Distribution and Identification of *Mycobacterium tuberculosis* Lineage in Kashgar Prefecture

Ai-Min Xu
The First People's Hospital of Kashgar
Chuan-Jiang He
The First People's Hospital of Kashgar
Xiang Chen
The First People's Hospital of Kashgar
Li Li
The First People's Hospital of Kashgar
AniKiz Abuduaini
The First People's Hospital of Kashgar
Zureguli Tuerxun
The First People's Hospital of Kashgar
Yin-Zhong Sha
The First People's Hospital of Kashgar
Aihemaitijiang Kaisaier
The First People's Hospital of Kashgar
Hong-Mei Peng
The First People's Hospital of Kashgar
Ya-Hui Zhen
The First People's Hospital of Kashgar
Su-Jie Zhang
The First People's Hospital of Kashgar
Jing-Ran Xu
The First People's Hospital of Kashgar
Xiao-Guang Zou (✉ zouxiaoguang2021@163.com)
The First People's Hospital of Kashgar

Research Article

Keywords: M.tb lineage, M.tb Sublineage, Branch-Specific SNP, Geographical distribution, Whole genome sequencing (WGS)

Posted Date: November 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-902617/v1

Version of Record: A version of this preprint was published at BMC Infectious Diseases on March 30th, 2022. See the published version at https://doi.org/10.1186/s12879-022-07307-4.

Abstract

Objectives

In order to understand the composition of *Mycobacterium tuberculosis*(*M.tb*) lineage and find specific tags to distinguish lineage of the *M.tb* in Kashgar prefecture, so as to provide a basis for the prevention of tuberculosis in this area.

Methods

Whole genome sequencing (WGS) of *M.tb* clinical strains (161 cases) was conducted. The phylogenetic tree was constructed by Maximum Likelihood (ML) on the basis of single nucleotide polymorphisms (SNPs) and verified via principal component analysis (PCA). The composition structure of *M.tb* in different regions was analyzed by combining geographic information.

Results

The *M.tb* clinical strains were composed of lineage 2 (73/161, 45.34%), lineage 3 (52/161, 32.30%) and lineage 4 (36/161, 22.36%) in Kashgar prefecture. And the 3 lineages were subdivided into 11 sublineages, among which lineage 2 includes lineage 2.2.2/Asia Ancestral 1(9/73, 12.33%),lineage 2.2.1-Asia Ancestral 2(9/73, 12.33%)，lineage 2.2.1-Asia Ancestral 3(18/73, 24.66%) and lineage 2.2.1-Modern Beijing(39/73, 53.42%).Lineage 3 includes lineage 3.2(14/52, 26.92%)and lineage 3.3(38/52, 73.08%)，lineage 4 includes lineage 4.1(3/36, 8.33%)，lineage 4.2(2/36, 5.66%)，lineage 4.4.2(1/36, 2.78%)，lineage 4.5(28/36, 77.78%) and lineage 4.8(2/36, 5.66%)，all of which were consistent with the PCA results. Among the identified 21438 SNPs ,there are 136 markers proposed to discriminate known circulating strains. Reconstruction of a phylogenetic tree using the 136 SNPs for all 161 samples resulted in a tree with the same number of delineated clades. Based on geographical location analysis, the composition of Lineage 2 in Kashgar prefecture (45.34%) is lower than other regions level in China(54.35%-90.27%), and the composition of Lineage 3 (32.30%)is much higher than other regions level in China (0.92%-2.01%), but it is lower than the bordering Pakistan (70.40%).

Conclusion

In summary, *M.tb* clinical strains from Kashgar prefecture were identifed 3 lineages and 11 sublineages, with 136 Branch-Specific SNP. Kashgar borders countries with a high incidence of tuberculosis such as Pakistan and India, resulting in a large difference between the *M.tb* lineage and sublineage distribution in this region and other provinces in China. This research provides a theoretical basis for the prevention and control of tuberculosis in Xinjiang.

Introduction

Tuberculosis patients in China accounted for 8.4% of total international cases (about 10 million).[1] Lineage plays an important role in disease prognosis, vaccine efficacy and drug resistance; therefore, it is crucial to determine and understand the *M.tb* lineage for the prevention and control of tuberculosis[2–3].Traditional *M.tb* genotyping methods (IS6100, Spoligotyping, MIRU-VNTR, etc.) was inadequate in tuberculosis prevention and control due to low resolutions and other factors. With the development of sequencing technology, whole genome sequencing (WGS) can identify sequence variations at the whole genome level with higher discrimination, and is more beneficial for understanding correlations among drug resistance, virulence and tuberculosis progression. Coll et al. [4] and Homolka et al.[5] found that single nucleotide polymorphism (SNP) can be used as a high resolution and stable typing technique, in addition to being used for phylogenetic and evolutionary analysis. Furthermore, 7 lineages and 55 sublineages of *M.tb* can be subdivided and denominated by 62 specific SNPs [4]. Many scholars had discussion the systematic genetic relationships between the *M.tb* lineages and sublineages on the basis of SNP. The human adaptive *M.tb* complex can be divided into 7 lineages (Lineage 1-Lineage 7), with each lineage indicating diversity in different regions, among which Lineage 2 and Lineage 4 were widely prevalent lineages in the world[6–7]. Stucki D et al. [8] found that Lineage 4 clinical strains (72 cases) were composed of 'generalists' (for example, Lineage 4.1.2, Lineage 4.3 and Lineage 4.10) and 'specialists' (for example, Lineage 4.1.3, Lineage 4.5, Lineage 4.6.1 and Lineage 4.6.2). Ajawatanawong P et al. [9] found that *M.tb* in Chiang Rai, Thailand (1170 cases) were composed of 4 lineages (Lineage 1-Lineage 4) and defined a new sublineage, specifically, Lineage 2.2.1.Ancestral 4. Zhang H T et al. [10] divided *M.tb* isolates from all over China (161 cases) into 3 lineages (Lineage 2-Lineage 4) and 5 sublineages (Lineage 2.1, Lineage 2.2, Lineage 2.3, Lineage 4.1 and Lineage 4.2). The results showed that the evolutionary transition from Ancestral to the Modern Beijing sublineages might be gradual. To summarize, the distribution of *M.tb* lineages and sublineages in different geographical regions is diverse. Currently, *M.tb* lineages are primarily by traditional typing methods in the Xinjiang region and the entire country, the results show the main epidemic isolate is Lineage 2[11–12], but there is insufficient research on *M.tb* lineage in terms of whole genome SNP.

