








# Genomic and Spatial Analysis on the Recent Transmission of *Mycobacterium tuberculosis* in Eastern China: A 10-Year Retrospective Population-Based Study

Xiwen Yin <sup>1,\*</sup>, Qiang Zhang <sup>1,\*</sup>, Yuting Wang <sup>1</sup>, Bilin Tao <sup>1</sup>, Xiaolong Zhang <sup>2</sup>, Jinyan Shi <sup>3</sup>, Xiaowei Deng <sup>1</sup>, Jianming Wang <sup>1,4,5</sup>

<sup>1</sup>Department of Epidemiology, Key Laboratory of Public Health Safety and Emergency Prevention and Control Technology of Higher Education Institutions in Jiangsu Province, Center for Global Health, School of Public Health, Nanjing, 211166, People's Republic of China; <sup>2</sup>Department of Tuberculosis Control, Center for Disease Control and Prevention, Suzhou, 215000, People's Republic of China; <sup>3</sup>Department of Clinical Laboratory, The Fourth People's Hospital of Lianyungang, Lianyungang, 222000, People's Republic of China; <sup>4</sup>Department of Tuberculosis, The Third People's Hospital of Changzhou, Changzhou, 215000, People's Republic of China; <sup>5</sup>National Vaccine Innovation Platform, Nanjing Medical University, Nanjing, 211166, People's Republic of China

\*These authors contributed equally to this work

Correspondence: Jianming Wang, Department of Epidemiology, Key Laboratory of Public Health Safety and Emergency Prevention and Control Technology of Higher Education Institutions in Jiangsu Province, Center for Global Health, School of Public Health, Nanjing, 211166, People's Republic of China, Tel +86-25-86868414, Email [jmwang@njmu.edu.cn](mailto:jmwang@njmu.edu.cn)

**Purpose:** Understanding the mode of *Mycobacterium tuberculosis* (*M. tuberculosis*) transmission is crucial for disease prevention and control. Compared to traditional genotyping methods, whole genome sequencing (WGS) provides higher resolution and comprehensive genetic information, enabling the tracing of infection sources and determining of transmission routes to resolve extensive tuberculosis (TB) outbreaks. We conducted a ten-year study on the transmission of *M. tuberculosis* in a population in eastern China.

**Patients and Methods:** We selected Lianyungang, an eastern city in China, as the study site. Patients diagnosed with active pulmonary TB from 2011 to 2020 were enrolled as the study subjects. We isolated and sequenced 2252 *M. tuberculosis*. Strains with pairwise genetic distances of less than 12 single nucleotide polymorphisms were defined as genomic clusters and which were considered recent transmissions. Kernel density estimation and K-function analysis were applied to explore the spatial distribution of recently transmitted strains.

**Results:** After excluding non-tuberculous mycobacteria and duplicated samples, 2114 strains were included in the final analysis. These strains comprised lineage 2 (1593, 75.35%) and 4 (521, 24.65%). There were 672 clustered strains, with a recent transmission rate of 31.79%. The logistic regression model showed that the risk of recent transmission was high in students [adjusted odds ratio (aOR): 2.68, 95% confidence interval (CI): 1.63–4.49,  $P < 0.001$ ] and people infected with L2.2.1 strains (aOR: 1.59, 95% CI: 1.20–2.12). Higher spatial aggregation of TB transmission has been concentrated in Haizhou, Donghai, and Guanyun for the past 10 years. Three outbreaks affecting 46 patients were spatially spaced, with 11 to 23 persons each. Different groups exhibited varying geographic distances between the initial and later cases.

**Conclusion:** There are areas with a high risk of transmission for *M. tuberculosis* in the research site, and the risk varies among different populations. Accurate prevention strategies targeted at specific regions and key populations can help curb the prevalence of TB.

**Keywords:** tuberculosis, transmission, cluster, whole genome sequencing, spatial analysis

## Introduction

Tuberculosis (TB) is a chronic respiratory infectious disease caused by *M. tuberculosis* and is one of the major infectious diseases that jeopardize human health.<sup>1</sup> China has the third highest burden of TB in the world (7.1%) after India and Indonesia.<sup>2</sup> In 2022, the estimated incidence of TB in China was 52/100,000, with a mortality rate of 2.0/100,000.<sup>2</sup> The

transmission of TB is influenced by environmental factors, the characteristics of source cases and contacts, and the nature of their exposure, with the most significant risk coming from individuals with positive sputum tests and high cough frequency.

In recent years, the prevention and treatment of TB in China has had remarkable success, with the overall incidence rate showing a downward trend. This outcome is partially attributed to investigating the spread dynamics of diseases to understand the risk factors leading to disease occurrence in different populations. Molecular epidemiology and spatial analysis make it possible to study the spread of *M. tuberculosis*, allowing us to understand the spatial distribution of the disease, identify the most affected areas, and formulate transmission hypotheses based on this.

Lianyungang is a city with a high TB burden in northern Jiangsu Province, China. The declining trend in incidence has tended to be slow in the last several years, and targeted strategies are needed to achieve further declines. We have previously conducted studies on TB transmission in Lianyungang using the Variable Number Tandem Repeats (VNTR) or high-resolution genotyping methods but without geospatial correlation analysis.<sup>3</sup> Traditional genotyping methods may not be sufficient to discriminate between closely related strains resulting from recent transmission. Whole Genome Sequencing (WGS) is a comprehensive method for analyzing entire genomes, which can provide higher resolution and comprehensive genetic information, enabling the tracing of infection sources and determining transmission routes. By differentiating strains into smaller and more precise clusters, WGS provides higher resolution for identifying TB outbreaks for identifying TB outbreaks. Therefore, we combined the genomic data of *M. tuberculosis* strains based on the WGS method and spatial analysis to explain the characteristics of TB transmission in Lianyungang.

## Materials and Methods

### Study Design and Study Subjects

We selected Lianyungang as the study site. Lianyungang is located on the eastern coast of China, with a total area of 7615 square kilometers and a population of 4.594 million. Patients diagnosed with active pulmonary TB from 2011 to 2020 were enrolled as the study subjects. We collected their demographic and clinical data, including age, sex, resident area, diagnosis delay, treatment delay, laboratory test results, drug resistance and treatment outcomes. We defined the diagnostic delay as the period between the onset of a patient's symptoms and the confirmed diagnosis. This study was conducted in accordance with the Declaration of Helsinki and approved by the ethics committee of Nanjing Medical University. Written informed consent was obtained from study participants.

### *M. tuberculosis* Culture and Sequencing

We collected sputum samples from each patient and cultured them for *M. tuberculosis* strain isolation. Each specimen was treated with an equal volume of 4% sodium hydroxide and then vigorously stirred to achieve homogenization. Subsequently, the resulting specimen was inoculated into the medium and incubated at 37°C. Isolated strains were stored at -80°C.

