (52)	179 (31)	177 (31)	189 (32)
Regulation and cell signaling	15 (0)	15 (0)	15 (0)	15 (0)	15 (0)	15 (0)
RNA processing	36 (0)	36 (0)	36 (0)	36 (0)	36 (0)	36 (0)
Stress response, defense, virulence	212 (7)	212 (12)	211 (22)	212 (6)	212 (7)	212 (9)
CRISPR	9(1)	9 (2)	9 (2)	9 (2)	9 (2)	9 (2)

Table 2. Frequency average distribution of Subsystems of 88 *Mtb* Ecuadorian isolates classified according to their sub-lineage. The numbers in brackets represent the number of proteins affected by variability according to the KEGG Database in the Ecuadorian isolates. 4.1.2, 4.1.2.1, 4.3.2/3, 4.8, 4.4.1.1, and 4.1.1 correspond to L4 sub-lineages Euro-American, Haarlem, LAM, mainly T, S-type, and X-type respectively.

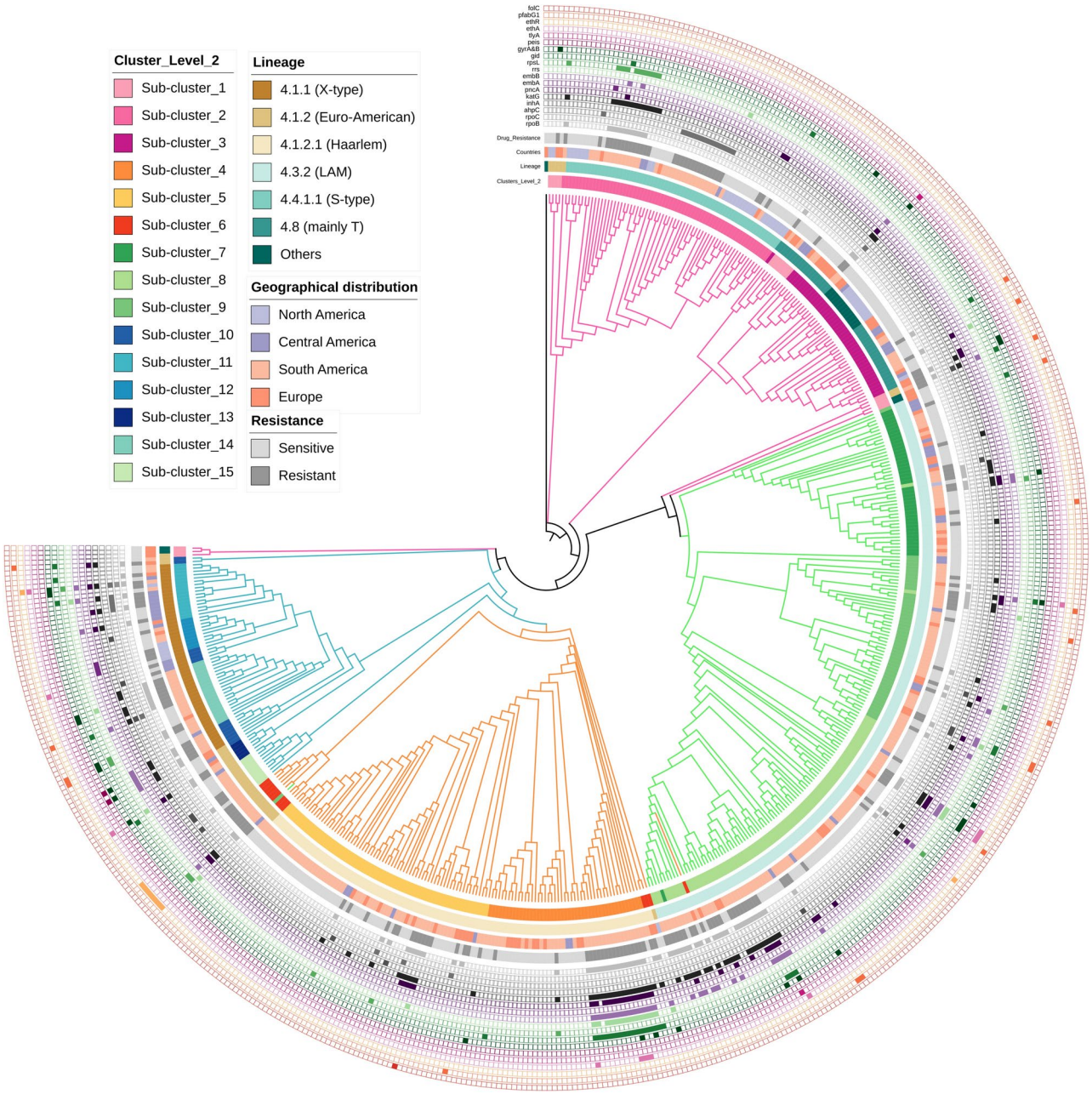

Fig. 1. Phylogenetic reconstruction of the 503 *Mtb* isolates. Circular representation of the phylogenetic tree reconstructed with 1437 SNPs using the ML method, the GTR + GAMMA substitution model, bootstrap support of 1000 replicates, and rooted with *M. microti*. Circular strips indicated the following metadata (inside out): (i) fifteen sub-clusters classified from the second clustering level, (ii) sub-lineage classification, (iii) the geographic zone of the samples, (iv) clinical drug-resistance classification, and (v) canonical variants associated with drug resistance. Color in branches on trees represents the first level of clustering (Lineages) of *Mtb* isolates generated by rhierBAPS.

To gain a deeper insight into the genomic epidemiology of TB in Ecuador, we incorporated an additional 415 *Mtb* samples from 19 countries categorized as both continental and transoceanic migratory nations (See "Methods" section). All these 503 *Mtb* genomes belonged to the L4 lineage, including 4.3.2/3 (35.4%), 4.1.2 (22.7%), 4.4.1 (12.7%), and 4.1.1 (10.7%) sub-lineages (Supplementary Table 2). Moreover, 63.8% (321/503) of *Mtb* samples were genotypically susceptible to all anti-TB drugs. The remaining 36.2% (180/503) were resistant to at least one antibiotic. The clinical classification of these drug-resistant samples showed that 16.7% were identified as MDR-TB, 8.5% as HR-TB, 4.2% as pre-XDR-TB, and 1.6% as RR-TB. Table 3 summarizes the drug-resistant clinical classification for the 503 *Mtb* samples per region and Table 4 provides an overview of their drug-resistant canonical mutations. An extended list of canonical mutations is provided in Supplementary Table 6.

Geographic region	HR-TB, n (%)	RR-TB, n (%)	MDR-TB, n (%)	Pre-XDR-TB, n (%)	Other, n (%)	Sensitive, n (%)
South America	33 (6.6)	4 (0.8)	71 (14.1)	19 (3.8)	12 (2.4)	207 (41.2)
4.1.1	2 (0.4)		7 (1.4)	7 (1.4)	1 (0.2)	14 (2.8)
4.1.2	3 (0.6)	2 (0.4)	4 (0.8)	1 (0.2)	6 (1.2)	10 (2.0)
4.1.2.1	5 (1.0)	1 (0.2)	20 (4.0)	2 (0.4)	2 (0.4)	65 (12.9)
4.3.2/3	8 (1.6)	1 (0.2)	29 (5.8)	8 (1.6)	2 (0.4)	95 (18.9)
4.4.1.1	15 (3.0)		10 (2.0)		1 (0.2)	8 (1.6)
4.8			1 (0.2)	1 (0.2)		15 (3.0)
Central America	7 (1.4)		7 (1.4)		4 (0.8)	38 (7.6)
4.1.2					1 (0.2)	2 (0.4)
H37Rv-like						1 (0.2)
4.1.2.1			2 (0.4)			4 (0.8)
4.3.2/3	2 (0.4)		1 (0.2)		1 (0.2)	16 (3.2)
4.8			1 (0.2)		1 (0.2)	7 (1.4)
4.4.1.1	2 (0.4)					1 (0.2)
4.1.1	3 (0.6)		3 (0.6)		1 (0.2)	7 (1.4)
North America					7 (1.4)	40 (8.0)
4.1.2						3 (0.6)
H37Rv-like						10 (2.0)
4.3.2/3						1 (0.2)
4.8					1 (0.2)	
4.4.1.1					2 (0.4)	24 (4.8)
4.1.1					4 (0.8)	2 (0.4)
Europe	3 (0.6)	4 (0.8)	6 (1.2)	2 (0.4)	1 (0.2)	38 (7.6)
Cameroon	1 (0.2)					1 (0.2)
4.1.2			1 (0.2)		1 (0.2)	2 (0.4)
H37Rv-like						2 (0.4)
4.1.2.1	1 (0.2)	1 (0.2)	1 (0.2)	1 (0.2)		9 (1.8)
4.3.2/3	1 (0.2)	1 (0.2)	3 (0.6)			9 (1.8)
M. microti						1 (0.2)
4.8		1 (0.2)	1 (0.2)	1 (0.2)		9 (1.8)
4.4.1.1						1 (0.2)
TUR						2 (0.4)
4.1.1		1 (0.2)				2 (0.4)

Table 3. Frequency distribution of the 503 *Mtb* samples, showing their sub-lineage and their drug resistance profile per geographical region. HR-TB isoniazid resistance, RR-TB Rifampicin resistance, MDR-TB multidrug resistance, Pre-XDR-TB pre-extremly drug resistance. TB tuberculosis. 4.1.2, 4.1.2.1, 4.3.2/3, 4.8, 4.4.1.1, and 4.1.1 corresponding to L4 sub-lineages Euro-American, Haarlem, LAM, mainly-T, S-type, and X-type respectively.

