Preplanned Studies

Genotypic Characteristics of *Mycobacterium Tuberculosis* Based on Whole Genome Sequencing — Southern Xinjiang Uygur Autonomous Region, China, 2021–2023

Palidanguli Abudureheman[1]; Senlu Wang[2]; Le Wang[2]; Gulina Badeerhan[2]; Kailibinuer Wumaier[2]; Nianqiang Liu[2]; Xiaowei Ma[1]; Xijiang Wang[2,#]

Summary

What is already known about this topic?

Currently, *Mycobacterium tuberculosis* is classified into 9 major lineages, each exhibiting distinct geographical distribution patterns and transmission characteristics. In China, Lineage 2 predominates, while Lineage 3 is primarily distributed in the Xinjiang region.

What is added by this report?

This study integrated multidimensional analyses incorporating patient characteristics, strain lineages, drug resistance profiles, and transmission networks, providing a comprehensive elucidation of *Mycobacterium tuberculosis* molecular epidemiology.

What are the implications for public health practice?

Molecular epidemiological insights into *Mycobacterium tuberculosis* transmission in Southern Xinjiang enable precision tuberculosis control.

ABSTRACT

Introduction: Southern Xinjiang Uygur Autonomous Region, China, experiences a substantial tuberculosis burden, yet comprehensive genotypic characterization of *Mycobacterium tuberculosis* strains in this region remains limited.

Methods: This study collected *Mycobacterium tuberculosis* strains and corresponding epidemiological data from patients between 2021 and 2023. Bacterial isolates underwent whole-genome sequencing using the Illumina next-generation sequencing platform. We constructed phylogenetic trees using iqtree and generated minimum spanning trees based on GraphSNP analysis, applying a clustering threshold of 12 single nucleotide polymorphisms (SNPs) to identify transmission clusters.

Results: Lineage 2 emerged as the predominant strain type, accounting for the majority of isolates, followed by Lineages 4 and 3. Notably, Lineage 2 demonstrated significantly elevated clustering rates compared to other lineages. Genetic diversity analysis revealed that Lineage 2 strains exhibited the most limited intra-lineage variation, whereas Lineage 3 displayed the greatest genetic heterogeneity among strains.

Conclusion: Our investigation demonstrates substantial genetic polymorphism among *Mycobacterium tuberculosis* strains circulating in southern Xinjiang. These findings highlight the critical need for enhanced transmission control strategies, with particular emphasis on intensive surveillance and prevention measures targeting Lineage 2 strains.

Tuberculosis (TB), a chronic infectious disease caused by *Mycobacterium tuberculosis* (MTB), continues to pose a substantial global public health challenge. According to the World Health Organization's 2023 Global TB Report, China ranked third in estimated TB incidence among 30 high-burden countries in 2022 (*1*). Within China, Xinjiang Uygur Autonomous Region has consistently maintained one of the highest tuberculosis burdens nationwide. Despite this epidemiological significance, the genotypic characteristics of circulating MTB strains in Xinjiang remain inadequately characterized. To date, nine human-adapted lineages of *Mycobacterium tuberculosis* have been identified (*2–3*), each demonstrating distinct geographical distribution patterns and unique global dissemination trajectories (*4–6*). Given Xinjiang's strategic geographical position at the crossroads of Central Asia, the region has likely developed distinctive MTB lineage characteristics that warrant comprehensive investigation.

We collected *M. tuberculosis* isolates and corresponding patient information from seven counties and cities in southern Xinjiang between 2021 and

2023. Written informed consent was obtained from all participants, and the study received approval from the central ethics review committee prior to sample collection. All isolates underwent subculturing on Lowenstein-Jensen media before DNA extraction using the cetyltrimethylammonium bromide (CTAB) method. Genomic DNA from each isolate was sequenced using Illumina PE 150 technology with 350 bp paired-end reads. For samples where the proportion of sequence classification to *Mycobacterium tuberculosis* complex (MTBC) fell below the established threshold of 90%, subsequent analytical modules were excluded to eliminate potential interference from nontuberculous mycobacteria and sample contamination.

Raw sequencing data underwent quality filtering using Fastp software. Strain identification was performed using Kraken 2, while lineage identification and drug resistance prediction were conducted using TB-Profiler. SNP detection was carried out using Samtools and bcftools. SNP differences were visualized through R-generated heatmaps, and population evolutionary analysis was performed using RAxML.