Before gene sequencing, the strain was retrieved from the refrigerator and revived on the Lowenstein-Jensen medium at 37°C. The genomic DNA of the culture-positive strains was extracted using the cetyl trimethyl ammonium bromide (CTAB) method.<sup>4</sup> We constructed a 300-base-pair double-ended DNA library for each strain and sequenced it with an expected depth of 300×. The whole genome of *M. tuberculosis* was sequenced on the Illumina NovaSeq 6000 platform. We obtained clean data by using fastp (v0.23.2) to control raw data quality and Bowtie2 (v2.3.1) to map the sequencing reads to the reference genome H37Rv (NC\_000962.3).<sup>5</sup> The SnpEff (v5.1) was then used to obtain single nucleotide polymorphisms (SNPs) in the target genome relative to H37Rv. Snippy (v4.6.0) and snp-dists (v0.7.0) were applied to generate a matrix of SNP distance. Following this, FreeBayes (v1.3.6), a tool for detecting genetic variants, was employed to identify SNPs, insertions, deletions (indels), and multi-nucleotide polymorphisms (MNPs), ensuring a minimum Phred base quality of 30, a mapping quality exceeding 30 or more, and a sequencing depth  $\geq 5$ .<sup>6</sup>

We only used fixed SNPs (frequency  $\geq 75\%$ ) not located in drug-resistant genes or repetitive regions of the genome to calculate the pairwise distance. We defined isolates with pairwise genetic distances less than 12 SNPs as genomic clusters, considered recent transmissions.<sup>7</sup> Clusters of  $\geq 10$  people were defined as an outbreak (Group A, B and C). We included unclustered cases as a comparison group (Group D). We established the phylogenetic tree using the maximum

likelihood method by MEGA(v11.0) and created the comment file with itol.toolkit (<https://github.com/TongZhou2017/itol.toolkit>).<sup>8</sup> Subsequently, we used the Interactive Tree of Life (<https://itol.embl.de/>) to visualize the phylogenetic tree.

## Spatial Analysis of *M. tuberculosis*

We geocoded the residential addresses of participants by ArcGIS (v10.8). We conducted the spatial analysis with the genotype data to identify geographic transmission clusters, including geographic median center point, standard deviational ellipse with directional distribution at 1 standard deviation (SD) and kernel density estimation. The median center identifies the location that minimizes the overall Euclidean distance to the features in a dataset.<sup>9</sup> Data outliers less influence the algorithm for the median center. We used the standard deviation ellipse to provide a graphical representation of the direction for the spatial distribution of strains.<sup>10</sup> Point data were used to map TB cases because of the limited number of strains with SNPs  $\leq 12$  and the high degree of randomness of recently transmitted strains at the township level.

The kernel density estimation method transforms discrete categorical variables into numerical continuous variables through the kernel function. Spatial densities are reflected by calculating the probability density of events occurring in the study area, and smooth maps showing denser areas are generated. The search radius is set to 10,000 meters. Temporal changes in areas of high spatial aggregation were observed by plotting kernel densities of clustered patient residences in a sliding 3-year window. Finally, we plotted three large clusters of kernel density maps to analyze the spatial aggregation of each outbreak.

## Statistical Analysis

All statistical analyses were performed with R software (v4.2.2). We used medians (interquartile range [IQR]) to represent continuous variables and numbers (percentages) to represent categorical variables. The logistic regression model was used to calculate the odds ratio (OR) and 95% confidence interval (CI) for risk factors connected to genomic clustering. Variables with  $P < 0.05$  in the univariable analysis were included in the multivariable analysis. Multivariable analysis was used to calculate the adjusted odds ratios (aOR). We used the Chi-squared and Fisher's exact tests to compare differences between clustered and unclustered groups. The difference was considered statistically significant at a criterion of  $P < 0.001$ .

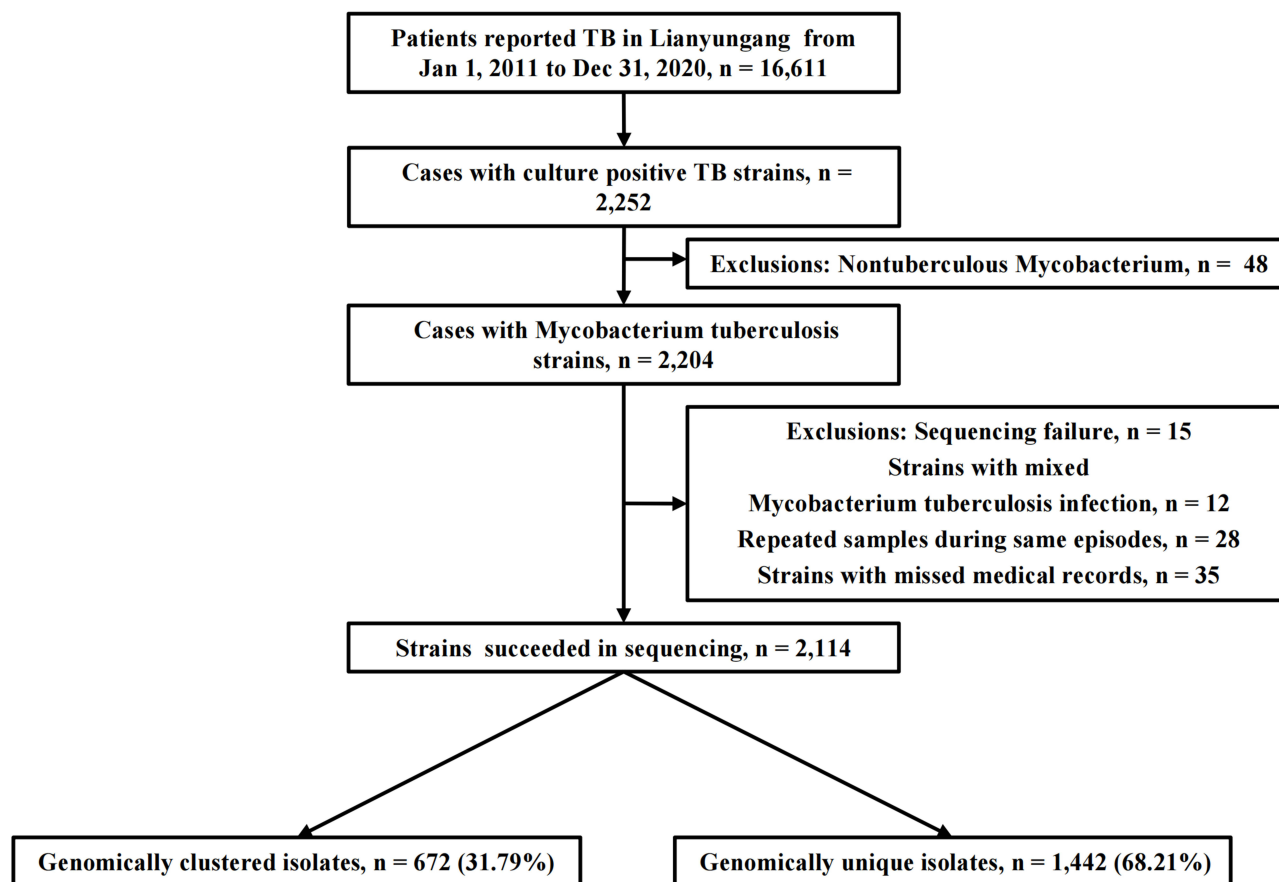
To estimate the spatial clustering among participants in each group, Ripley's K-function analysis, a representative spatial point pattern analysis approach, was employed to quantitatively evaluate the spatial dispersion characteristics of point patterns at a certain distance.<sup>11,12</sup> We compared relative clustering in individuals of clustered and unclustered groups by estimating the difference in K-functions over a range of distances to explain potential clustering caused by underlying transmission.<sup>13</sup> To aid our interpretation, we created plots showing K-function estimations along the y-axis and distances indicated along the x-axis. Next, we looked at the form and behavior of the observed K-function values. To get 95% CIs, we employed 999 random permutations. We examined lines where observations fell outside the upper or lower confidence interval to detect statistically significant differences.