Since multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) can increase domestic transmission[40], we determined the frequency of resistance-associated canonical mutations in 503 *Mtb* isolates. The most frequent mutations were *katG* Ser315Thr (n = 92), *rpoB* Ser450Leu (n = 68), *rpsL* Lys43Arg (n = 21), *embB* Met306Ile (n = 17), *pncA* Gln10Pro (n = 14), and *gyrA* Ala90Val (n = 11), which confer resistance to INH, RIF, STR, EMB, PZA, and FQs, respectively. Other common canonical mutations included fabG*1* − 15C > T (n = 16), *ahpC* -74G > A (n = 15), and *gid* 329_330delTG (n = 16), which are associated with INH and STR resistance, respectively.

TB population structure and transmission clusters

To better understand the genetic diversity and potential transmission dynamics of TB in Ecuador, we performed a population structure analysis of 503 *Mtb* samples (Fig. 2). Clusterization based on the rhierBAPS approach revealed four primary ***clusters*** at the initial hierarchical level, which were further divided into 15 ***sub-clusters*** at the second hierarchical level. Phylogenetic reconstruction of 503 isolates using 1,437 SNPs revealed four main clusters with different branch colors. Additionally, the sub-clusters (level 2), lineages, geographical distribution, and resistance are represented by colored concentric lines. Cluster 1 (pink) encompassed 128 isolates divided into three sub-clusters (1–3), containing samples belonging to 4, 4.4.1.1, 4.2.2, 4.6.2, and 4.7/8 sub-lineages. Cluster 2 (in orange) comprises 125 isolates, including three sub-clusters (4–6), with the majority belonging to the 4.1.2.1 sub-lineage, and some isolates belonging to the 4.1.2 sub-lineage (sub-cluster 6). Cluster 3 (in green) was composed of 178 isolates, distributed across three sub-clusters (7–9), primarily associated with

Drug	Gene name	Canonical variants (n)
RIF	*rpoB*	Ser450Leu (68), Asp435Val (13), His445Asn (6)
INH	*ahpC*	−74G > A (15)
	fabG1	−15C > T (16)
	inhA	−154G > A (9)
	katG	Ser315Thr (92)
PZA	*pncA*	Gln10Pro (14)
EMB	*embB*	Met306Ile (17), Gly406Ala (15)
STR	*gid*	329_330delTG (16)
	rpsL	Lys43Arg (21)
FQ	*gyrA*	Ala90Val (11)
	gyrB	Asp461His (2)
AMG	*rrs*	1401A > G (18)
	eis	−12C > T (2)
	tlyA	Gly232Asp (1), Lys69Glu (1)
ETH	*ethA*	1222delT (3)
	ethR	Phe110Leu (8)
PAS	*folC*	Glu40Gly (1)

Table 4. Canonical variants associated with drug resistance distributed within the 503 *Mtb* samples. *#inhA* promoter mutations include mutations in *fabG1* open reading frame (ORF) because they create alternative promoters for *inhA* and mutations upstream of *fabG1* because they act as promoters of the entire operon, which includes *inhA*. (*) Stop codon. (†) Variants associated with INH and ETH resistance. Abbreviature: RIF (Rifampicin), INH (Isoniazid), PZA (Pyrazinamide), EMB (Ethambutol), STR (Streptomycin), FQ (Fluoroquinolones), AM (Amikacin), ETH (Ethionamide), PAS (Para-aminosalicylic acid). *katG*: Catalase-peroxidase, *rpoB*: RNA polymerase beta subunit, *rpsL*: Ribosomal protein S12, *embB*: Arabinosyl transferase B, *pncA*: Pyrazinamidase/nicotinamidase, *gyrA*: DNA gyrase subunit A, *gyrB*: DNA gyrase subunit B, *rrs*: 16S ribosomal RNA, *eis*: Enhanced intracellular survival protein (aminoglycoside acetyltransferase), *fabG1*/*inhA*: Enoyl-ACP reductase, *ahpC*: Alkyl hydroperoxide reductase C, tlyA: RRNA methyltransferase associated with ribosomal RNA modification, *ethA*: Monooxygenase involved in the activation of ethionamide, *ethR*: Transcriptional repressor regulating *ethA* expression, *folC*: Dihydrofolate synthase/folylpolyglutamate synthase and *gid*: Ribosomal small subunit methyltransferase G.

4.3.2, 4.3.3, 4.3.4.1, and 4.3.4.2, and, unexpectedly, a 4.1.2 isolate within sub-cluster 9. Finally, Cluster 4 (in blue) encompasses six sub-clusters (10–15), including 71 isolates predominantly from the 4.1.1, 4.1.1.1, and 4.1.1.3 sub-lineages. Notably, the isolates formed subclusters 13 and 15, with a small portion of subcluster 10 corresponding to the 4.1 sub-lineage.

In addition, we analyzed the isolates for transmission clusters by pairwise comparisons using a minimum distance of 12 SNPs with MTBseq. Thus, strains in the same *transmission genomic cluster (TGC)* had fewer than 12 SNPs. Using this strategy, 92.8% (n = 467) of the isolates were classified into 51 TGCs, ranging from two to 63 members, whereas 35 isolates were not classified. Based on the number of members per TGC (see "Methods" section), we classified seven members as *small*, 20 as *medium*, and 24 as *large* (Supplementary Table 7). Most of the 51 TGCs resolved using this methodology were within Cluster 3 (37.3%) and Cluster 1 (29.4%), as determined using rhierBAPS (Fig. 2A).

In particular, *Mtb* samples identified as susceptible were more prevalent in larger TGC than in resistant samples. Most pre-XDR isolates exhibited clonal relationships either among themselves or with isolates from other countries. Of the 35 isolates not grouped by MTBseq, 28.6% belonged to sub-lineage 4.3.2/3/4 and 25.7% to sub-lineage 4.8 (Fig. 2A and Table 5).

Upon closer examination of the five largest TGCs, specifically, TGC_1, TGC_2, TGC_9, TGC_18, and TGC_19, distinct geographical profiles were identified. Notably, TGC_1 (n = 63 isolates) predominantly comprised *Mtb* isolates from Paraguay (46.3%), Peru (19.0%), and Spain (15.9%). TGC_18 (n = 56) included isolates from Ecuador (41.0%) and Paraguay (26.8%), whereas TGC_19 (n = 31) consisted solely of isolates from Spain (38.7%). TGC_2 (n = 30) was predominantly isolated from Argentina (53.3%), whereas TGC_9 (n = 24) was isolated mainly from Canada (54.1%).