Therefore, in this study, WGS of *M.tb* clinical strains in Kashgar prefecture (161 cases) was performed, analysis of *M.tb* lineage and sublineage composition on the basis of specific SNP was conducted and geographical distribution characteristics were examined to provide a theoretical basis for epidemiological investigation.

Materials And Methods

Samples

A total of 161 *M.tb* clinical strains were collected from 2018 to 2019 in Kashgar prefecture, Xinjiang, China. The clinical strains were collected from the First People's Hospital of Kashgar, Shufu County

People's Hospital, Shule County People's Hospital, Payzawat County People's Hospital, Yengisar County People's Hospital, Yarkant County People's Hospital and Poskam County People's Hospital. Sputum from patients' lower respiratory tracts was collected for the clinical strains, and general information of each patient was acquired and sorted (Additional file 1). All patients were clinically tested for etiology, drug sensitivity and IGRA. Informed consent from each patient was also obtained.

Whole Genome Sequencing (WGS)

M.tb genomic DNA was extracted and purified by magnetic bead extraction kit (MGIEasy, 1000006988) and the concentration of nucleic acid was quantified by Qubit 3.0 fluorescence quantitator (ThermoFisher, Q33216). The qualified clinical strains were treated with MGIEasy Digesting DNA Library Preparation Kit (MGIEasy, V2.0) for library construction, and library fragment size was checked by an Agilent 2100 Bioanalyzer (Agilent Technologies, G2939AA). After qualified libraries were mixed, WGS was conducted on the MGI 2000 Platform (MGI, PE100).

Sequencing Data Process and Mutation Detection

The quality of the raw reads was checked using FastQC toolkit (V0.11.8) followed by trimming of adapters, low-quality bases with a Phred quality score of less than 20 and fragments with large fluctuation in the beginning of each sequence. Reads shorter than 30 bp were excluded from the downstream analysis and the effective sequence length of reads was controlled at about 80bp. The Coverage depth of *M.tb* genome were analyzed by depth function of SAMtools (V1.10) [13]. Samples with a coverage of more than 95% were qualified for sequencing data. Then, reads were then mapped on to the reconstructed ancestral sequence of *M.tb*[14] using Burrows-Wheeler Alignment Tool (BWA) (V0.7.17) [14]. There is no reconstruction available for an ancestral *M.tb* chromosome and thus the chromosome coordinates and the annotation used is that of H37Rv (NC_000962.3). Duplicated reads were marked by the Mark Duplicates module of Picard (V1.119) and excluded, which generated by PCR that are not the genome itself. SNPs were called from each alignment file using GATK (V4.0) [15]. All SNPs were annotated with H37Rv by ANNOVAR (V2.1.1) [16]. The annotation embodied the amino acid changes at the SNP site, the position information of the antigen peptide, and the gene name and Rv number.

Phylogenetic Analysis and PCA

Based on SNPs f the whole-genome sequencing (WGS) of *M.tb* clinical strains (161 cases), an ML phylogenetic tree was constructed via IQ-tree (V1.6.12)[17] in a way of the ultra-fast bootstrap(bootstrap=1000) method. Then using KvarQ (V0.12.2)[18] determines the *M.tb* complex lineage/sublineage (Additional file 3) by analyzing the spoligotyping of the sample. The phylogenetic tree was drawn and beautified using FigTree (V1.4.4). In terms of all SNP, PCA was conducted for *M.tb* clinical strains using Plink 2.0[19] and adegenet package of R(V4.0.5) in order to verify accuracy of the lineages and sublineages.

Branch-Specific SNPs and Classification of Lineages and Sublineages

For each lineage and sublineage, the dataset was split into two populations: one containing all samples descending the clade-defining node and the other with remaining samples. The different SNPs are obtained by comparing the two branches. To ensure that branch-specific SNPs can also be used as markers for strain typing, we adopt the following filtering criteria: (1) Only synonymous mutations are retained due to reduce selective under external pressure. (2) SNPs in the coding region are retained, because which is lower frequency of insertions and deletions in the coding region. (3) The basic genes related to the growth of *M.tb* were used. (4) When comparing two branches difference of SNP sites, we select the site with F-statistics (Fst)> 0.99, Fst can be calculated by hierfstat of R package (value range 0~1, 0 means that the two populations are random Mating, 1 means that the two populations are completely isolated). (5)Refer to the classification basis proposed by Coll et al. [4] and Shitikov et al. [20] for the classification of Lineage 2, Lineage 3, Lineage 4 and its sublineages, The Branch-Specific SNPs is selected for *M.tb* lineages and sublineages in this area(see Additional file 3 for details based on the Branch-Specific SNPs of *M.tb* lineage and sublineage proposed by many scholars [4,13,20–22].)