We investigated possible spatiotemporal trends by measuring the geographic distance between the first participant diagnosed with TB and those subsequently diagnosed in each group. We applied the R package of spatstat (<https://cran.r-project.org/web/packages/spatstat/index.html>) for K-function analysis and ggplot2 to display scatter plots.<sup>14</sup>

## Results

From 2011 to 2020, 16,611 TB cases were reported in Lianyungang, and 2252 strains were culture-positive and sent for WGS. After excluding 48 non-tuberculous mycobacteria strains, 12 mixed infections, 28 duplications, 15 sequencing failures, and 35 strains with missing data, a total of 2114 *M. tuberculosis* strains were involved in the final analysis. Among them, there were 372 genomically clustered isolates, with the clustering rate at 31.79% (Figure 1). Strains were classified into 310 separate clusters, with each cluster varying in size from 2 to 23, indicating recent transmission of *M. tuberculosis*. There were 223 (71.9%) small clusters with 2 strains, 84 (27.0%) medium clusters with 3–9 strains, and 3 (0.97%) large clusters with more than 10 strains.

The incidence rate and the lineage proportion of TB patients in each county per year are shown in Figure S1. According to the phylogenetic tree constructed based on the WGS data, Lineage 2.2, also known as the Beijing genotype



**Figure 1** Flow chart of study subject enrollment.

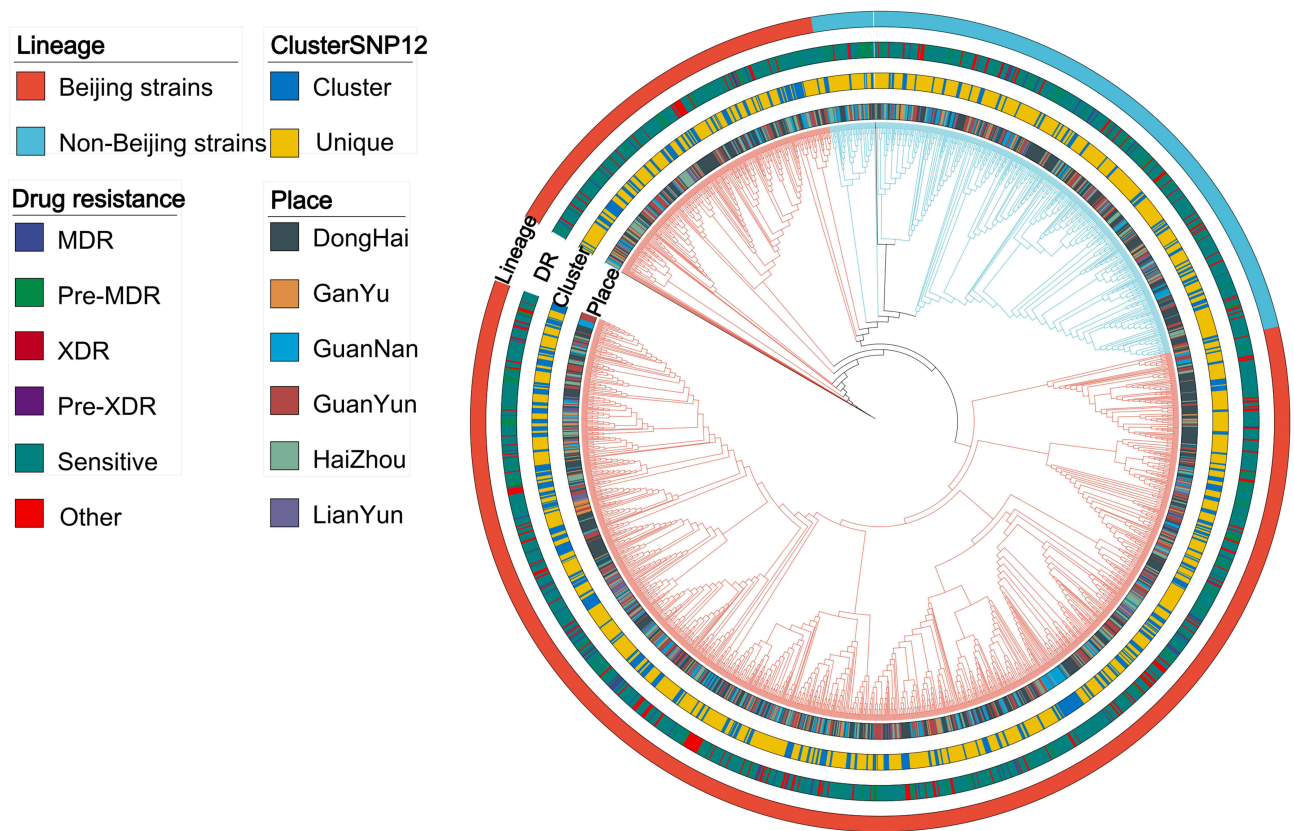
strain, accounted for the majority (75.2%, 1590/2114). L2.2.1 (1568 strains) was the predominant sub-lineage, comprising 64.1% of the total isolations. The remaining 523 strains, representing 24.7% of the total (523/2114), were classified as Lineage 4 strains, also known as Euro-American lineage (Figure 2).

The multivariable analysis revealed that students (aOR: 2.68, 95% CI: 1.63–4.49) have the highest clustering risk as compared to farmers. Using the sublineage L4.4.2 as the reference, L2.2.1 strains (aOR: 1.59, 95% CI: 1.20–2.12) were at a higher risk of clustering. In addition, the risk of clustering was lower in older patients as compared to patients under the age of 25 (25–44 years: aOR: 0.71, 95% CI: 0.52–0.97; 45–64 years: aOR: 0.51, 95% CI: 0.37–0.69;  $\geq 65$  years: aOR: 0.24, 95% CI: 0.17–0.34) (Table 1).