Among the 88 Ecuadorian *Mtb* isolates, 84 were grouped into 15 of 51 identified TGCs. These included five medium-sized and ten large TGCs. Notably, ten of these TGCs (TGC_2, TGC_3, TGC_5, TGC_14, TGC_15, TGC_18, TGC_20, TGC_21, TGC_37, and TGC_47) included isolates from Ecuador alongside isolates from various countries, unveiling previously unrecognized potential continental and intercontinental connections. In particular, TGC_37 and TGC_14 exhibited a close genetic relationship among their members, showing distances of ≤ 9 SNPs and ≤ 6 SNPs, respectively. In addition, TGC_37 comprised five isolates from Ecuador and one from an Ecuadorian migrant in Spain, highlighting the impact of migration on TB spread. Conversely, TGC_14

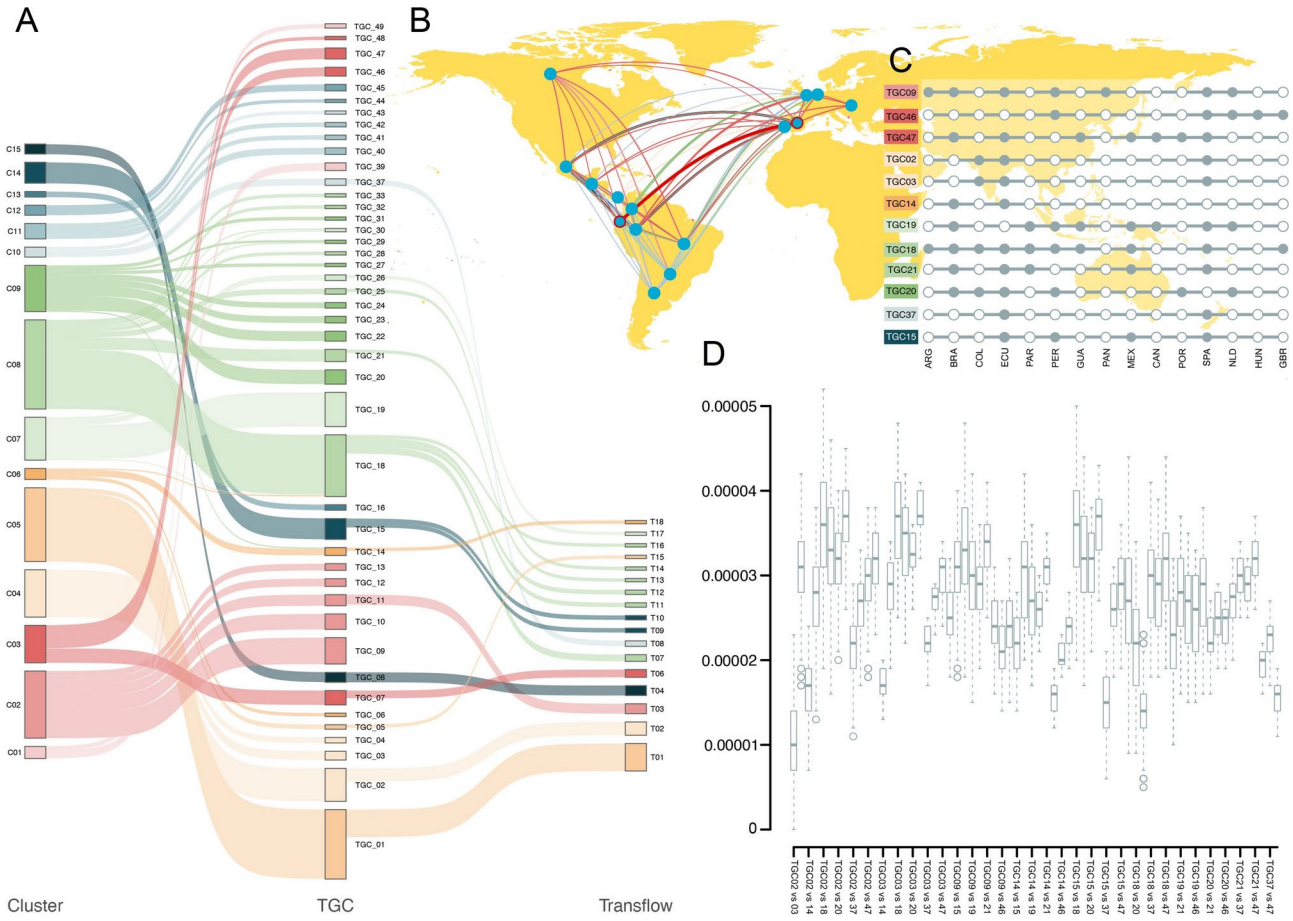

Fig. 2. Genomics clustering network of the 503 *Mtb* samples. (**A**) The plot showed the relationship between sub-clusters (C01—C15) defined from the population structure analysis with rhierBAPs, the TGC (< 12 SNPs) determined by MTBseq, and Transmission networks identified by Transflow analysis. (**B**) TGCs Networks considered the Ecuadorian isolates analyzed in this study, in red highlighting the connection confirmed by Transflow. The width of the links is due to the genetic distance (wider, less distance). The color of the links is according to the TGCs. More details at https://microreact.org/project/phylo-tb. (**C**) The geographical distribution of TGCs from the B panel revealed 12 networks. Nonetheless, two of these TGCs did not incorporate isolates from Ecuador. Interestingly, isolates from Ecuadorian migrants in Spain were identified within these TGCs. (**D**) Boxplot displaying genetic distances within and between the TGCs involving Ecuadorian isolates.

Clusters classification	*n* of isolates	Distribution of transmission genomic clusters (TGCs)	Largest TGCs *n* isolate, lineage
Cluster 1	128	15 TGCs + 14 ungrouped isolates	TGC_9: 24 isolates. S-type (4.4.1.1)
Cluster 2	125	7 TGCs + 5 ungrouped isolates + 1 isolate TGC18 (Cluster3)	TGC_1: 63 isolates. Haarlem (4.1.2.1) TGC_2: 30 isolates. Haarlem (4.1.2.1, *n* = 14; 4.1.2.1.1, *n* = 16)
Cluster 3	178	19 TGCs + 10 ungrouped isolates + 1 isolate TGC14 (Cluster2)	TGC_18: 56 isolates. LAM (4.3.3) TGC_19: 31 isolates. LAM (4.3.2)
Cluster 4	71	10 TGCs + 6 ungrouped isolates	

Table 5. Distribution of 503 isolates of *Mtb* according to Clustering (rhierBAPS) and TGCs approach. *TGC* transmission genomic cluster, *LAM*: Latin-American.

grouped two isolates from Ecuador with five from Brazil, indicating active transmission between these nations (Supplementary Table 2).

The Ecuadorian isolates within each TGC exhibited a high degree of clonality. For example, approximately 90% of the Ecuadorian isolates in TGC_11 and 70% in TGC_15 and TGC_18 had a genetic distance of zero SNPs (Supplementary Tables 7, 8 and Supplementary Figs. 2, 3). Similarly, in some TGCs, many isolates remained distinct and did not cluster with other isolates (singletons) in various proportions. (Supplementary Table 9).

Regarding the sub-lineage composition, the most representative sub-lineages in the Ecuadorian isolates among the TGCs were 4.3.3 (46.9%, 23/49; TGC_18), 4.1.1 (32.7%, 16/49; TGC_15), and 4.4.1.1 (20.4%, 10/49; TGC_11). Concerning the potential transmission of isolates, we identified four pre-XDR TB cases in TGC_15 and three pre-XDR isolates in TGC_15 and TGC_18, corresponding to clonal samples (genetic distance < 1 SNP). Among these three TGCs, 41% were from previously treated patients with TB, 34% were from untreated patients, and 25% were from patients currently undergoing treatment.

TB transmission networks analysis

TB transmission networks, that is, the suggested transmission routes of TB considering the genetic distances between samples, their geographical location, and year of isolation, were inferred for the 503 *Mtb* isolates using TransFlow. Using this strategy, 46 transmission networks were identified; however, only 18 of these networks had more than three members, as shown in Fig. 2A. As can be seen, most of these transmission networks (with more than three members) correspond to different TGCs, meaning they present more than 12 SNPs between them. However, six transmission networks (7, 11–14, and 16) corresponded to TGC_18, indicating the low genetic distance between them (Fig. 2A).

Considering the three clustering strategies employed (rhierBAPS, MTBseq, and TransFlow), we identified 17 convergent transmission groups to define potential TB transmission networks between Ecuadorian and non-Ecuadorian isolates. The geographic distribution of networks involving Ecuadorian isolates is shown in Fig. 2B, where TGCs that were confirmed and congruent with TransFlow are highlighted in red. Notably, some of these Ecuadorian samples (collected between 2019 and 2021) showed clonality with isolates from Colombia (n = 2) and Latin American migrants in Spain (n = 4), which were detected in 2014 and 2015, respectively (Fig. 2A, Supplementary Figs. 3, 4 and Supplementary Table 7). We identified several transmission networks showing connections among countries, including Paraguay, Argentina, and Brazil, with specific samples acting as potential index cases in the corresponding transmission networks (Supplementary Fig. 3). The distribution of local and migrant Ecuadorian isolates revealed diverse representations spread across 12 TGCs (Fig. 2C). Some of these isolates showed clonality with samples from Colombia, Latin America, and Ecuadorian migrants in Spain, or were identified as potential index cases within transmission networks (Fig. 2C and Supplementary Fig. 4). Notably, TGC18 and 19 exhibit diverse nationalities, with 10 and 8 different origins represented within these clusters, respectively. The short genetic distance between different TGCs indicated a close relationship, suggesting a possible joint origin. This relationship was supported by the genetic distances observed between the Ecuadorian isolates, implying a shared ancestry among the TGCs (Fig. 2D). These findings highlight the complex transmission dynamics of TB, with evidence of cross-border and international transmission networks involving Ecuadorian populations and those from other Latin American countries. The diversity of isolates and potential index cases underscore the need for targeted public health interventions to address the multifaceted nature of TB epidemiology.