Geographical Distribution of Lineages and Sublineages

M.tb lineage and sublineage information were correlated with the geographical information of the 161 tuberculosis cases. The composition and differences between *M.tb* lineages and sublineages in Kashgar (including one city and six counties) were analyzed. To compared with those in: Xinjiang's neighboring provinces (Tibet, Gansu and Qinghai), and the other regions in China and neighboring countries to Xinjiang (Pakistan, India, etc.). The differences in *M.tb* lineage composition between Kashgar prefecture (including one city and six counties) and the above regions have previously been discussed.

Results

Lineage and Sublineage Analysis of 161 M.tb Clinical Strains

Based on 21,438 SNPs, 161 *M.tb* clinical strains in Kashgar prefecture were clearly divided into three main branches using the phylogenetic tree constructed via ML method (Fig. 1A). One branch samples were clustered with corresponding lineage reference strains, namely Lineage 2 (73/161, 45.34%), Lineage 3 (52/161, 32.30%) and Lineage 4 (36/161, 22.36%). The 3 lineages were further divided into 11 sublineages according to branch-specific SNPs (Fig. 1A). 73 *M.tb* of lineage 2 were entirely from Lineage 2.2, of which 64 *M.tb* were Lineage 2.2.1 (87.67%), and other 9 *M.tb* were Lineage 2.2.2/Asia Ancestral 1 (12.33%). Lineage 2.2.1 was further divided into 3 sublineages, corresponding to Asia Ancestral 2 (9/73, 12.33%), Asia Ancestral 3 (16/73, 21.92%) and the Modern Beijing sublineage (39/73, 53.42%), respectively. 52 *M.tb* of Lineage 3 were divided into two main branches in the phylogenetic tree without Lineage 3.1 sublineage. Therefore, Lineage 3.2 (14/52, 26.92%) and Lineage 3.3 (38/52, 73.08%) were named anticlockwise in phylogenetic tree, respectively. 36 *M.tb* of Lineage 4 were divided into 5 sublineages, corresponding to Lineage 4.1 (3/36, 8.33%), Lineage 4.2 (2/36, 5.66%), Lineage 4.4.2 (1/36, 5.66%).2.78%), Lineage 4.5 (28/36, 77.78%) and Lineage 4.8 (2/36, 5.66%), respectively. Among the 9 sublineages mentioned, Lineage 2.2.1-Modern Beijing sublineages (39/161, 24.22%), Lineage 3.3

(38/161, 23.60%) and Lineage 4.5 (28/161, 17.39%) had the highest proportion in each lineage, which were the main epidemic strains in Kashgar prefecture (Fig. 1B). PCA was consistent with the above results (Fig. 1C). The results showed that the samples could be divided into three main lineages, namely L2, L3 and L4. Among them, PC1 and PC2 are the most important and cumulatively explain 52.64% (Fig. 1C). PCA of the three main lineages was further performed and the sublineages could be clearly divided (Fig. 1D-1F).

Specific SNPs of 161 M.tb Clinical Strains

136 branch-specific SNPs were obtained by screening. Reconstruction of a phylogenetic tree using the 136 SNPs for all 161 samples resulted in a tree with the same number of delineated clades (Additional file 2).The Branch-Specific SNPs of each lineage and sublineage are shown (Table 1, Additional file 3). There were 14 Branch-Specific SNPs in Lineage 2, among which 9 SNPs had not been reported before. In addition, 12 Branch-Specific SNPs is screened in 4 sublineages (Lineage 2.2.2/Asia Ancestral 1, Lineage 2.2.1-Asia Ancestral 2, Lineage 2.2.1-Asia Ancestral 3 and Lineage2.2.1-Modern Beijing sublineagewere 6, 4, 1 and 1, respectively), and these Branch-Specific SNPs of *M.tb* sublineages were found for the first time; There were 14 Branch-Specific SNPs in Lineage 3, among which 9 SNPs were different from the reported SNPs. In addition, There was also three Branch-Specific SNPs in sublineages 3.2 and sublineages 3.3; There were 10 Branch-Specific SNPs in Lineage 4, among which 7 SNPs had not been reported before. In addition, 72 Branch-Specific SNPs is screened in 5 sublineages (Lineage 4.1, Lineage 4.2, Lineage 4.4.2, Lineage 4.5 and Lineage 4.8 were 18, 18, 20, 13 and 3), and the Branch-Specific SNPs of *M.tb* strain were found for the first time. It also has the same sublineage Branch-Specific SNPs as the Coll's report. To summarize, there were Branch-Specific SNPs for 161 *M.tb* clinical strains in Kashgar prefecture (including one city and six counties).

Geographical Distribution of Lineages/Sublineages

M.tb lineage/sublineage information were correlated with the geographical information of tuberculosis patients (Fig. 2A). *M.tb* clinical strains of Kashgar prefecture (including one city and six counties) were composed of 3 lineages (Lineage 2, Lineage 3 and Lineage 4). The proportion of *M.tb* lineages in the six Kashgar prefecture counties were similar. Compared with neighboring provinces such as Tibet, Qinghai and Gansu (Lineage 2: 85.00%-92.12%, Lineage 3: 0.20%-5.00%, and Lineage 4: 4.98%-10.00%), Kashgar prefecture (including one city and six counties) had a lower proportion of Lineage 2 (45.34%) and a higher proportion of Lineage 3 (32.30%) and Lineage 4 (22.36%)(Fig. 2B). In Kashgar prefecture (including one city and six counties), the proportion of Lineage 2 (45.34%) was lower than the national average (70.00%), the proportion of Lineage 3 was higher than the national average (0.92%-2.01%), and the proportion of Lineage 4 was similar to the national average (24.82%-25.25%)(Fig. 2C). Compared with neighboring countries, the proportions of main *M.tb* epidemic lineages in Pakistan and India were Lineage 3 (70.40%) and Lineage 1 (70.15%), respectively. The proportion of Lineage 3 (32.30%) was relatively low and there was no Lineage 1 in Kashgar prefecture (including one city and six counties) (Fig. 2C). *M.tb* lineage was more complex than that of in other regions of China but similar to that of in

neighboring countries in Kashgar prefecture (including one city and six counties). Thus, it is speculated that the Lineage 3 strains may be introduced from neighboring countries in Kashgar prefecture (including one city and six counties).