Supplementary Table S1 (available at https://weekly. chinacdc.cn/) presents the demographic characteristics of 472 isolates across population, temporal, and geographical parameters. The majority of cases (62.5%) occurred in the 56–75 age group, with a relatively balanced gender distribution (46.2% male,
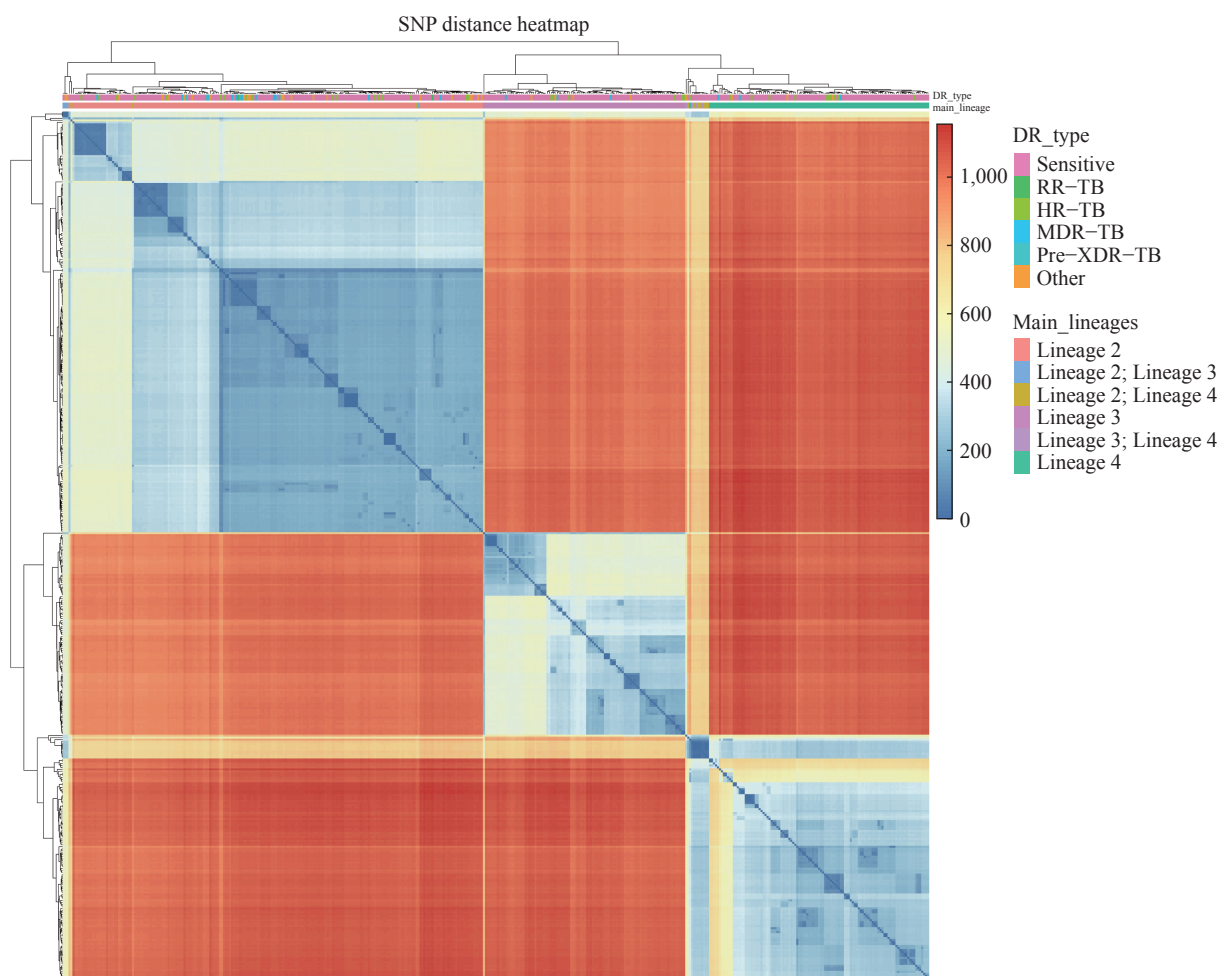

FIGURE 1. SNP distance heatmap.
Note: The hierarchical clustering dendrograms displayed on the left and top margins represent sample clustering patterns based on SNP differences. The upper panel illustrates drug resistance classifications and lineage information for each sample, with corresponding color annotations detailed in the legend. The central heatmap visualizes the SNP difference matrix, where each cell represents the SNP difference between samples corresponding to the horizontal and vertical axes. The color scale on the right indicates the magnitude of SNP differences. The heatmap clearly demonstrates the correlation patterns between lineage distributions and SNP differences.
Abbreviation: SNP=single nucleotide polymorphism.