The areas of spatial aggregation were relatively stable during the 10-year study period. Haizhou, located in the center of Lianyungang, has always been the high spatial aggregation area for cluster-forming strains. High-density areas of cluster-forming strains gradually appeared in Donghai, and the high-density areas in Guanyun gradually moved toward Haizhou (Figure 3).

There were 46 cases involved in the clusters of A, B, and C, and 1442 cases were distributed in the control group D (unclustered). In group D, 85.23% of participants were primary TB patients, with a median age of 58 years, and 77.95% were male. Older adults (aged 65 and above) made up 35.71% of the group, with the majority being farmers (76.49%). Additionally, 71.29% of group D participants experienced delayed diagnosis, and 79.75% of the group were from rural areas.

Three clustered groups all belonged to the 2.2.1 lineage. The number of males exceeds females in each group, with median ages ranging from 20 to 37 years. Groups B and C were predominantly composed of younger individuals, with the proportion of students was 33.33% and 34.78%. Participants in groups A and B mainly come from urban areas, while



**Figure 2** Phylogenetic tree of *M. tuberculosis* isolates in Lianyungang.

those in group C mainly come from rural areas. Most of the three groups are newly diagnosed patients, with a higher incidence of delayed diagnosis (Table 2).

As shown in Figure 4, Kernel density estimations, median center points, and directional distributions for each clustered and unclustered group participant are presented on the map. For every group, the center point was located in

**Table 1** Univariate and Multivariable Logistic Regression of Risk Factors for Genomic Clustering

Variables	Unclustered N=1442	Clustered N=672	Total N=2114	Univariate Analysis		Multivariate Analysis	
				Odds Ratio (95% CI)	P-value	Odds Ratio (95% CI)	P-value
Gender							
Male	1124 (68.7)	513 (31.3)	1637 (77.4)	Ref.			
Female	318 (66.7)	159 (33.3)	477 (22.6)	1.10 (0.88, 1.36)	0.410		
Age group (years)							
≤24	158 (46.2)	184 (53.8)	342 (16.2)	Ref.			
25~44	294 (61.1)	187 (38.9)	481 (22.8)	0.55 (0.41, 0.72)	<0.001	0.71 (0.52, 0.97)	0.030
45~64	475 (70.3)	201 (29.7)	676 (32.0)	0.36 (0.28, 0.48)	<0.001	0.51 (0.37, 0.69)	<0.001
≥65	515 (83.7)	100 (16.3)	615 (29.1)	0.17 (0.12, 0.22)	<0.001	0.24 (0.17, 0.34)	<0.001
Occupation							
Farmer	1103 (71.3)	443 (28.7)	1546 (73.1)	Ref.			
Student	31 (30.7)	70 (69.3)	101 (4.8)	5.62 (3.67, 8.81)	<0.001	2.68 (1.63, 4.49)	<0.001
Service sector	136 (61.8)	84 (38.2)	220 (10.4)	1.54 (1.14, 2.06)	<0.001	1.11 (0.82, 1.51)	0.500
Laborer	74 (63.2)	43 (36.8)	117 (5.5)	1.45 (0.97, 2.13)	0.060	0.95 (0.63, 1.42)	0.810
Retire	77 (83.7)	15 (16.3)	92 (4.4)	0.49 (0.27, 0.83)	0.010	0.71 (0.83, 1.24)	0.250
Other	21 (55.3)	17 (44.7)	38 (1.8)	2.02 (1.04, 3.85)	0.030	1.65 (0.84, 3.22)	0.140

(Continued)

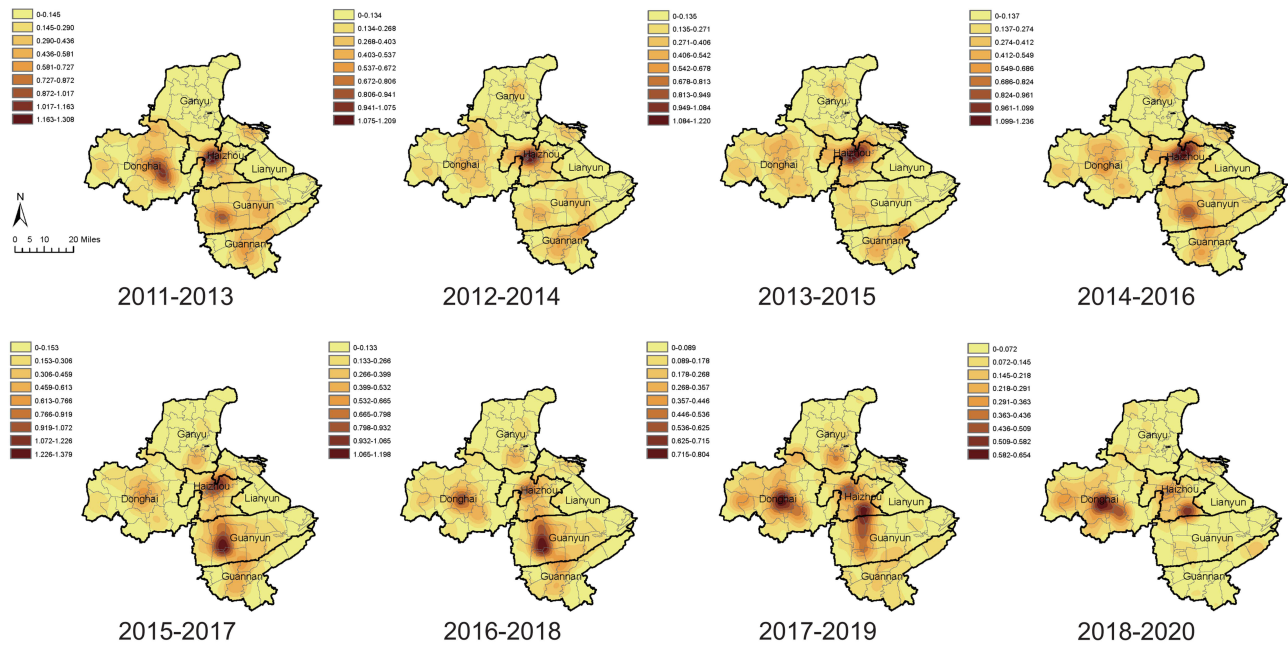
**Table I** (Continued).