The Lineage2.2.1-Modern Beijing sublineage of Lineage 2 was distributed in all counties except Shule county, which accounted for the highest proportion (33.33%-86.67%) in the Lineage2 of Kashgar prefecture (including one city and six counties) (Fig. 2D). The Lineage2.2.1-Modern Beijing sublineage (57.45%) was also predominant in other regions of China (Fig. 2E). Except for Poskam County, the proportion of Lineage 3.3 (42.86%-93.75%) was higher than that of Lineage 3.2 (6.25%-57.14%) in all other cities and counties in this region. Moreover, Lineage 3.1 sublineage was predominant in other regions of China(Fig. 2D). As for Lineage 4, the Lineage 4.5 sublineage was the dominant sublineage, which had the highest proportion (33.33%-100%) in each city and county in Kashgar prefecture (including one city and six counties) (Fig. 2D).Additionally,3 Lineage 4.1 strains were found in Payzawat, Yengisar and Shache County, respectively. 2 Lineage 4.2 strains were found in Kashgar city. 1 Lineage 4.4.2 strain was found in Shache County. 2 Lineage 4.8 strains were found in Poskam County, and the remaining 28 *M.tb* were all Lineage 4.5 strains. Lineage 4 was mainly composed of Lineage 4.2 (7.78%), Lineage 4.4 (31.11%) and Lineage 4.5 (60.00%) in other regions in China (Fig. 2F).In a nutshell, the Lineage 2.2.1-Modern Beijing sublineage and Lineage 4.5 were the primary dominant *M.tb* strains and other regions in China, and the sublineage composition of Lineage 3 was different other regions in China in Kashgar prefecture (including one city and six counties).

Discussion

Kashgar prefecture is located in Northwest China, and borders with many countries in Central Asia. It is a crucial transportation hub for cultural exchange, tourism, and economic trade between China and other countries in Central Asia. The annual incidence of tuberculosis in China is about 1.3 million, Ranking second in the world [23]. Since Kashgar prefecture borders Pakistan, India and other countries with high tuberculosis rates, it may have an impact on the local *M.tb* distribution. And there are vast differences in the pathogenicity of *M.tb* across different lineages [23, 24]. Understanding the distribution of *M.tb* lineages is beneficial for the prevention and control of tuberculosis.

In this study, 161 cases of *M.tb* clinical strains were composed of Lineage 2, Lineage 3 and Lineage 4, which is consistent with Chen H on the distribution of *M.tb* isolates in the Xinjiang region[25]. This study confirmed that Lineage 2 is dominant in all provinces in China (the national average is 70%, and the proportion in Xinjiang was only 44%) [26, 27], and the composition of Lineage 2(73/161, 45.34%)in Kashgar prefecture (including one city and six counties) is lower than the national average and consistent with the above results. As per the study, Lineage 3 was spread through the overland Silk Road [27] and was concentrated in northwestern China. 62% of Lineage 3 (CAS/Delhi) strains in China were found in Xinjiang [28].Lineage 3 strains were also found in provinces and cities adjacent to Xinjiang (Tibet and Qinghai Provinces), and the composition of Lineage 3 is higher than other prefecture[26]. In addition,

Lineage 3 (70.40%) was dominant in Pakistan, the country with a high tuberculosis rate bordering Kashgar prefecture [29]. In this study, Lineage 3 in Kashgar prefecture (including one city and six counties) accounted for 32.30%, which was much higher than surrounding provinces (0.20%-5.00%) and other provinces in China (0.92%-2.01%), but was similar to neighboring countries.It is speculated that the Kashgar prefecture borders Pakistan, and its border crossing (Khunjerab Port) is also located here, resulting in frequent movement of people between the two places, resulting in the spread of Lineage 3 strains and further spread to other provinces. Lineage 4 was highly prevalent in western China [25, 30], and mainly consisted of the Lineage 4.5 sublineage (primarily in Xinjiang for the geographical restriction), but absent in the Americas and Africa [8]. In this study, in regards to Lineage 4(36/161, 22.36%), the Lineage 4.5 sublineage(28/36, 77.78%)was also prevalent in Kashgar prefecture (including one city and six counties), with a higher proportion than that of other regions in China, which is consistent with the above study. Xinjiang is located in the middle of Eurasia and is a crucial transportation hub for the Silk Road Economic Belt. The Lineage 4.5 may have been spread to the region through the ancient Silk Road, which led to the prevalence of this sublineage in the area.

WGS can rapidly provide genotypes and drug-resistant types for epidemiological surveillance. Coll [4] proposed the classification markers for 7 lineages of *M.tb* and their sublineages based on analysis of 1,601 *M.tb* genomic data worldwide. *M.tb* clinical strains could be accurately classified on the basis of these markers. Prasit P [31] classified 480 cases of Lineage 1 clinical strains in Thailand into 18 sublineages based on the above-mentioned markers. In this study, 3 Lineages of *M.tb* clinical strains were further divided into 11 sublineages based on the reported specific SNP [4,20–22]. After screening, a total of 136 Branch-Specific SNPs were obtained, among which 89 SNPs(89) were different from Coll et al. reported. Furthermore, the newly identified Lineage 3.2 and Lineage 3.3 may be specific sublineages in Kashgar prefecture (including one city and six counties), and their specific classifications will be further investigated in future studies. The above results enriches *M.tb* lineage and sublineage marker SNP, and provides a theoretical basis for the prevention of tuberculosis in this region and a deeper understanding of *M.tb*-specific SNP in Kashgar prefecture (including one city and six counties).