53.8% female). Of the 472 isolates analyzed, 437 were confirmed as *M. tuberculosis*. Supplementary Table S2 (available at https://weekly.chinacdc.cn/) details the distribution of *M. tuberculosis* lineages and their respective clustering rates (SNP distance <12 served as the clustering criterion). Lineage 2 predominated (46.9%), followed by lineage 4 (25.6%) and lineage 3 (23.3%). Although lineage 2 exhibited a higher clustering rate (25.9%) compared to lineages 3 and 4 (both 19.6%), these differences were not statistically significant (χ^2=2.136, *P*=0.314).

A comprehensive heatmap was constructed to visualize SNP differences among all samples,

incorporating drug resistance classifications and lineage identification data (Figure 1). Additionally, a population evolutionary tree was generated based on whole genome sequencing data, integrating lineage identification and drug resistance prediction results (Figure 2).

Figure 3 illustrates the genetic distance networks within transmission clusters. While most transmission clusters remained geographically localized, four clusters demonstrated cross-regional transmission, exemplified by the largest cluster (C1) spanning Aksu City, Wushi County, and Shache County. Among all clusters, 10 contained 22 strains exhibiting genotypic resistance.
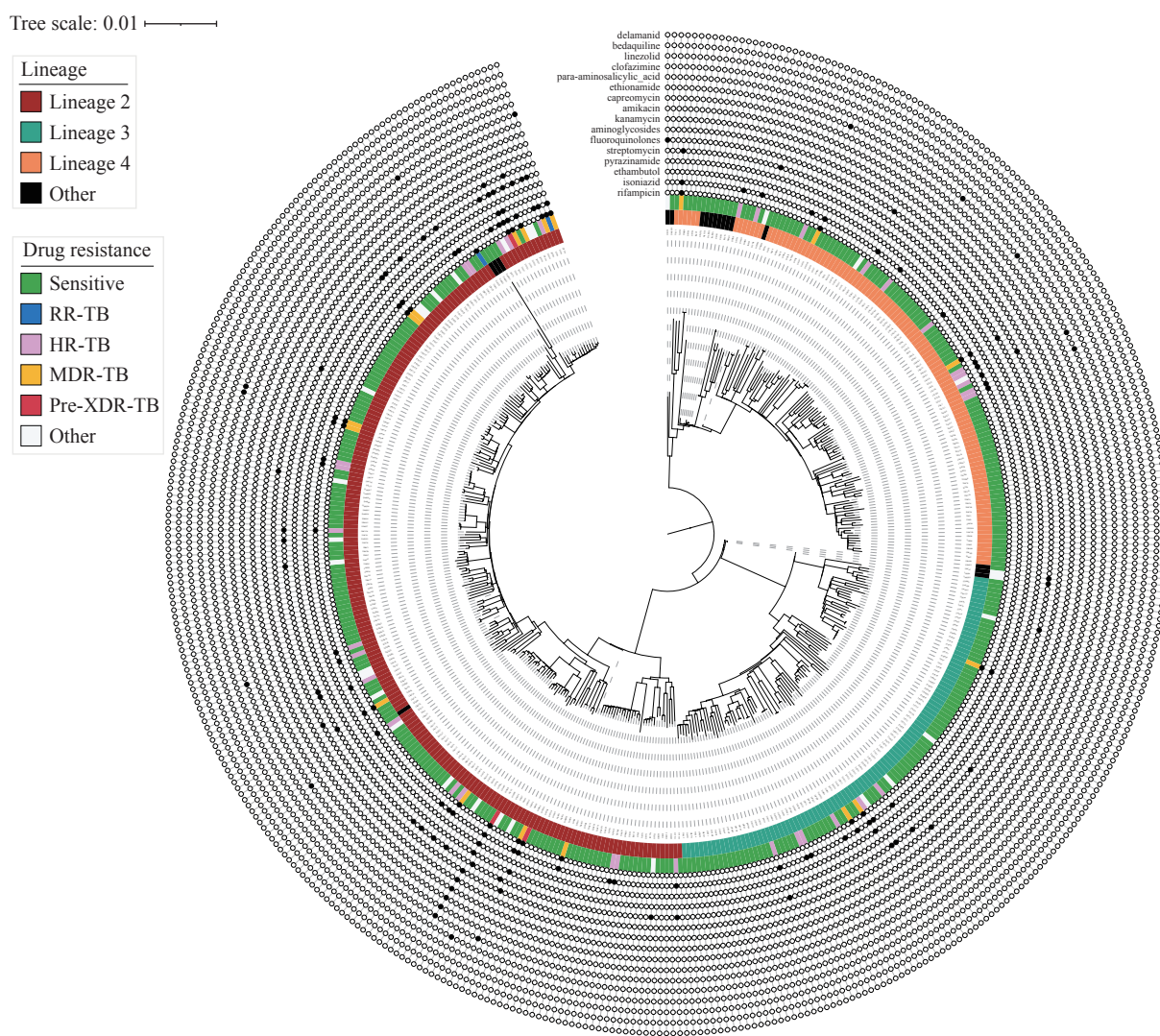