Variables	Unclustered N=1442	Clustered N=672	Total N=2114	Univariate Analysis		Multivariate Analysis	
				Odds Ratio (95% CI)	P-value	Odds Ratio (95% CI)	P-value
Disease history							
New cases	1229 (68.2)	574 (31.8)	1803 (85.3)	Ref.			
Retreated cases	213 (68.5)	98 (31.5)	311 (14.7)	0.99 (0.76, 1.27)	0.910		
Diabetes							
No	1380 (68.2)	643 (31.8)	2023 (95.7)	Ref.			
Yes	62 (68.1)	29 (31.9)	91 (4.3)	1.00 (0.63, 1.56)	0.990		
Diagnostic delay							
No	414 (66.24)	211 (33.76)	625 (29.6)	Ref.			
Yes	1028 (69.0)	461 (31.0)	1489 (70.4)	0.88 (0.72, 1.07)	0.210	0.85 (0.69, 1.05)	0.120
Delay time							
No delay	414 (66.24)	211 (33.76)	625 (29.6)	Ref.			
≤2 weeks	294 (69.3)	130 (30.7)	424 (20.1)	0.87 (0.66, 1.13)	0.290		
2–4 weeks	283 (66.4)	143 (33.6)	426 (20.2)	0.99 (0.76, 1.29)	0.950		
4–8 weeks	212 (74.1)	74 (25.9)	286 (13.5)	0.68 (0.50, 0.93)	0.020		
>8 weeks	239 (67.7)	114 (32.3)	353 (16.7)	0.94 (0.71, 1.23)	0.640		
DR type							
MDR	59 (69.4)	26 (30.6)	85 (4.0)	Ref.			
Pre-MDR	114 (64.4)	63 (35.6)	177 (8.4)	1.25 (0.73, 2.21)	0.420		
XDR	1 (100.0)	0 (0.0)	1 (0.0)	0.00	0.970		
Pre-XDR	2 (66.7)	1 (33.3)	3 (0.1)	1.13 (0.05, 12.36)	0.920		
Sensitive	1111 (69.0)	499 (31.0)	1610 (76.2)	1.02 (0.64, 1.66)	0.940		
Other	155 (65.1)	83 (34.9)	238 (11.3)	1.22 (0.72, 2.09)	0.470		
Urban/Rural							
Rural	1057 (68.8)	480 (31.2)	1537 (72.7)	Ref.			
Urban	385 (66.7)	192 (22.3)	577 (27.3)	1.10 (0.89, 1.35)	0.370		
Sublineage							
L4.4.2	267 (77.2)	79 (22.8)	346 (16.4)	Ref.			
L2.2.1	892 (65.8)	463 (34.2)	1508 (64.1)	1.83 (1.40, 2.41)	<0.001	1.59 (1.20, 2.12)	<0.001
L2.2.2	55 (67.9)	26 (32.1)	81 (3.8)	1.60 (0.93, 2.69)	0.080	1.55 (0.88, 2.66)	0.120
L4.2.2	40 (78.4)	11 (21.6)	51 (2.4)	0.93 (0.44, 1.84)	0.840	0.78 (0.35, 1.62)	0.530
L4.5	85 (77.3)	25 (22.7)	110 (5.2)	0.99 (0.59, 1.64)	0.980	0.83 (0.48, 1.41)	0.510
Other	16 (0.8)	2 (0.1)	18 (0.9)	0.42 (0.07, 1.53)	0.260	0.47 (0.07, 1.74)	0.320

**Abbreviation:** Ref, reference.

different places. The median center point of group A and the unclustered group D were located on Ninghai Street and Gangbu Farm in Haizhou District. The median center point of group B was located in Sucheng Street, Lianyung District. The median center point of group C was located in Xinan Town, Guannan County. Participants in 3 cluster groups were all dispersed in a north-south pattern. However, the distribution of group A was the most compact. Clustered groups had a more compact distribution than unclustered group D (Figure 4).

Kernel density estimations reveal the potential spatial clusters of participants of each group, which is demonstrably similar to the median center results. Groups A and B had similar case concentration areas in Haizhou District, while cases of group C were mainly distributed in the southern townships of Guanyun County. High-density areas of unclustered and cases were mainly distributed in Haizhou District, separately followed by Ganyu District and Guanyun District (Figure 4). The results of the K-function analysis display that the partial curve is outside of the envelope, which also supports the phenomenon of spatial clustering. K-function differences showed that groups A had a larger scale clustering pattern (up to  $\approx 7.5$  km) than groups B and C (Figure 5). Thirty cases (65.2%) in the three clustered groups had an identifiable epidemiologic link. Seven patients in group A lived in the same village in Ninghai Street, and two with the same surname and address were presumed to be paternally related. Six cases in group B and eight cases in group C lived



**Figure 3** Kernel density maps of clustered strains in 3-year sliding window intervals.

in the same towns. The 22 cases in group C resided in adjacent townships or streets (Figure 6). All cases diagnosed later in group A were situated near the initial participant, while group C exhibited a general trend of increasing distance as the diagnosis time progressed (Figure 7).

**Table 2** Characteristics of Patients of Large-Cluster Groups and Unclustered Group in Lianyungang, 2011–2020

Variables	Group				P
	A (n=11)	B (n=12)	C (n=23)	D (n=1442)	
Lineage	2.2.1	2.2.1	2.2.1		
Gender (M/F)	10/1 (90.91/ 9.09)	8/4 (66.67/ 33.33)	21/2 (91.30/ 8.70)	1124/318 (77.95/ 22.05)	<0.001
Age (years), median	37.00 (27.50, 49.50)	21.50 (19.75, 27.25)	20.00 (19.00, 23.50)	58.00 (37.00, 70.00)	<0.001
Age group					
≤ 24	1 (9.09)	7 (58.33)	18 (78.26)	158 (10.96)	
25–44	6 (54.55)	4 (33.33)	4 (17.39)	294 (20.39)	
45–64	4 (36.36)	1 (8.33)	1 (4.35)	475 (32.94)	
≥ 65	0 (0.00)	0 (0.00)	0 (0.00)	515 (35.71)	
Occupation					<0.001
Farmer	5 (45.45)	1 (8.33)	13 (56.52)	1103 (76.49)	
Student	1 (9.09)	4 (33.33)	8 (34.78)	31 (2.15)	
Service sector	4 (36.36)	5 (41.67)	1 (4.35)	136 (9.43)	
Laborer	1 (9.09)	2 (16.67)	1 (4.35)	74 (5.13)	
Retire	0 (0.00)	0 (0.00)	0 (0.00)	77 (5.34)	
Other	0 (0.00)	0 (0.00)	0 (0.00)	21 (1.46)	
Location					<0.001
Donghai	0 (0.00)	0 (0.00)	0 (0.00)	558 (38.70)	
Ganyu	0 (0.00)	0 (0.00)	0 (0.00)	93 (6.45)	
Guannan	0 (0.00)	0 (0.00)	21 (91.30)	212 (14.70)	
Guanyun	1 (9.09)	2 (18.18)	1 (4.35)	287 (19.90)	
Haizhou	9 (81.82)	3 (27.27)	1 (4.35)	223 (15.46)	
Lianyungang	1 (9.09)	6 (54.55)	0 (0.00)	69 (4.79)	