Genetic diversity within the virulence-associated genes

To characterize mutations in the genes involved in host adaptability, we analyzed SNPs in genes commonly associated with MTBC virulence in Ecuadorian isolates[64]. A total of 303 SNPs were identified in these virulence genes, and all 88 isolates had at least one SNP each in the mce1F, mmpL4, phoR, ctpV, pepD, mce3F, fadD13, and nuoG genes. The pks12, fadD5, mce3C, pks12, nuoG, and katG genes were mutated with at least one SNP in more than 50% of the isolates. The top six genes with the most significant SNPs were plcA, plcB, pks7, pks12, phoR, and PPE46 (Supplementary Tables 10, 11).

Among the isolates, 46.6% (41/88) had more than 40 SNPs, with a maximum of 79 SNPs identified within virulence genes. Isolates corresponding to sub-lineages 4.1.1 (TGC_15 and TGC_37), 4.4.1.1 (TGC_11), and 4.3.2/3/4 (TGC_18, _20, _21, _26, _29, and _30) presented higher numbers of virulence-associated polymorphisms. Additionally, two members of TGC_14 (S0017 and S0039) and one ungrouped sample (S2193) showed a high number of mutations in the virulence genes (70, 71, and 74 SNPs, respectively). Interestingly, the five isolates associated with sub-lineage 4.7/8 showed the lowest number of SNPs in the virulence genes (15–25 SNPs).

Phylodynamics of TB in Ecuador

The divergence time to the most recent common ancestor (TMRCA) estimated for the Ecuadorian isolates fell within a density interval between 697 and 1,475 years, with an estimated time of 1,054 years. We also estimated TMRCA for different TGCs, identifying TGC_15 as the oldest (218 years before present, YBP) and TGC_03 as the youngest (130 YBP). Notably, isolates corresponding to TGC_02, TGC_21, and TGC_47, each of which included only one Ecuadorian isolate, had TMRCAs greater than 150 YBP (191, 208, and 229 YBP, respectively). TMRCAs were also estimated for the ungrouped strains, ranging from 157 to 404 YBP (Supplementary Fig. 5).

Because most Ecuadorian *Mtb* isolates were closely related to the 4.3.2/3 and 4.1.1 sub-lineages, we focused on all isolates belonging to these sub-lineages. The TMRCA estimates for the 4.3.2/3 and 4.1.1 sub-lineages were 470 YBP (95% HPD, 358 to 720) and 450 YBP (95% HPD, 349 to 647), respectively. Within the 4.1.1 sub-lineage, TGC_15 members have a TMRCA of 291 YBP, while those from TGC_37 have a TMRCA of 254 YBP. Isolates from TGC_14 had a TMRCA of 238 YBP, whereas ungrouped isolates S0516 and S2193 had TMRCA estimates of 482 and 354 YBP, respectively. In the 4.3.2/3 sub-lineage, TMRCA estimates ranged from 242 to 295 YBP, covering six TGCs: TGC_26 (289 YBP), TGC_18 (242–265 YBP), TGC_30 (249 YBP), TGC_20 (271–293 YBP), TGC_21 (295 YBP), and TGC_29 (292 YBP), and one ungrouped isolate (S2192, 289 YBP). Notably, isolate S2192 was included in the clade containing TGC_26, and TGC_18 was divided into three clades.

In the 4.1.2.1 sub-lineage, isolates have a TMRCA estimated to be between 237 and 292 YBP, including 263 YBP for isolates corresponding to TGC_05, 237 to 273 YBP for TGC_03, and 292 YBP for TGC_02. The S-type

sub-lineage had a TMRCA of 257 YBP (TGC_11), while isolates corresponding to the 4.8 sub-lineage had a TMRCA of 278 YBP (TGC_48) and 269 YBP (TGC_47). Isolate S0137, which formed a clade with TGC_48, had a TMRCA of 340 YBP (Supplementary Fig. 6).

Discussion

Human-adapted MTB strains show a high degree of genomic conservation but vary in geographic distribution, virulence, transmissibility, and drug resistance patterns[41]. To better understand the transmission pathways within our study population, we analyzed the sequences of 503 *Mtb* isolates of the L4 lineage, mostly from neighboring regions in Ecuador and other areas worldwide. It should be noted that the number of isolates varies significantly between countries. For instance, El Salvador and Chile have only one sample each, while countries such as Brazil and Paraguay have more than one hundred. This heterogeneity may affect the analysis of geographic distribution and limit the interpretation of the results. Our analysis of high-quality genomes revealed that 63.8% of the isolates were sensitive to all drugs used for TB treatment, whereas the remaining 36.2% were resistant to at least one drug. The genomes of all the isolates studied belonged to the Euro-American lineage, with the most common sub-lineages being 4.3.2/3 (35.4%), 4.1.2.1 (22.7%), 4.4.1 (12.7%), and 4.1.1. (10.7%). These results were congruent with those of previous studies in Ecuador that used the MIRU-VNTR strategy for genotyping circulating MTBC strains and also identified other *Mtb* lineages, including L2.1, and L4.2[32,33,37]. The distribution of these sub-lineages suggests that both historical European roots[24,42,43] and migratory processes[18,44,45] contributed to their presence in South and Central America and the Caribbean. Furthermore, we identified 19 genes that harbor mutations associated with resistance, predominantly linked to resistance to first-line drugs, thus enhancing the local genetic data for TB research[13,46,47].

Functional characterization of genes is crucial for understanding how microorganisms such as *M. tuberculosis* adapt and survive in their hosts. Annotation of protein-encoding genes revealed that genes mainly associated with Cofactors, Vitamins, and Prosthetic groups, fatty acids, Lipids, and Isoprenoids, and Amino acids, and derivatives were most representative. Similar findings have been reported in other *M. tuberculosis* populations, where genes related to energy production and conversion, amino acid transport and metabolism, and lipid transport and metabolism are highly represented[48–50]. The wide conservation of these genes indicates their importance in interactions between bacteria and their human hosts. Specifically, during mycobacterial persistence, when the host–pathogen struggles for nutrient and immune recognition, these genes play a crucial role in ensuring the survival and adaptability of bacteria. On the other hand, we found that certain genes associated with the virulence of *Mycobacterium tuberculosis* like ESAT-6-like protein EsxS (63.6%), Acid and phagosome-regulated protein Apr AB (69.3%), and Chorismate mutase I were absent in the Ecuadorian isolates which would suggest possible affectation in adaptation and persistence of *Mtb*, due ESAT-6-like protein EsxS is related with the modulation of host immune responses[51,52]; the Apr ABC locus modulates pH-driven adaptation to the macrophage phagosome[53,54], and Chorismate mutase I is involved in inhibiting intrinsic apoptotic cell death of macrophages, playing a key role in the pathogenesis of TB[55,56].

Utilizing WGS in spanning network analysis enables researchers to connect TB transmission events[57,58], providing insights into the spatial and temporal dynamics of TB transmission and identifying individuals and locations that play a critical role[59,60]. A recent population-based sequencing approach realizes a critical analysis of the utility of a pairwise distance threshold of < 12 SNPs and suggests that in scenarios with higher transmission rates, it is necessary to comprehend long-term transmission dynamics, adhering to a strict transmission 12 SNPs threshold is not advisable[61]; however, contrary to several studies supporting the utility of pairwise distance threshold of < 12 SNPs to identify recent transmission events[61–64]. We examined the possible transmission networks between *Mtb* isolates from Ecuador and from other 19 countries. Evidence suggests the potential spread of Ecuadorian *Mtb* isolates to individuals in different countries, based on the most significant clustering proportion confined to specific geographical locations. Combining genetic and epidemiological data could facilitate TB transmission management, particularly in migrant communities where socio-epidemiological changes due to migration may increase transmission complexity[65].