As per the results of this study, it can be seen that most patients suffering from tuberculosis in China are caused by *M.tb* Lineage 2 and Lineage 4 [26], which are both more pathogenic than other lineage isolates [24], and among them, there may be possible correlations between Lineage 4 and non-Han populations [32], and between Lineage 3 and the Uygur nationality [25]. Kashgar prefecture as a crucial transportation hub for the ancient Silk Road, and is an international trading port where Chinese and foreign merchants gather. In addition, it borders Pakistan, India and other countries frequently encountering tuberculosis, and experiences frequent movement from people from neighboring countries, which may be one of the key reasons for the different proportion of *M.tb* lineage between Kashgar prefecture (including one city and six counties) and other regions in China.

In conclude, 161 cases of *M.tb* clinical strains from Kashgar prefecture (including one city and six counties) were divided into 3 lineages and 11 sublineages, with region-specific SNP. In consideration of

the geographical distribution of *M.tb*, it was found that the composition of *M.tb* lineage in Kashgar prefecture was more complex than other regions in China, and the proportion of *M.tb* lineage in Kashgar City was different from the other six counties in this region. Lineage 3 was the main prevalent strains in Pakistan, but it was only prevalent in the Xinjiang region in China, which may explain why this lineage strains is thought to have been spread from neighboring countries. Through this study, a fundamental basis is provided for the study of *M.tb* lineages and the prevention and control of tuberculosis in Kashgar prefecture.

Declarations

Ethics approval

All experiments in this present study were approved by the ethics committee of the First People's Hospital of Kashgar (approval number:*2019. No. (55)*, *2020. No. (57), 2020. No. (58)*).We confirmed that all methods were performed in accordance with the Declaration of Helsinki and relevant regulations. We declared that all participants have signed the Informed Consent Form for this research

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files]. The datasets generated and analysed during the current study are available in the [28860] repository, [PERSISTENT WEB link TO datasets].All the raw sequencing data and sample related information have been uploaded to NGDC database,[The assigned accession of the submission is: CRA005180] [https://ngdc.cncb.ac.cn/].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by The Special project for the construction of the autonomous region's innovative environment (talents, bases): Tianshan Cedar Project (*grant numbers. 2019XS22*), Tianshan Innovation Team (*grant numbers. 2021D14003)*. and the Natural Science Foundation of Xinjiang Uygur Autonomous Region (*grant numbers. 2020D01C013*).

Authors' contributions

Xiao-Guang Zou and Ai-Min Xu conceptualized the study, designed the study protocol. Ai-Min Xu, Chuan-Jiang He and Xiang Chen wrote the manuscript. Li Li, AniKiz Abuduaini and Zureguli Tuerxun responsible

for collecting clinical data. Aihemaitijiang Kaisaier, Hong-Mei Peng and Ya-Hui Zhen were contributed to whole genome sequencing. Yin-Zhong Sha, Su-Jie Zhang and Jing-Ran Xu were contributed in data analysis and created figures. The authors thank the anonymous reviewers and all of the Editors for their helpful comments and suggestions on this manuscript.

Acknowledgements

Thanks to Lu Liu, Wenbo Zhao and Chaoyang Chen of Xinjiang Dingju Biotechnology Co., Ltd. for their technical support in whole-genome sequencing and data analysis.

References

1. World Health Organization(WHO). Global tuberculosis report, 2020[R]; WHO.
2. Lo´pez B, Aguilar D, Orozco H, et al., A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes[J]. Clin Exp Immunol 2003; 133:30–37. doi: 10.1046/j.1365-2249.2003.02171.x.
3. Ford CB, Shah RR, Maeda MK, et al., *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis[J]. Nat Genet 2013; 45:784–790. doi: 10.1038/ng.2656.
4. Coll F, McNerney R, Guerra-Assuncao JA, et al., A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains[J]. Nat Commun 2014; 5:4812–4816. doi: 10.1038/ncomms5812.
5. Homolka S, Projahn M, Feuerriegel S, **et al.**, **High Resolution Discrimination of Clinical** *Mycobacterium tuberculosis* **Complex Strains Based on Single Nucleotide Polymorphisms[J]**. PLoS one 2012; **7**:**e39855**. **doi**: 10.1371/journal.pone.0039855.
6. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*[J]. Nat Rev Microbiol 2018; 16:202–213. doi: 10.1038/nrmicro.2018.8.
7. Comas I, Coscolla M, Luo T, et al., Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans[J]. Nat Genet 2013; 45:1176–1182. doi: 10.1038/ng.2744.
8. Stucki D, Brites D, Leïla J, et al., *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages[J]. Nat Genet 2016; 48:1535–1543. doi: 10.1038/ng.3704.
9. Ajawatanawong P, Yanai H, Smittipat N, et al., A novel Ancestral Beijing sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sublineages[J]. Entific Reports 2019; 9:1–12. doi: 10.1038/s41598-019-50078-3.
10. Zhang HT, Li DF, Zhao LL, et al., Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance[J]. Nat Genet 2013; 45:1255–1260. doi: 10.1038/ng.2735.
11. Li WM, Wang SM, Li CY, et al., Molecular epidemiology of *Mycobacterium tuberculosis* in China: A nationwide random survey in 2000[J]. Int J Tuberc Lung Dis 2005; 9:1314–1319.