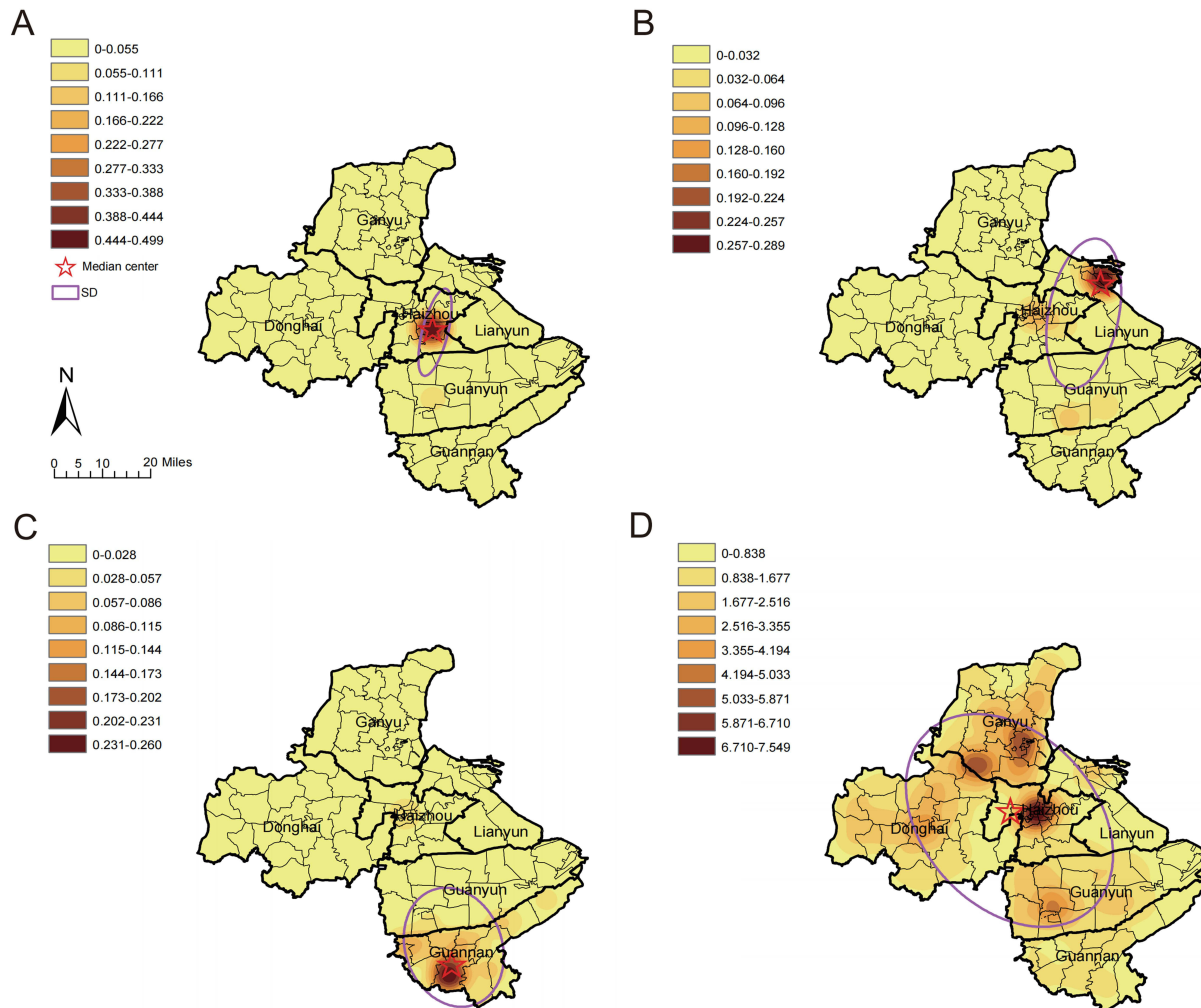
(Continued)

**Table 2** (Continued).

Variables	Group				P
	A (n=11)	B (n=12)	C (n=23)	D (n=1442)	
Rural/Urban	1/10 (9.09/90.91)	2/10 (16.67/83.33)	22/1 (95.65/4.35)	1150/292 (79.75/20.25)	<0.001
Initial/Relapse	11/0 (100.00/0.00)	10/1 (90.91/9.09)	21/2 (91.30/8.70)	1229/213 (85.23/14.77)	0.548
Diagnostic delay (No/Yes)	5/6 (45.45/54.55)	1/11 (8.33/91.67)	9/14 (39.13/60.87)	414/1028 (28.71/71.29)	0.174
Delay time					0.246
≤ 2 weeks	1 (9.09)	3 (25.00)	6 (26.09)	294 (20.39)	
2–4 weeks	2 (18.18)	3 (25.00)	5 (21.74)	283 (19.63)	
4–8 weeks	1 (9.09)	0 (0.00)	0 (0.00)	212 (14.70)	
> 8 weeks	2 (18.18)	5 (41.67)	3 (13.04)	239 (16.57)	

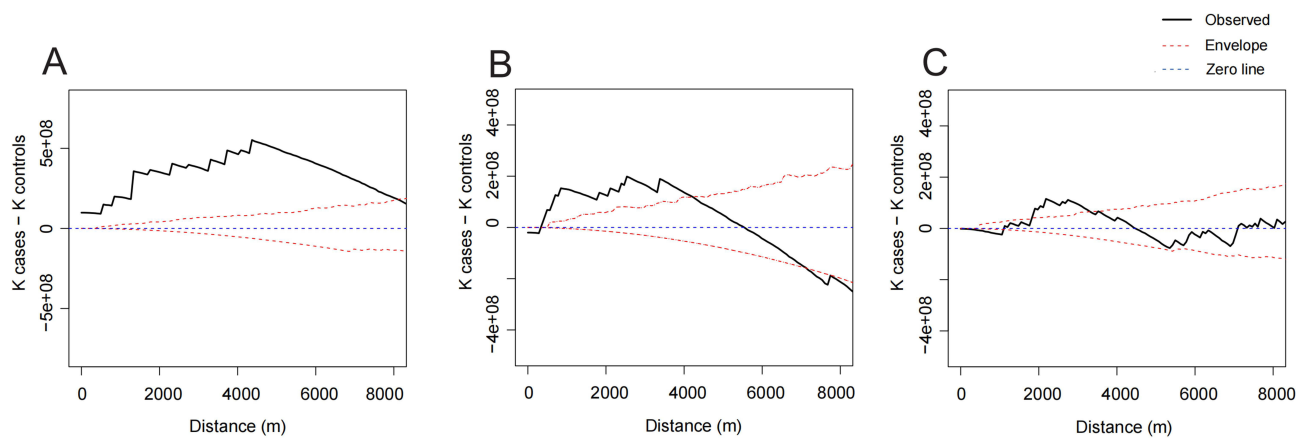
## Discussion

Ending the TB epidemic in high-burden settings is critical while interrupting TB transmission is one of the key strategies for achieving this goal. In this 10-year retrospective study in eastern China, we observed areas with a high risk of transmission for *M. tuberculosis*, and the risk varied among different populations.