Migration fluxes to and from high-burden countries significantly influence TB incidence, potentially leading to disease reactivation[61,66]. TB case clusters may involve autochthonous, mixed multinational, or cases among foreign-born individuals concentrated in a specific country[67,68]. Furthermore, *Mtb* can exhibit clonal transmission between hosts and establish clonal infections within a single host, with limited genetic diversification during infection or reactivation[69–72]. Our study observed clonality among Ecuadorian isolates with Colombian or Latin American migrants in Spain, indicating a potential transmission route involving direct contact between migrants, particularly in shared workplaces, such as plantations, factories, or restaurants. Surprisingly, three isolates corresponding to female patients were found in both scenarios, indicating that there might have been a relationship among the individuals, which might have been the cause of disease transmission. These findings suggest possible transnational transmission events involving Ecuador and its border countries and frequent migration destinations, highlighting the need to strengthen disease surveillance to reduce the possibility of more dangerous strains of tuberculosis entering the country[73,74].

Genome comparisons offer valuable insights into the molecular mechanisms that bacteria employ to survive and multiply within intracellular or extracellular host environments and to induce lesions and diseases[75]. However, our understanding of the virulence factors expressed by *Mtb* is limited, and genetic variations resulting from selection pressure may affect their expression. Despite these challenges, performing this type of analysis may contribute to our understanding of how these factors function under local conditions[76]. Our study identified 303 SNPs located within 103 genomic regions associated with virulence, with particular emphasis on genes such as *plcA, plcB, pks7, pks12, phoR*, and *PPE46* which displayed the highest number of SNPs. These genes play pivotal roles in lipid metabolism, which is crucial for *Mtb* virulence of *Mtb* and is an integral component of the complex mycobacterial cell envelope[77,78]. Variations in these genes could affect their functionality, affecting the

ability of the bacterium to evade host immune defense, induce necrosis in macrophages, and modulate virulence, particularly the PhoP-PhoR two-component system, potentially contributing to the pathogenic capabilities of the bacterium[79,80].

Phylodynamic analysis is useful for understanding TB strain variation and dynamics in various countries[81]. The TMCRA and many TGC datasets suggest that TB isolates sampled in Ecuador can trace their ancestry back hundreds of years, indicating a complex evolutionary situation and implying the possibility that different TB isolates have caused infections at various time intervals from varied sources. The Ecuadorian *Mtb* data reported frequencies of 4.3.2/3 and 4.1.1, with TMRCA estimated to range from 450 to 470 YBP. These data indicate the possible ancient roots of the isolates. Nonetheless, it should be mentioned that these calculations only depict TB ancestral isolates within each sub-lineage, and the origin of TB in Ecuador could have been even earlier, before the arrival of Europeans, indicating a long history of human-pathogen co-evolution in the region, consistent with that reported in the ancient Andean population[82]. While some isolates appear to have signs of long-living existence and evolution, and thus historical transmission, others seem to have appeared more recently, suggesting current transmission. Genetic diversity among Ecuadorian *Mtb* isolates is likely to arise from transcontinental interactions, human migration, and various factors that promote the circulation of local and global TB isolates. Thus, additional studies integrating genomic data from various geographic sites as well as comprehensive epidemiological information could help trace the origin and dissemination routes of TB isolates in Ecuador more accurately.

Although our study offers valuable insights, it had several limitations. The primary limitation is the modest sample size of 88 *Mtb* isolates from cultured samples, a small fraction compared to the total number of TB cases reported in Ecuador in 2021 (5595). Furthermore, the lack of comprehensive epidemiological information, including contact tracing details, restricted our analysis of transmission dynamics. To enhance our understanding, future research should aim to combine genomic data with additional epidemiological details to uncover potential *Mtb* transmission pathways for *Mtb*. Moreover, our sequencing focused on cultured sputum isolates, a standard approach in *Mtb* genomic epidemiology that may only partially capture the complete spectrum of mycobacteria in the lungs, thus limiting our ability to capture the intricacies of within-host variations in individual infections[83] Longitudinal studies have the potential to enhance sequencing analyses and to uncover variations in resistance genes and virulence factors, thereby assisting in refining treatment strategies. These insights hold promise for alleviating the global TB burden.

To the best of our knowledge, this is the first study to establish local *Mtb* transmission networks in Ecuador using whole-genome analysis. Our findings reinforce and contribute to the knowledge of transmission networks previously characterized in Ecuador based on the MIRU-VNTR approach[32,33,37,84]. Our study provides valuable insights into the genomic and epidemiological characteristics of *Mtb* isolates from Ecuador. Additionally, our analysis identified drug-resistant isolates and transmission events between individuals and across borders, underscoring the need for more extensive whole-genome sequencing and network analyses to guide public health interventions.

Methods

Genome database of the Mtb samples

This study analyzed raw reads from the genomes of 88 clinical *Mtb* isolates of the L4 lineage, sequenced in a previous study[38], corresponding to BioProject PRJNA827129. These isolates were collected conveniently between 2019 and 2021 from private laboratories and the National Reference of Mycobacteria at the National Institute of Public Health Research "Leopoldo Izquieta Pérez" (INSPI-LIP) across the different provinces in Ecuador. A significant proportion of the samples came from Guayaquil, accounting for 81.8% of the isolates. This city represents epidemiologically more than half of the TB cases in the country[85]. The remaining samples were collected from Babahoyo (5.6%), El Empalme (3.4%), and Quito (2.3%), highlighting the geographical spread and prevalence of TB in these regions. A small number of isolates came from Chone, Duran, Guaranda, Machala, and Nueva Loja, accounting for 1.1% of the total isolates.

Additionally, the study included 415 publicly available sequences previously characterized as *Mtb* isolates of the L4 lineage from 19 countries identified as significant in the Ecuadorian context of continental and transoceanic migration. Among the continental countries, the study included isolates from Argentina (n = 18), Brazil (n = 84), Canada (n = 41), Colombia (n = 8), Guatemala (n = 16), Mexico (n = 35), Panama (n = 4), Paraguay (n = 67), Peru (n = 44), and the USA (n = 6). Isolates from Hungary (n = 5), the Netherlands (n = 12), Portugal (n = 8), Spain (n = 59), and the United Kingdom (n = 5) were included. Notably, among the 59 Spanish isolates, some were identified previously from Latin American migrants who had settled in Spain years earlier, including individuals from Bolivia (n = 20), Colombia (n = 8), Ecuador (n = 8), Chile (n = 1), and Honduras (n = 1), as detailed in reference[61]. The accession numbers and distributions of these countries are listed in Supplementary Table 1, 2. All protocols used in this article were approved by The University Espiritu Santo Review Board under code 2022-001A.

Genome assembly and annotation

To ensure high-quality genomic data, raw reads from the 503 *Mtb* genomes were initially processed with rigorous quality control using Fastp v0.23.4[86,87] and Kraken v2[88] for species confirmation and contamination screening, ensuring that only *Mtb*-specific reads were processed.

Pangenome construction

Pangenomic analysis was carried out using the Panaroo pipeline[89] from the GFF archives annotated by Prokka v1.14.16[90] using the H37Rv (NC_000962.3) *M. tuberculosis* reference genome and default parameters for

clustering to define gene families and identify core genes present in 99% of the isolates. The pangenome was divided into core, accessory, and unique genes, which were categorized based on their presence in isolates.

Variant calling analysis

The cleaned reads were processed using the MTBseq pipeline[91] with standard input parameters to map the reads to the *Mtb* H37Rv reference genome (NC_000962.3). Briefly, this tool involves BWA-mem and SAMtools for mapping, GATK v3 for base call recalibration, and realigning reads around insertions and deletions (InDels), followed by SAMtools mpileup for Single Nucleotide Polymorphisms (SNPs) and InDel calling (using parameters B and d 1000). Only high-quality genomes were processed after accomplishing the following criteria: mean coverage greater than 20x, read depth less than 5x, and reference genome coverage of > 95%. MTBseq analysis facilitated sub-lineage classification and the construction of a genetic distance matrix to identify transmission groups.

Virulence analysis

Ecuadorian *Mtb* isolates were analyzed to identify mutations in the genes involved in host adaptability and virulence. Genomes from the isolates were assembled from high-quality reads using the Unicycler assembly pipeline[92] with a minimum contig size of 300 bp and polished using Pilon[93]. SNPs within high-confidence and repetitive genomic regions, such as the PE/PPE gene family, were confirmed via visualization using Integrated Genome Viewer[94,95]. Virulence-related proteins were analyzed using the Virulence Factor Database (VFDB) in the Pathosystems Resource Integration Center (PATRIC) online (https://www.patricbrc.org/).