12. Zhao XQ, Dong HY, Liu ZG, et al., Preliminary analysis of the distribution of Beijing genotype strains of *Mycobacterium tuberculosis* in parts of China[J]. Practical Preventive Medicine 2012; 19:662–664.
13. Heng Li. Astatistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data[J]. Bioinformatics, 2011, 27(21):2987–93. doi: 10.1093/bioinformatics/btr509.
14. Heng, Li, Richard, et al. Fast and accurate short read alignment with Burrows-Wheeler transform.[J]. Bioinformatics, 2009. doi: 10.1093/bioinformatics/btp324.
15. Mckenna A,Hanna M ,Banks E,et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data[J]. Genome Research, 2010, 20(9):1297–1303. doi: 10.1101/gr.107524.110.
16. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data[J]. Nucleic Acids Research, 2010, 38:e164. doi: 10.1093/nar/gkq603.
17. Lam-Tung N, Schmidt H A, Arndt V H, et al. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies[J]. Molecular Biology & Evolution, 2015(1):268–274. doi: 10.1093/molbev/msu300.
18. Steiner, Andreas,Stucki, et al. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes[J]. Bmc Genomics, 2014, 15. doi: 10.1186/1471-2164-15-881.
19. Chang C C, Chow C C, Cam T L,et al. Second-generation PLINK: rising to the challenge of larger and richer datasets[J]. GigaScience,4,1(2015-02-25)(1):7. doi: 10.1186/s13742-015-0047-8.
20. Shitikov E, Kolchenko S, Mokrousov I, et al., Evolutionary pathway analysis and unified classification of East Asian lineage of *Mycobacterium tuberculosis*[J]. Scientific Reports 2017; 7:9227–9236. doi: 10.1038/s41598-017-10018-5.
21. Merker M, Niemann S, Mona S, et al., Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage[J]. Nature Genetics 2015; 47:242–24*9*. doi: 10.1038/ng.3195.
22. López B, Aguilar D, Orozco H, et al., A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes[J]. Clinical & Experimental Immunology 2003; 133:30–37. doi: 10.1046/j.1365-2249.2003.02171.x.
23. Chen S M. Overview of Tuberculosis Research in China[J]. Chinese Journal of Antituberculosis 2011; 23:525–526.
24. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*[J]. Semin Immunol 2014; 26:431–444. doi: 10.1016/j.smim.2014.09.012.
25. Chen HX, He L, Huang HR, et al., *Mycobacterium tuberculosis* Lineage Distribution in Xinjiang and Gansu Provinces, China[J]. Sci Rep 2017; 7:1068. doi: 10.1038/s41598-017-00720-9.
26. Liu QY, Ma AJ, Wei LH, et al., China's tuberculosis epidemic stems from historical expansion of four strains of *Mycobacterium tuberculosis*[J]. Nature ecology & evolution 2018; 2:1982–1992. doi: 10.1038/s41559-018-0680-6.

27. Zhou Y, van den Hof S, Wang SF, et al., Association between genotype and drug resistance profiles of *Mycobacterium tuberculosis* strains circulating in China in a national drug resistance survey[J]. PLoS one 2017; 12:e0174197. doi: 10.1371/journal.pone.0174197.
28. O'Neill M B, Shockey A, Zarley A, et al., Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia[J]. Molecular Ecology 2019, 28:3241–3256. doi: 10.1111/mec.15120.
29. Abdul J, Jody EP, Paola Florez de Sessions, et al., Whole genome sequencing of drug resistant *Mycobacterium tuberculosis* isolates from a high burden tuberculosis region of North West Pakistan[J]. Sci Rep 2019; 9:14996. doi: 10.1038/s41598-019-51562-6.
30. Li Y, Fu Y, Yuan M, et al., Study on the population-genetics of *Mycobacterium tuberculosis* from Sichuan Basin in China[J]. Chinese Journal of Epidemiology 2015; 36:374–378.
31. Prasit P, Pravech A, Wasna V, et al., Evidence for Host-Bacterial Co-evolution via Genome Sequence Analysis of 480 Thai *Mycobacterium tuberculosis* Lineage 1 Isolates[J]. Sci Rep 2018; 8:11597–11610. doi: 10.1038/s41598-018-29986-3.
32. Guan YB. National Geographic Distribution and characteristics of China[J]. National Forum 1996; 3:19–23.
33. Manson AL, Thomas A, Galagan JE, et al., *Mycobacterium tuberculosis* Whole Genome Sequences From Southern India Suggest Novel Resistance Mechanisms and the Need for Region-Specific Diagnostics[J]. Clinical Infectious Diseases 2017:1494–1501. doi: 10.1093/cid/cix169.
34. Brites D, Gagneux S. The Nature and Evolution of Genomic Diversity in the *Mycobacterium tuberculosis* Complex[J]. Adv Exp Med Biol 2017:1–26. doi: 10.1007/978-3-319-64371-7_1.

Tables

Table. 1 Specific SNP of 3 Lineages and 11 Sublineages in this Study

Note: Number of specific SNP referred to SNP with *M.tb* mostly appearing in 3 lineages and 11 sublineages. Common SNP indicated the same specific SNP reported in this study as Coll et al. [4]

Lineage	Sublineage			Name	N	Number of specific SNP	Number of specific SNPs in Coll's study	Common SNP
2				East Asian	73	14	6	5(5/14, 35.71%)
	2.1			Protobeijing	-	-	12	-
	2.2			Beijing 2.2	-	-	5	-
		2.2.1		Beijing 2.2.1	64	8	2	2(2/8, 25%)
			*	Asia Ancestral 2	9	4	-	-
			*	Asia Ancestral 3	16	1	-	-
			*	Modern Beijing	39	1	-	-
		2.2.2		Asia Ancestral 1	9	6	6	5(5/6, 83.33%)
3				India and East Africa	52	14	9	5(5/14, 35.71%)
	3.1				-	-	-	-
		3.1.1			-	-	3	-
		3.1.2			-	-	2	-
	3.2				14	3	-	-
	3.3				38	3	-	-
4				Euro-American	36	10	3	3(3/10, 30.00%)
	4.1				3	18	1	1(1/18, 5.56%)
	4.2				2	18	8	8(8/18, 44.44%)
	4.4				-	-	2	-
		4.4.2			1	20	11	11(11/20, 55.00%)
	4.5				28	13	8	6(6/13, 46.15%)
	4.8				2	3	1	1(1/3,