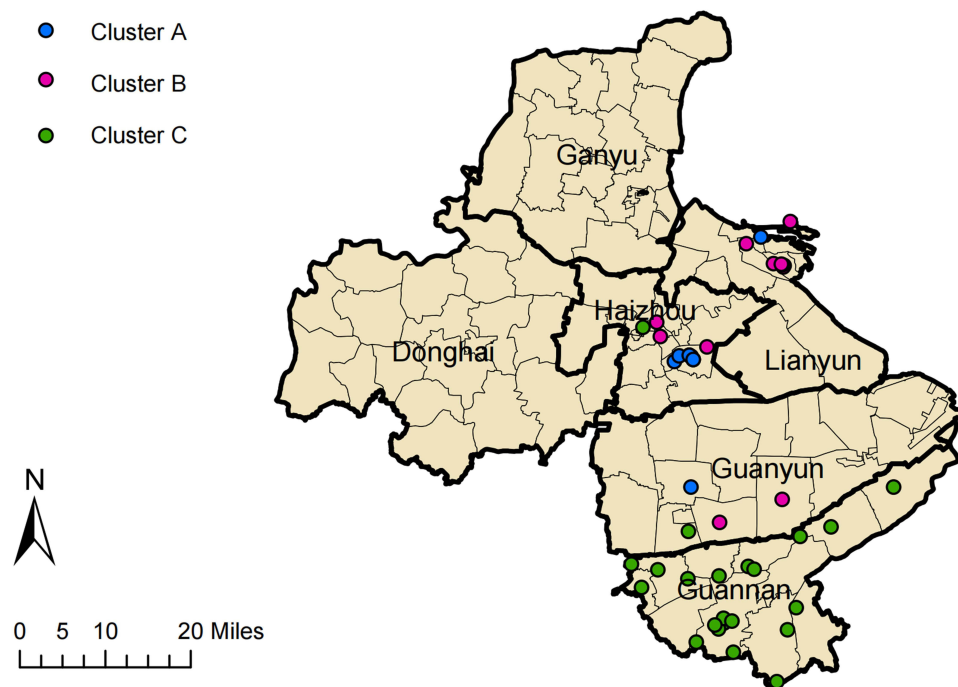


**Figure 4** Kernel density map, directional distribution, and median center point for clustered and unclustered groups. Purple oval shapes cover the region inside the standard deviation ellipse, representing the geographical distance and the directional orientation of the participants' positions in each group. Red stars represent the median center point of each group. Clustered groups: (A-C); unclustered group: (D).



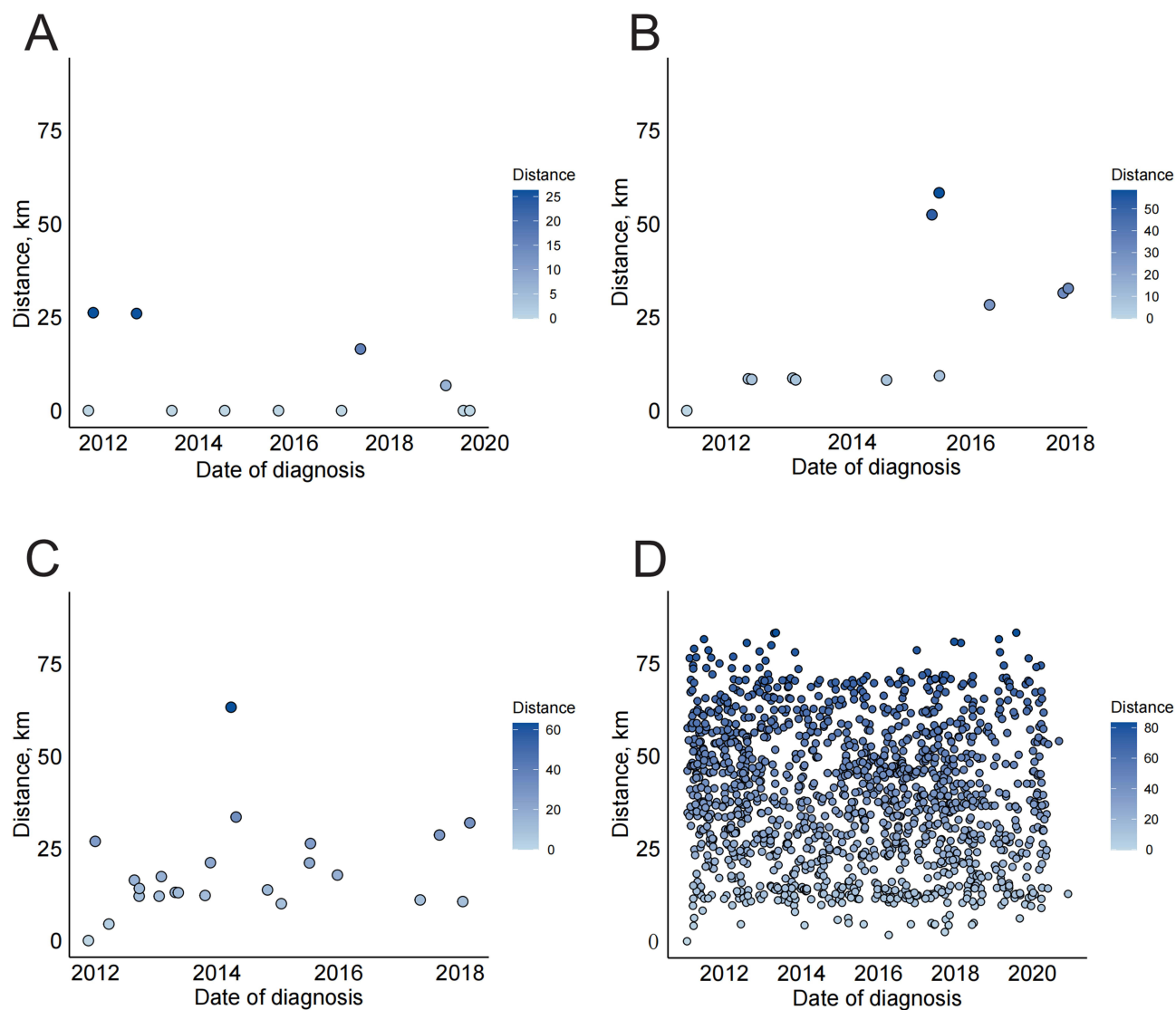


**Figure 5** K-function differences for assessing geospatial clustering of clustered groups compared with unclustered group. The x-axis represents distances between points; the y-axis represents the K-value. The dotted lines represent the 95% confidence envelope of complete spatial randomness, while the solid line indicates the obtained values. Observations that fall above the upper 95% envelope indicate significant spatial clustering. Clustered groups: (A-C).