Drug resistance prediction and lineage classification

The TB-Profiler v4 pipeline uses the BCF tool for variant calling and predicts drug resistance-associated mutations. These include resistance to first-line drugs such as rifampicin (RIF), isoniazid (INH), pyrazinamide (PZA), and ethambutol (EMB), and second-line drugs such as fluoroquinolones (FQs), streptomycin (STR), ethionamide (ETH), and aminoglycosides (including second-line drugs, amikacin, kanamycin, and capreomycin). According to World Health Organization guidelines, isolates exhibiting resistance solely to INH are classified as isoniazid-resistant TB (HR-TB), those resistant to RIF as rifampicin-resistant TB (RR-TB), and those with resistance to both INH and RIF as multidrug-resistant TB (MDR-TB). Isolates that demonstrated resistance to any FQ in addition to MDR or RR status were categorized as pre-extensively drug-resistant TB (pre-XDR-TB).

Phylogenetic analysis and population structure of Mtb isolates

Phylogenetic reconstruction was conducted using concatenated SNPs aligned genomically using the MTBseq tool. Repetitive regions (PPE/PE-PGRS genes), consecutive indels, and genes involved in antibiotic resistance were excluded from further phylogenetic analysis. The optimal substitution model for this SNP alignment was determined using ModelTest-NG v0.1.7[96]. Following this, the phylogenetic tree was reconstructed using the maximum-likelihood (ML) method via RAxML-NG[97], utilizing the general time-reversible model with gamma-distributed rate heterogeneity (GTR + GAMMA) and supported by 1000 bootstrap replicates. To address the potential ascertainment bias resulting from the exclusive use of polymorphic sites and consequent rescaling of tree branches, ascertainment bias correction was applied by specifying the number of invariant sites[98], as outlined at https://github.com/conmeehan/pathophy. The phylogenetic tree was visualized using the Interactive Tree Of Life (iTOL) v6.6[99]. The genome of *Mycobacterium microti* genome (accession number: SRR3647357) was used as an outgroup to root the tree.

Additionally, a population structure analysis was conducted using the rhierBAPS package[100], which was set to a maximum depth of two and n.pops of 20. This analysis employed a hierarchical nested clustering approach based on genetic data, specifically prioritizing SNP loci that displayed a minor allele in at least two sequences, to effectively identify subpopulations or clusters.

Transmission cluster analysis and modeling of the genetic clustering network

Transmission genomic clusters (TGCs) were identified using concatenated high-quality SNPs processed using MTBseq. We applied a pairwise distance threshold of 12 SNPs between isolates to achieve optimal population clustering, a standard established in previous genomic TB studies[61–63,101]. TGCs were categorized based on their size into small (fewer than three isolates), medium (three to five isolates), and large (more than five isolates). Visualization of the resulting SNP alignment for each cluster was used to infer a genetic network. We used a parsimony-based algorithm for network reconstruction using PopART software[102] because of the monomorphic and non-recombining behavior of *Mtb,* as well as the potential of the sample dataset, including the original genotype. We compared the distribution network of isolates using a median-spanning network (MSN) and median-joining network (MJN).

For clusters of at least three samples, we utilized TransFlow[103] to reconstruct the transmission network, enhancing our understanding of local transmission dynamics. TransFlow integrates genomic data with epidemiological factors, such as sampling dates and geographic coordinates, to map spatial connectivity among isolates. Additionally, to mitigate the bias arising from the use of lineage-specific reference genomes, TransFlow incorporates the PANPASCO pipeline[104]. This pipeline employs a computational pan-genome consisting of 146 complete MTBC genomes from major lineages 1–4, facilitating accurate pairwise SNP distance calculations. This methodology ensures a thorough and representative analysis of genetic variation across MTBC populations.

Phylodynamics of TB transmission clusters

To study the temporal dynamics of the different sub-lineages identified in the Ecuadorian *Mtb* L4 isolates, with a particular focus on uncovering historical introduction events, a time-calibrated phylogeny was inferred using

BEAST v1.10.4[105], utilizing collection and tip dates from the isolates. The XML input file necessary for analysis was generated by concatenating SNPs derived from MTBseq and processed using BEAUTi. This file was adjusted to specify the number of invariant sites following guidance from the BEAST User Forum.

To assess the temporal signal of the sequence alignments, we used TempEst v1.5.3[106]. The dating analysis employed the general time-reversible plus gamma distribution (GTR + GAMMA) substitution model coupled with a strict molecular clock and coalescent constant-size demographic model. Markov chain Monte Carlo (MCMC) simulations were conducted for 250 million iterations, with a 10% burn-in phase and samples taken every 10,000 generations. This approach facilitates independent evaluation of chain convergence.

The analysis results were summarized and convergence was confirmed using Tracer v1.6[107], ensuring that all essential parameters achieved an effective sample size (ESS) of over 200. The Maximum Clade Credibility (MCC) tree was then computed using TreeAnnotator v2.5.0, providing a statistically supported phylogenetic tree with time-calibrated estimates.

Data availability

The raw reads from *Mtb* isolates isolated in Ecuador and other 19 countries analyzed in this study are available in the SRA database (http://www.ncbi.nlm.nih.gov/sra) under BioProject accession numbers PRJNA827129, PRJEB23245, PRJEB23681, PRJEB27366, PRJEB29069, PRJEB44165, PRJEB48543, PRJEB50999, PRJEB7669, PRJNA37301, PRJNA599957, PRJNA628024, PRJNA227755, PRJNA227756, PRJNA755956, PRJNA272873, PRJNA824124, PRJNA707145, PRJNA870648, PRJNA270004, PRJNA422870, and PRJNA438689. In addition, *Mtb* sequences from different countries with the SRA IDs listed in Table S1-2 were used. The protocols used are described in this article. Eleven supplementary tables and six supplementary figures are available in the online version of the manuscript.

Received: 30 September 2024; Accepted: 8 January 2025
Published online: 30 January 2025

References

1. World Health Organization. Global Tuberculosis Report, 2022 [Internet]. Available from: https://www.who.int/teams/global-tuberculosis-programme/tb-reports (2022).
2. Lima, S. V. M. A. et al. Impact of the COVID-19 pandemic on the diagnosis of tuberculosis in Brazil: Is the WHO end TB strategy at risk?. *Front. Pharmacol.* https://doi.org/10.3389/fphar.2022.891711 (2022).
3. Jones, A. J. et al. Impact of COVID-19 on diagnosis and testing for TB in a high-resource, low-burden setting. *Int. J. Tuberc. Lung Dis.* **26**(9), 888–890 (2022).
4. World Bank. Incidence of tuberculosis (per 100,000 people) - Ecuador | Data [Internet]. https://data.worldbank.org/indicator/SH.TBS.INCD?end=2021&locations=EC&name_desc=false&start=2002&view=chart (2023).
5. Mushomi, J. A. et al. Impact of coronavirus disease (COVID-19) crisis on migrants on the move in Southern Africa: Implications for policy and practice. *Health Syst. Reform.* **8**(1), e2019571 (2022).
6. Mellado-Sola, I. et al. Impact of coronavirus pandemic on tuberculosis and other imported diseases screening among migrant minors in Spain. *Trop. Med. Infect. Dis.* **8**(1), 28 (2023).
7. Jokisch B. migrationpolicy.org. 2007. Ecuador: Diversity in Migration. https://www.migrationpolicy.org/article/ecuador-diversity-migration (2023).
8. Herrera G. Migration and migration policy in Ecuador in 2000–2021 (2022).
9. Jokisch BD. migrationpolicy.org. 2023. Ecuador Juggles Rising Emigration and Challenges Accommodating Venezuelan Arrivals (2023).
10. Cavalcante, S. C. et al. Community-randomized trial of enhanced DOTS for tuberculosis control in Rio de Janeiro, Brazil. *Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis.* **14**(2), 203–209 (2010).
11. Yang, T. et al. SAM-TB: a whole genome sequencing data analysis website for detection of Mycobacterium tuberculosis drug resistance and transmission. *Brief. Bioinform.* **23**(2), bbac030 (2022).
12. Lam, C. et al. Value of routine whole genome sequencing for *Mycobacterium tuberculosis* drug resistance detection. *Int. J. Infect. Dis.* **113**, S48-54 (2021).
13. Wang, L. et al. Whole-genome sequencing of *Mycobacterium tuberculosis* for prediction of drug resistance. *Epidemiol. Infect.* **150**, e22 (2022).
14. Brown, A. C. Whole-genome sequencing of Mycobacterium tuberculosis directly from sputum samples. In *Mycobacteria Protocols* (eds Parish, T. & Kumar, A.) 459–80 (Springer US, 2021). https://doi.org/10.1007/978-1-0716-1460-0_20.
15. Torres Ortiz, A. et al. Genomic signatures of pre-resistance in *Mycobacterium tuberculosis*. *Nat. Commun.* **12**(1), 7312 (2021).
16. Freschi, L. et al. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. *Nat. Commun.* **12**(1), 6099 (2021).
17. Stucki, D. et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sub-lineages. *Nat. Genet.* **48**(12), 1535–1543 (2016).
18. Brynildsrud, O. B. et al. Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Sci. Adv.* **4**(10), eaat586 (2018).
19. Ajawatanawong, P. et al. A novel Ancestral Beijing sub-lineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sub-lineages. *Sci. Rep.* **23**(9), 13718 (2019).
20. Ektefaie, Y., Dixit, A., Freschi, L. & Farhat, M. R. Globally diverse *Mycobacterium tuberculosis* resistance acquisition: a retrospective geographical and temporal analysis of whole genome sequences. *Lancet Microbe* **2**(3), e96-104 (2021).
21. Napier, G. et al. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med.* **12**(1), 114 (2020).
22. Coll, F. et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5**(1), 4812 (2014).
23. Senghore, M. et al. Evolution of *Mycobacterium tuberculosis* complex lineages and their role in an emerging threat of multidrug resistant tuberculosis in Bamako, Mali. *Sci. Rep.* **10**(1), 327 (2020).
24. Demay, C. et al. SITVITWEB – A publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect. Genet. Evol.* **12**(4), 755–766 (2012).
25. Santos-Lazaro, D., Gavilan, R. G., Solari, L., Vigo, A. N. & Puyen, Z. M. Whole genome analysis of extensively drug resistant *Mycobacterium tuberculosis* strains in Peru. *Sci. Rep.* **11**(1), 9493 (2021).
26. Tatara, M. B. et al. Genetic diversity and molecular epidemiology of *Mycobacterium tuberculosis* in Roraima state, Brazil. *Am. J. Trop. Med. Hyg.* **101**(4), 774–779 (2019).