33.33%)

Figures

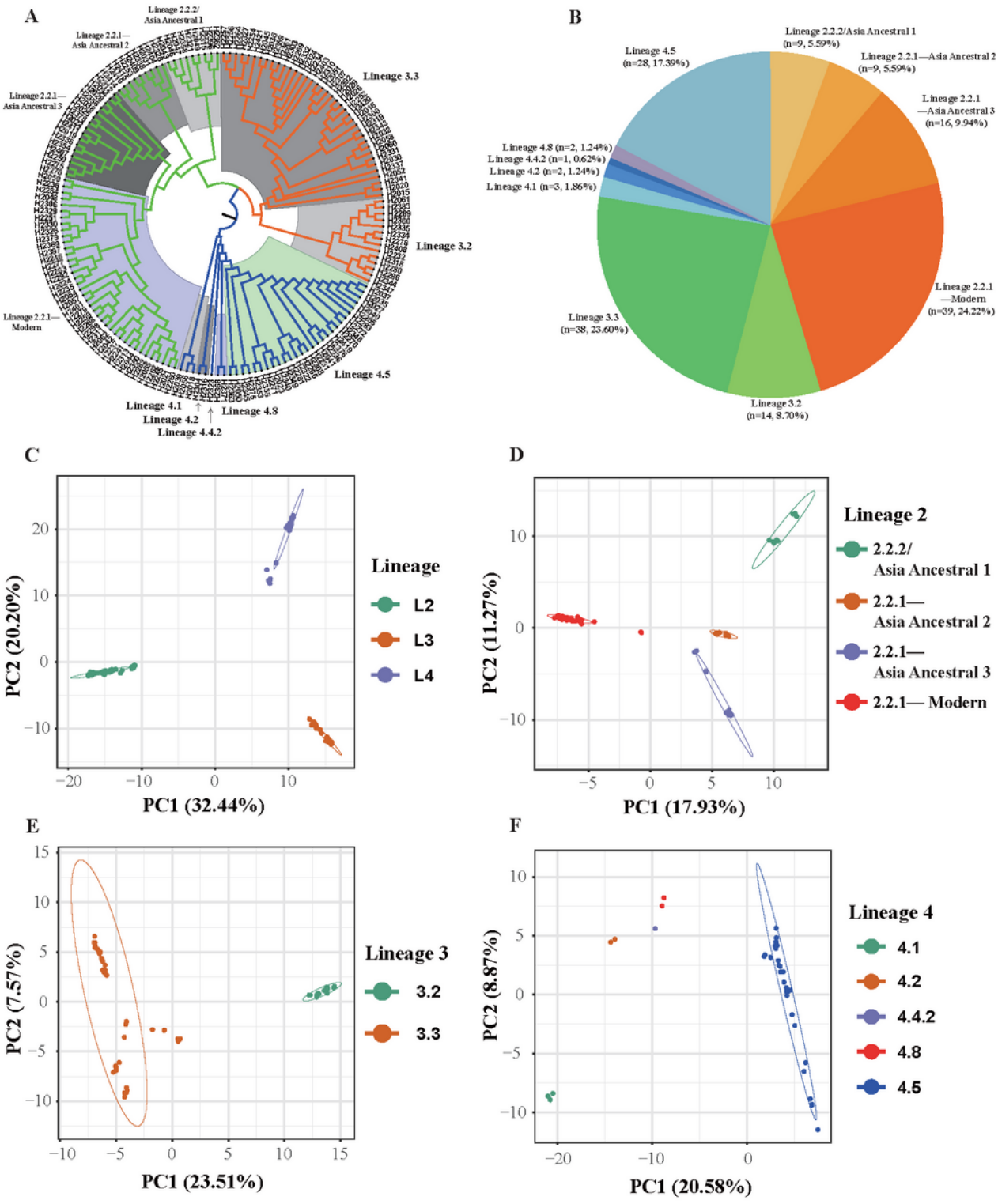

Figure 1

Lineage and Sublineage Analysis of 161 Cases of M.tb Clinical strains Note: (A) The phylogenetic tree of 161 cases of M.tb clinical strains in Kashgar prefecture (including one city and six counties) was constructed by ML method. 161 cases of M.tb clinical strains were labeled as per the classification basis proposed by Coll et al. [4] and Shitikov et al. [10]. A total of 3 lineages (Lineage 2-Lineage 4, marked with orange, green and blue, respectively) and 11 sublineages (marked by the shaded areas) were defined, and among them, the Lineage 3 sublineage was designated in an anticlockwise order (as Lineage 3.2 and Lineage 3.3, respectively). and the N 161 cases of M.tb clinical strains are marked with black font. Excluding Lineage 2.2.2 sublineages, all clinical strains in M.tb Lineage 2 were of Lineage 2.2.1 sublineages. (B) The proportion of 9 sublineages of M.tb clinical strains. (C) PCA Diagram of M.tb Clinical strains of the three lineages. (D) PCA Diagram of M.tb Clinical strains in Lineage 2. (E) PCA Diagram of M.tb Clinical strains in Lineage 3. (F) PCA Diagram of M.tb Clinical strains in Lineage 4.