FIGURE 2. Population evolutionary tree diagram.
Note: A population evolutionary tree constructed from whole-genome sequencing SNP data of *Mycobacterium tuberculosis*, with branch lengths representing evolutionary distances. The right panel displays lineage identification and drug resistance prediction results. The color scheme for lineage identification and drug resistance classification is detailed in the legend. For individual drug resistance results, black indicates resistance while gray indicates susceptibility.
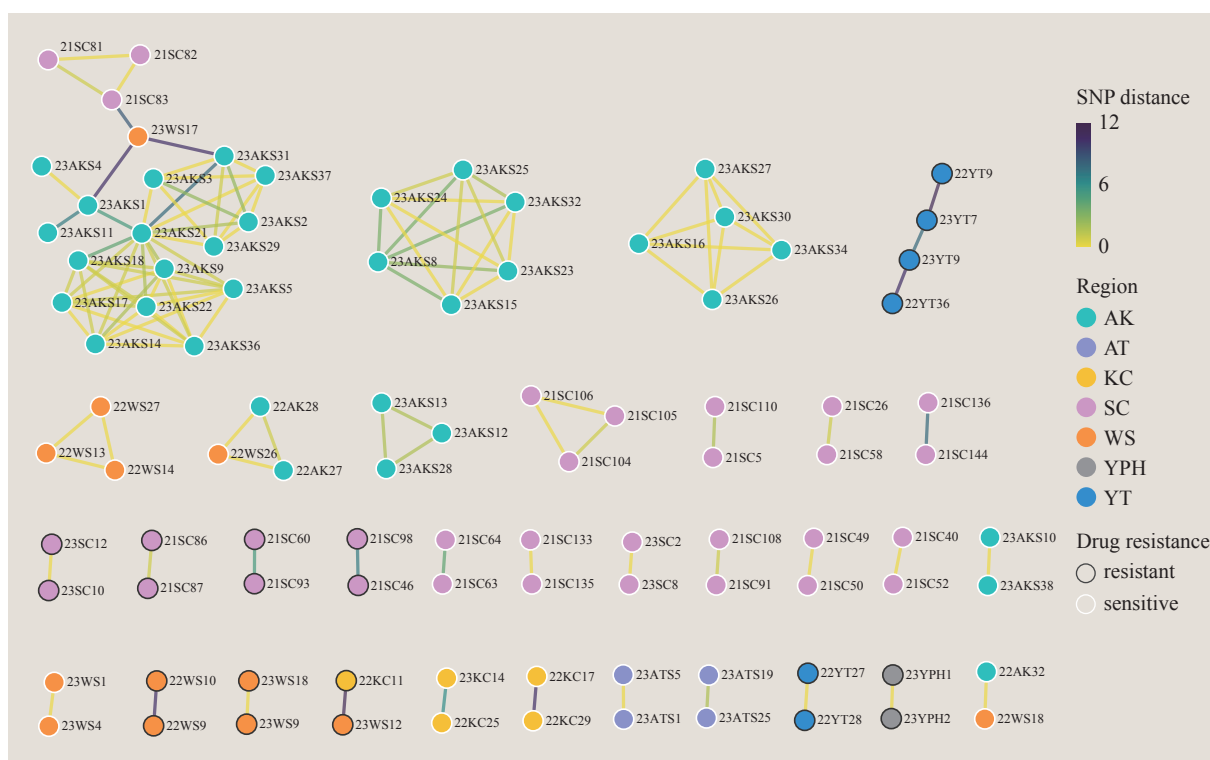Abbreviation: SNP=single nucleotide polymorphism.