27. Lagos, J. et al. Analysis of *Mycobacterium tuberculosis* genotypic lineage distribution in Chile and Neighboring Countries. *PLoS One* **11**(8), e0160434 (2016).
28. Jiménez-Ruano, A. C. et al. Whole genomic sequencing based genotyping reveals a specific X3 sub-lineage restricted to Mexico and related with multidrug resistance. *Sci. Rep.* **11**(1), 1870 (2021).
29. Mejía-Ponce, P. M. et al. Genomic epidemiology analysis of drug-resistant Mycobacterium tuberculosis distributed in Mexico. *PLoS One* **18**(10), e0292965 (2023).
30. Arias, A.P.J. et al. Comparative study of the genetic diversity of Mycobacterium tuberculosis Complex by Simplified Amplified Fragment Length Polymorphism and Mycobacterial Interspersed Repetitive Unit Variable Number Tandem Repeat Analysis. *Rev. Ecuat. Med. Cienc. Biol.*, 39(1) (2018).
31. Zurita, J. et al. Genetic diversity and drug resistance of *Mycobacterium tuberculosis* in Ecuador. *Int. J. Tuberc. Lung Dis.* **23**(2), 166–173 (2019).
32. Garzon-Chavez, D. et al. Prevalence, drug resistance, and genotypic diversity of the Mycobacterium tuberculosis Beijing family in Ecuador. *Microb. Drug Resist.* **25**(6), 931–937 (2019).
33. Garzon-Chavez, D. et al. Population structure and genetic diversity of *Mycobacterium tuberculosis* in Ecuador. *Sci. Rep.* **10**(1), 6237 (2020).
34. Xu, G., Mao, X., Wang, J. & Pan, H. Clustering and recent transmission of *Mycobacterium tuberculosis* in a Chinese population. *Infect. Drug Resist.* **6**(11), 323–330 (2018).
35. Ng, I. C., Wen, T. H., Yang, S. T., Fang, C. T. & Hsueh, P. R. Detecting tuberculosis clusters in urban neighborhoods, Taipei, Taiwan: Linking geographic and genotyping evidence. *Appl. Geogr.* **1**(104), 56–64 (2019).
36. Morey-León, G., Andrade-Molina, D., Fernández-Cadena, J. C. & Berná, L. Comparative genomics of drug-resistant strains of *Mycobacterium tuberculosis* in Ecuador. *BMC Genom.* **23**(1), 844 (2022).
37. Castro-Rodriguez, B. et al. A first insight into tuberculosis transmission at the border of Ecuador and Colombia: a retrospective study of the population structure of Mycobacterium tuberculosis in Esmeraldas province. *Front. Public Health* 12 (2024).
38. Morey-León, G. et al. A precision overview of genomic resistance screening in Ecuadorian isolates of Mycobacterium tuberculosis using web-based bioinformatics tools. *PLoS One* **18**(12), e0294670 (2023).
39. Smith, I. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin. Microbiol. Rev.* **16**(3), 463–496 (2003).
40. Leung, E. C. C. et al. Transmission of multidrug-resistant and extensively drug-resistant tuberculosis in a metropolitan city. *Eur. Respir. J.* **41**(4), 901–908 (2013).
41. Zakham, F., Sironen, T., Vapalahti, O. & Kant, R. Pan and core genome analysis of 183 *Mycobacterium tuberculosis* strains revealed a high inter-species diversity among the human adapted strains. *Antibiotics* **10**(5), 500 (2021).
42. Realpe, T. et al. Population Structure among *Mycobacterium tuberculosis* isolates from pulmonary tuberculosis patients in Colombia. *PLoS One* **9**(4), e93848 (2014).
43. Díaz Acosta, C. C. et al. Exploring the "Latin American Mediterranean" family and the RDRio lineage in *Mycobacterium tuberculosis* isolates from Paraguay, Argentina and Venezuela. *BMC Microbiol.* **19**(1), 131 (2019).
44. Woodman, M., Haeusler, I. L. & Grandjean, L. Tuberculosis genetic epidemiology: A Latin American perspective. *Genes* **10**(1), 53 (2019).
45. Wiens, K. E. et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Med.* **30**(16), 196 (2018).
46. Chernyaeva, E. et al. Genomic variations in drug resistant *Mycobacterium tuberculosis* strains collected from patients with different localization of infection. *Antibiotics* **10**(1), 27 (2020).
47. Vianna, J. S. et al. Whole-genome sequencing as a tool for studying the microevolution of drug-resistant serial *Mycobacterium tuberculosis* isolates. *Tuberculosis* **131**, 102137 (2021).
48. Periwal, V. et al. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of *Mycobacterium tuberculosis* pangenome. *PLoS One* **10**(4), e0122979 (2015).
49. de Cassio Barreto Oliveira, M. & Balan, A. The ATP-binding cassette (ABC) transport systems in *Mycobacterium tuberculosis*: Structure, function, and possible targets for therapeutics. *Biology* **9**(12), 443 (2020).
50. Yang, Z., Zeng, X. & Tsui, S. K. W. Investigating function roles of hypothetical proteins encoded by the *Mycobacterium tuberculosis* H37Rv genome. *BMC Genom.* **20**(1), 394 (2019).
51. Sreejit, G. et al. The ESAT-6 protein of *Mycobacterium tuberculosis* interacts with beta-2-microglobulin (β2M) affecting antigen presentation function of macrophage. *PLoS Pathog.* **10**(10), e1004446 (2014).
52. Mustafa, A. S. Immunological characterization of proteins expressed by genes located in *Mycobacterium tuberculosis*-specific genomic regions encoding the ESAT6-like proteins. *Vaccines* **9**(1), 27 (2021).
53. Abramovitch, R. B., Rohde, K. H., Hsu, F. F. & Russell, D. G. aprABC: a *Mycobacterium tuberculosis* complex-specific locus that modulates pH-driven adaptation to the macrophage phagosome. *Mol. Microbiol.* **80**(3), 678–694 (2011).
54. Singh, P. R. et al. Dual functioning by the PhoR sensor is a key determinant to *Mycobacterium tuberculosis* virulence. *PLoS Genet.* **19**(12), e1011070 (2023).
55. Khanapur, M. et al. *Mycobacterium tuberculosis* chorismate mutase: A potential target for TB. *Bioorg. Med. Chem.* **25**(6), 1725–1736 (2017).
56. Lee, M. H., Kim, H. L., Seo, H., Jung, S. & Kim, B. J. A secreted form of chorismate mutase (Rv1885c) in Mycobacterium bovis BCG contributes to pathogenesis by inhibiting mitochondria-mediated apoptotic cell death of macrophages. *J. Biomed. Sci.* **30**(1), 95 (2023).
57. Yassine, E. et al. Assessing a transmission network of Mycobacterium tuberculosis in an African city using single nucleotide polymorphism threshold analysis. *MicrobiologyOpen* **10**(3), e1211 (2021).
58. López-Cortés, A. et al. Clinical, genomics and networking analyses of a high-altitude native American Ecuadorian patient with congenital insensitivity to pain with anhidrosis: a case report. *BMC Med. Genom.* **17**(13), 113 (2020).
59. Cudahy, P. G. et al. Spatially targeted screening to reduce tuberculosis transmission in high incidence settings: a systematic review and synthesis. *Lancet Infect. Dis.* **19**(3), e89-95 (2019).
60. Alaridah, N. et al. Transmission dynamics study of tuberculosis isolates with whole genome sequencing in southern Sweden. *Sci. Rep.* **9**(1), 4931 (2019).
61. Cancino-Muñoz, I. et al. Population-based sequencing of *Mycobacterium tuberculosis* reveals how current population dynamics are shaped by past epidemics. *eLife* **11**, e76605 (2022).
62. Séraphin, M. N. et al. Direct transmission of within-host *Mycobacterium tuberculosis* diversity to secondary cases can lead to variable between-host heterogeneity without de novo mutation: A genomic investigation. *EBioMedicine* **1**(47), 293–300 (2019).
63. Hall, M. B. et al. Evaluation of Nanopore sequencing for *Mycobacterium tuberculosis* drug susceptibility testing and outbreak investigation: a genomic analysis. *Lancet Microbe* **4**(2), e84-92 (2023).
64. Walker, T. M. et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir. Med.* **2**(4), 285–292 (2014).
65. Abascal, E. et al. Whole genome sequencing–based analysis of tuberculosis (TB) in migrants: rapid tools for cross-border surveillance and to distinguish between recent transmission in the host country and new importations. *Eurosurveillance* **24**(4), 1800005 (2019).