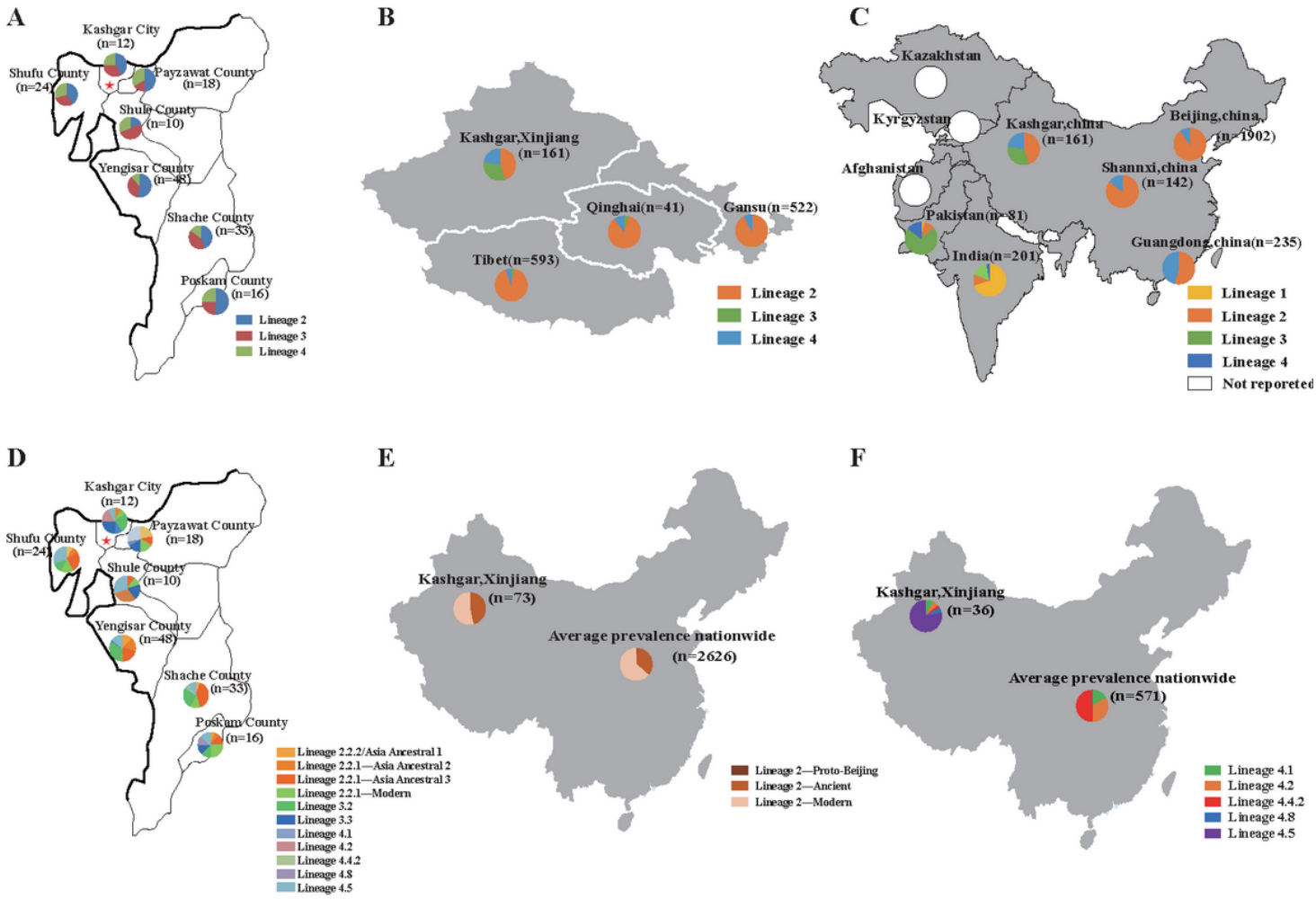

Figure 2

Comparison and Geographical Distribution of M.tb Lineages and sublineages Note: (A) Lineage Composition and Distribution of 161 Cases of M.tb Lineage 2-Lineage 4 in Kashgar prefecture (including One City and Six Counties); (B) M.tb Lineage Composition of Kashgar prefecture (including One City and Six Counties) and Adjacent Provinces (Tibet, Qinghai and Gansu Provinces) (C) M.tb Lineage

Composition of Kashgar prefecture (including One City and Six Counties), certain Domestic Provinces and Cities (Shanxi Province, Beijing, and Guangdong Province), and Neighboring Countries (Kazakhstan, Kyrgyzstan, Tajikistan, Afghanistan, Pakistan and India). There was no report on M.tb lineages in Kazakhstan, Kyrgyzstan, Tajikistan and Afghanistan, which are respectively marked with white circles in Fig. 2C. (D) Sublineage Composition and Distribution of 161 Cases of M.tb Lineage 2-Lineage 4 in Kashgar prefecture (including One City and Six Counties); (E) Proportion of M.tb Lineage 2 sublineages in Kashgar prefecture (including One City and Six Counties) and the Whole Country; (F) Proportion of M.tb Lineage 4 sublineages in Kashgar prefecture (including One City and Six Counties) and the Whole Country. All data (excluding data on Kashgar prefecture) were obtained from Liu Q Y et al. (2018)[26], Abdul J et al. (2019)[29], Manson A L et al. (2017)[33], Brites D et al. (2017)[34]. Kashgar City is marked with a red five-pointed star in the figure, which indicates the location of the administrative office of Kashgar prefecture.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.Thedetailofthe161Samplesinformation.xlsx
- Additionalfile2.Phylogenetictreebasedon136specificSNPs.pdf
- Additionalfile3.136specificSNPsinM.tb.xlsx