FIGURE 3. Visual network of all transmission clusters.
Note: Network visualization depicting strain relationships where individual nodes represent bacterial strains and connecting edges indicate SNP differences between strains. Node interior colors denote sampling regions, while outer ring colors indicate the presence of drug-resistance mutations. Edge colors transition from yellow to green, representing SNP differences ranging from 0 to 12.
Abbreviation: AK=Aksu; WS=Wushi; KC=Kuche; YT=Yutian; YP=Yuepuhu; SC=Shache; AT=Atushi; SNP=single nucleotide polymorphism.

Figure 4A reveals significant lineage-specific variations in genetic diversity, with lineage 2 demonstrating the lowest intra-lineage genetic variation and lineage 3 displaying the highest. Figure 4B demonstrates that while lineage 2 generally exhibits higher clustering rates, at lower thresholds (e.g., 5 SNPs), lineage 3 shows the highest clustering rate and lineage 2 the lowest. These patterns suggest sustained long-term transmission of lineage 2 in the region, whereas lineages 3 and 4 demonstrate evidence of increased recent transmission activity (within the past decade).

DISCUSSION

Understanding the molecular epidemiology of tuberculosis is essential for achieving the World Health Organization's "End TB" strategy goals. This research provides comprehensive molecular-level insights into TB transmission dynamics across southern Xinjiang, revealing important patterns that inform regional control strategies. The Beijing genotype (Lineage 2) predominates in southern Xinjiang, accounting for 46.9% of isolates. This prevalence is notably lower than previous regional studies, including reports by Yuan et al. (7) (57.5%) and Anwaierjiang, A et al. (8) (58.4%), as well as findings from northern and southern China (76.5% and 53.2%, respectively) (9). While Lineages 2 and 4 demonstrate widespread global distribution (10–11), our findings reflect Xinjiang's distinctive position as a historical crossroads along the ancient Silk Road Economic Belt. Notably, Lineage 3, which has been extensively documented throughout historical Silk Road regions and northwestern China (12), was also identified in neighboring provinces, including Tibet and Qinghai (13). The rapid global expansion of Lineage 2 over the past two centuries may be attributed to its enhanced transmission capacity and propensity for drug resistance development, which correlates with the higher clustering rate observed for Lineage 2 in our study (14).