66. Meumann, E. M. et al. Tuberculosis in Australia's tropical north: a population-based genomic epidemiological study. Lancet Reg Health – West Pac (2021).
67. Yang, C. et al. Internal migration and the transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, and genomic analysis. *Lancet Infect. Dis.* **18**(7), 788–795 (2018).
68. Ayabina, D. et al. Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population. *Microb. Genom.* **4**(10), e000219 (2018).
69. Verma, S. et al. Transmission phenotype of *Mycobacterium tuberculosis* strains is mechanistically linked to induction of distinct pulmonary pathology. *PLoS Pathog.* **15**(3), e1007613 (2019).
70. Vadwai, V. et al. Clonal population of *Mycobacterium tuberculosis* strains reside within multiple lung cavities. *PLoS One* **6**(9), e24770 (2011).
71. Bjorn-Mortensen, K. et al. Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. *Sci. Rep.* **6**(1), 33180 (2016).
72. Rodwell, T. C., Kapasi, A. J., Barnes, R. F. W. & Moser, K. S. Factors associated with genotype clustering of *Mycobacterium tuberculosis* isolates in an ethnically diverse region of southern California, United States. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **12**(8), 1917–1925 (2012).
73. Pescarini, J. M. et al. Migration and tuberculosis transmission in a middle-income country: a cross-sectional study in a central area of São Paulo, Brazil. *BMC Med.* **16**(1), 62 (2018).
74. Sanabria, G. E. et al. Phylogeography and transmission of *Mycobacterium tuberculosis* spanning prisons and surrounding communities in Paraguay. *Nat. Commun.* **14**(1), 303 (2023).
75. Sapriel, G. & Brosch, R. Shared pathogenomic patterns characterize a new phylotype, revealing transition toward host-adaptation long before speciation of *Mycobacterium tuberculosis*. *Genome Biol. Evol.* **11**(8), 2420–2438 (2019).
76. Forrellad, M. A. et al. Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* **4**(1), 3–66 (2013).
77. Quadri, L. E. N. Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. *Crit. Rev. Biochem. Mol. Biol.* **49**(3), 179–211 (2014).
78. Le Chevalier, F. et al. Revisiting the role of phospholipases C in virulence and the lifecycle of Mycobacterium tuberculosis. *Sci. Rep.* **5**(1), 16918 (2015).
79. Cimino, M. et al. Identification of DNA binding motifs of the *Mycobacterium tuberculosis* PhoP/PhoR two-component signal transduction system. *PLoS One* **7**(8), e42876 (2012).
80. Malaga, W. et al. Natural mutations in the sensor kinase of the PhoPR two-component regulatory system modulate virulence of ancestor-like tuberculosis bacilli. *PLoS Pathog.* **19**(7), e1011437 (2023).
81. Kühnert, D. et al. Tuberculosis outbreak investigation using phylodynamic analysis. *Epidemics* **25**, 47–53 (2018).
82. Joseph, S. K. et al. Genomic evidence for adaptation to tuberculosis in the Andes before European contact. *iScience* **26**(2), 106034 (2023).
83. Kaplan, G. et al. *Mycobacterium tuberculosis* growth at the cavity surface: a microenvironment with failed immunity. *Infect. Immun.* **71**(12), 7099–7108 (2003).
84. Garcés E, Cifuentes L, Franco G, Romero-Sandoval N, Arias PJ. Study of the distribution of lineages of Mycobacterium tuberculosis in a prison in Guayaquil, Ecuador (2023)
85. Quillupangui S. El Comercio. 2021. Más de la mitad de pacientes con tuberculosis están en Guayas. Available from: https://www.elcomercio.com/actualidad/ecuador/tuberculosis-guayas-enfermedad-pacientes-salud.html (2024).
86. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**(2), e107 (2023).
87. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890 (2018).
88. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**(3), R46 (2014).
89. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **22**(21), 180 (2020).
90. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**(14), 2068–2069 (2014).
91. Kohl, T. A. et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of *Mycobacterium tuberculosis* complex isolates. *PeerJ.* **13**(6), e5895 (2018).
92. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**(6), e1005595 (2017).
93. Walker, B. J. et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**(11), e112963 (2014).
94. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**(2), 178–192 (2013).
95. Dixit, A. et al. Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. *Sci. Rep.* **9**(1), 5602 (2019).
96. Darriba, D. et al. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* **37**(1), 291–294 (2020).
97. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**(21), 4453–4455 (2019).
98. Coscolla, M. et al. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microb. Genom.* **7**(2), 000477 (2021).
99. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**(W1), W293–W296 (2021).
100. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**(5), 1224–1228 (2013).
101. Jajou, R. et al. Towards standardisation: comparison of five whole genome sequencing (WGS) analysis pipelines for detection of epidemiologically linked tuberculosis cases. *Eurosurveillance* **24**(50), 1900130 (2019).
102. Leigh, J. W. & Bryant, D. popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**(9), 1110–1116 (2015).
103. Pan, J. et al. TransFlow: a Snakemake workflow for transmission analysis of Mycobacterium tuberculosis whole-genome sequencing data. *Bioinformatics* **39**(1), btac785 (2023).
104. Jandrasits, C., Kröger, S., Haas, W. & Renard, B. Y. Computational pan-genome mapping and pairwise SNP-distance improve detection of *Mycobacterium tuberculosis* transmission clusters. *PLoS Comput Biol.* **15**(12), e1007527 (2019).
105. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 110. *Virus Evol.* **4**(1), vey016 (2018).
106. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**(1), vew007 (2016).
107. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst Biol.* **67**(5), 901–4 (2018).

Acknowledgements
We thank all the researchers from the different BioProjects used in this study for the availability of sequences on the SRA website. Omics Sciences Laboratory of the Faculty of Health Sciences for server availability.

Author contributions
GML and PMMP conceived and designed the study, performed the data analysis (bioinformatics processing of the raw sequencing data), and drafted the original manuscript. DAM, EGM, and JCFC performed data analysis and drafted the manuscript. CLC provided advice on data analysis. PF and LB conceived and designed the study, wrote the main manuscript and reviewed the manuscript. GM and DAM performed the funding acquisition. All the authors have read and approved the final manuscript.

Funding
GM is a doctoral student in the PEDECIBA program. LB and PF are members of PEDECIBA and the Sistema Nacional de Investigadores (SNI) of ANII. PMMP received financial support from the Mexican National Council for Humanities Science and Technology (CONAHCYT). This work was partially supported by the Ciencia de Frontera (CONAHCYT) – 319590 projects granted to CLC. The funders played no role in the study design, data collection, analysis, decision to publish, or manuscript preparation. This study received no specific grants from any funding agencies.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
The need to obtain informed consent was waived by The University Espíritu Santo Review Board under code 2022-001A, because the study included only isolates from a collection. The ethics committee ruled out informed consent because the data of the isolates were anonymized, and no patient data were disclosed. This study was conducted under the ethical principles outlined in the Declaration of Helsinki.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86079-8.

Correspondence and requests for materials should be addressed to G.M.-L. or L.B.