Genetic distance analysis enables precise characterization of tuberculosis transmission patterns and identification of potential multiple transmission
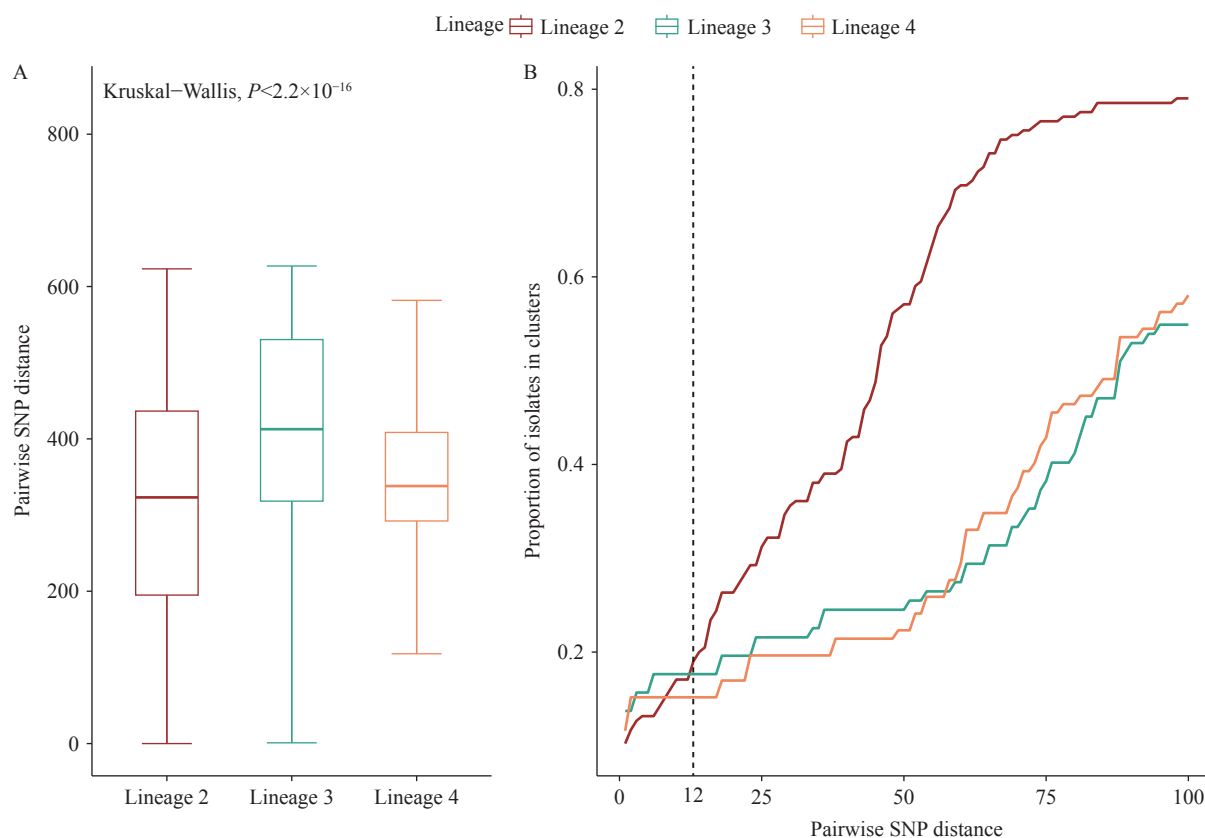
FIGURE 4. Comparison of genetic distance and clustering among different lineages. (A) Box plot distribution of inter-strain SNP differences across the three lineages. (B) Cumulative clustering ratio curves for the three bacterial lineages using varying SNP difference thresholds (1–100) for transmission clustering.
Note: For (B), the clustering rates at different thresholds reflect lineage-specific clustering patterns and indirectly indicate transmission scales across different time periods.

sources within communities. The overall clustering rate of 22.2% observed in southern Xinjiang approximates that reported in Shanghai (25.2%) while substantially exceeding rates documented in Shenzhen (12.2%) (15–16). Our investigation revealed mixed lineage infections across all seven studied counties and cities, indicating elevated levels of recent transmission activity. This transmission pattern likely reflects the complex interplay of local cultural practices, housing conditions, and socioeconomic determinants that facilitate disease spread, warranting comprehensive epidemiological investigation.

As Xinjiang assumes an increasingly pivotal role in China's Belt and Road Initiative, expanding economic cooperation and human mobility with the five Central Asian republics necessitates enhanced infectious disease surveillance systems. This evolving landscape demands strengthened international collaboration in tuberculosis control, particularly through research and development of innovative diagnostic technologies, targeted support for border regions, and establishment of robust cross-border prevention mechanisms. Such coordinated multilateral efforts are fundamental to achieving the ambitious "End TB" strategy objectives by 2030. Future research should incorporate larger sample sizes and extended temporal frameworks to strengthen these preliminary findings and guide evidence-based policy development.

Corresponding author: Xijiang Wang, wxj@xjcdc.com.

¹ School of Public Health, Xinjiang Medical University, Urumqi City, Xinjiang Uygur Autonomous Region, China; ² Xinjiang Uygur Autonomous Region Center for Disease Control and Prevention, Urumqi City, Xinjiang Uygur Autonomous Region, China.

REFERENCES

1. Bagcchi S. WHO's global tuberculosis report 2022. Lancet Microbe 2023;4(1):e20. https://doi.org/10.1016/S2666-5247(22)00359-7.

2. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. Nat Commun 2020;11(1):2917. https://doi.org/10.1038/s41467-020-16626-6.

3. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. Microb Genom 2021;7(2):000477. https://doi.org/10.1099/mgen.0.000477.

4. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 2006;103(8):2869 – 73. https://doi.org/10.1073/pnas.0511240103.

5. Pasipanodya JG, Moonan PK, Vecino E, Miller TL, Fernandez M, Slocum P, et al. *Allopatric* tuberculosis host-pathogen relationships are associated with greater pulmonary impairment. Infect Genet Evol 2013;16:433 – 40. https://doi.org/10.1016/j.meegid.2013.02.015.

6. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu QY, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 2016;48(12):1535 – 43. https://doi.org/10.1038/ng.3704.

7. Yuan L, Mi LG, Li YX, Zhang H, Zheng F, Li ZY. Genotypic characteristics of *Mycobacterium tuberculosis* circulating in Xinjiang, China. Infect Dis 2016;48(2):108 – 15. https://doi.org/10.3109/23744235.2015.1087649.

8. Anwaierjiang A, Wang Q, Liu HC, Yin CJ, Xu M, Li MC, et al. Prevalence and molecular characteristics based on whole genome sequencing of *Mycobacterium tuberculosis* resistant to four anti-tuberculosis drugs from southern Xinjiang, China. Infect Drug Resist 2021;14:3379 – 91. https://doi.org/10.2147/IDR.S320024.

9. Pang Y, Zhou Y, Zhao B, Liu G, Jiang GL, Xia H, et al. Spoligotyping and drug resistance analysis of *Mycobacterium tuberculosis* strains from national survey in China. PLoS One 2012;7(3):e32976. https://doi.org/10.1371/journal.pone.0032976.

10. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol 2018;16(4):202 – 13. https://doi.org/10.1038/nrmicro.2018.8.

11. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet 2013;45(10):1176 – 82. https://doi.org/10.1038/ng.2744.

12. Zhou Y, van den Hof S, Wang SF, Pang Y, Zhao B, Xia H, et al. Association between genotype and drug resistance profiles of *Mycobacterium tuberculosis* strains circulating in China in a national drug resistance survey. PLoS One 2017;12(3):e0174197. https://doi.org/10.1371/journal.pone.0174197.

13. O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. Mol Ecol 2019;28(13):3241 – 56. https://doi.org/10.1111/mec.15120.

14. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nat Genet 2015;47(3):242 – 9. https://doi.org/10.1038/ng.3195.

15. Li M, Lu LP, Jiang Q, Jiang Y, Yang CG, Li J, et al. Genotypic and spatial analysis of transmission dynamics of tuberculosis in Shanghai, China: a 10-year prospective population-based surveillance study. Lancet Reg Health West Pac 2023;38:100833. https://doi.org/10.1016/j.lanwpc.2023.100833.

16. Yang TT, Wang YX, Liu QY, Jiang Q, Hong CY, Wu LK, et al. A population-based genomic epidemiological study of the source of tuberculosis infections in an emerging city: Shenzhen, China. Lancet Reg Health West Pac 2021;8:100106. https://doi.org/10.1016/j.lanwpc.2021.100106.

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY TABLE S1. Demographic characteristics of the 472 isolates in Xinjiang Uygur Autonomous Region, China.

Variables	No. of patients	Constituent ratio (%)
Age (Years)		
<25	27	5.7
26–55	83	17.6
56–75	295	62.5
>75	67	14.2
Sex		
Male	218	46.2
Female	254	53.8
Year		
2021	143	30.3
2022	148	31.4
2023	181	38.3
Region		
Aksu	70	14.8
Kuche	60	12.7
Wushi	59	12.5
Shache	165	35.0
Yuepuhu	21	4.4
Yutian	62	13.1
Atushi	35	7.4

SUPPLEMENTARY TABLE S2. Distribution of lineages and clustering rates among *M. tuberculosis* strains.

Lineages	No. of isolates (constituent ratio)	No. of cases in cluster (clustering rate)	χ^2	P
Lineage 2	205 (46.9%)	53 (25.9%)	2.136	0.314
Lineage 3	102 (23.3%)	20 (19.6%)		
Lineage 4	112 (25.6%)	22 (19.6%)		