**Figure 6** Distribution of three large clustered groups of *M. tuberculosis* strains.

Clusters of *M. tuberculosis* strains are often considered evidence of recent transmission. In the current study, the recent TB transmission rate was estimated to be 31.79%, which was significantly higher than that reported in Songjiang (25.2%), Shenzhen (12.2%) and Ghana (24.7%).<sup>15–17</sup> Younger people had a higher risk of recent transmission. Strains were more likely to be clustered in the population aged  $\leq 45$  years, consistent with the results of a similar study conducted in Botswana.<sup>18</sup> The localized transmission in association with younger people may be due to the frequency and intensity of their social activities throughout life.<sup>19</sup> Younger patients may have more social contact and engagement with non-family members,<sup>19</sup> further illustrating the role in TB transmission. Older patients, on the other hand, may develop TB with non-genotypically aggregated strains by developing infections from the distant past.<sup>20</sup>



**Figure 7** Geographic distance between the first participant diagnosed with TB (shown in each plot at a distance of 0 km) and those subsequently diagnosed in clustered and unclustered groups. Clustered groups: (A-C); unclustered group: (D).

More notably, our results found a higher risk of transmission among students. Jiang et al<sup>21</sup> noted an increased risk of multidrug-resistant TB among students in Shenzhen. Faccini et al<sup>22</sup> reported that TB outbreaks most often occurred in schools due to delayed diagnosis, persistent exposure, and school overcrowding. Therefore, TB prevention and control in schools should be paid attention to, and infectious tuberculosis patients should be identified and treated in a timely manner through symptom screening and routine physical examination.<sup>23</sup>

L2.2.1 strains had a higher risk of transmission, which is consistent with the widely reported predominance of Beijing strains that are more likely to transmit. In particular, the modern Beijing strains (L2.3) have dominated the *M. tuberculosis* population in China for the past century.<sup>24</sup> This could be attributed to the inherent increase in virulence caused by mutations in *ppe38*, which block the secretion of ESX-5 substrate,<sup>6</sup> along with the historical population growth and widespread migration of the Beijing population, which promoted the transmission and spread of this sublineage.<sup>24</sup>

Identifying high-density areas of recent TB transmission helps TB control programs target specific interventions to areas at the highest risk. The kernel density map shows that Haizhou District, Donghai County, and Guanyun County are high-density areas with clusters of cases. Previous studies have shown that geographical aggregation of TB cases can be linked to genotypic clustering.<sup>25,26</sup> In our study site, three large clusters displayed extremely close geographic proximity.

The distance between the first and subsequent cases varied geographically among the groups. Patients within clusters were concentrated in specific areas, while the non-clustered patients were more widely distributed.

Despite the continuous progress and decreasing costs of WGS-based typing techniques, some critical challenges remain, such as a lack of standardization of genomic distances (SNP distances) for cluster definition.<sup>27</sup> Peru and some other countries used SNPs  $\leq 5$  as the threshold for the definition of gene clustering,<sup>28</sup> while Malawi<sup>29</sup> and Ghana<sup>17</sup> used SNPs  $\leq 10$  as the threshold value. In the current study, we applied an SNP threshold of 12, which was also used in Beijing and Shenzhen.<sup>30,31</sup> But it might overestimate the recent transmission.

We did not observe clusters of multidrug-resistant TB strains were detected in any of the three outbreak cohorts of our study, which supported previous findings that the transmission ability of drug-resistant *M. tuberculosis* is lower than that of sensitive strains.<sup>32</sup> However, some studies in Beijing and Shenzhen have shown different results. Yin et al<sup>30</sup> found that about 63% of multidrug-resistant TB cases were caused by recent transmission. Jiang et al<sup>21</sup> suggested that patients with multidrug-resistant TB were more likely to have genomically clustered isolates.

Our study has some limitations. Firstly, we could not obtain detailed social contact data of all cases, thus limiting our ability to infer epidemiologic links. Secondly, spatial analysis was limited to the patient's residential address at diagnosis. Actual transmission may have occurred in other settings, such as workplaces, social and recreational venues, and transportation facilities. Additional WGS and epidemiological data combined with spatial and social network analyses may help us to reconstruct potential transmission chains better, and more complete data may lead to the discovery of larger or more outbreak clusters or altered geospatial patterns.

## Conclusion

There are areas with a high risk of transmission for *M. tuberculosis* in the research site, and the risk varies among different populations. Accurate prevention strategies targeted at specific regions and key populations, such as a 1–2 year follow-up for close contacts of TB patients, can help curb the prevalence of TB.

## Data Sharing Statement

All data are available from the corresponding author upon reasonable request.

## Ethics Approval

This study was conducted in accordance with the Declaration of Helsinki and approved by the ethics committee of Nanjing Medical University. After obtaining informed consent from all participants, we collected their demographic data.

## Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

## Funding

This study was funded by the National Natural Science Foundation of China (82473693, 81973103), Medical Research Project of Jiangsu Health Commission (ZDB2020013), Nanjing Major Science and Technology Specific Project (2021–11005), and Jiangsu Province Graduate Research and Innovation Program (KYCX21\_1547). The funding agencies did not play a role in the study.

## Disclosure

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

## References

1. Lawn SD, Zumla AI. Tuberculosis. *Lancet*. 2011;378(9785):57–72. doi:10.1016/s0140-6736(10)62173-3
2. Bagcchi S. WHO's global tuberculosis report 2022. *Lancet Microbe*. 2023;4(1):e20. doi:10.1016/s2666-5247(22)00359-7
3. Xu G, Mao X, Wang J, Pan H. Clustering and recent transmission of Mycobacterium tuberculosis in a Chinese population. *Infect Drug Resist*. 2018;11:323–330. doi:10.2147/idr.S156534
4. Schiebelhut LM, Abboud SS, Gómez Daglio LE, Swift HF, Dawson MN. A comparison of DNA extraction methods for high-throughput DNA analyses. *Mol Ecol Resour*. 2016;17(4):721–729. doi:10.1111/1755-0998.12620
5. Kohl TA, Utpatel C, Schleusener V, et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. *PeerJ*. 2018;6:e5895. doi:10.7717/peerj.5895
6. Kay GL, Sergeant MJ, Zhou Z, et al. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun*. 2015;6(1):6717. doi:10.1038/ncomms7717
7. Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med*. 2016;14(1):21. doi:10.1186/s12916-016-0566-x
8. Zhou T, Xu K, Zhao F, et al. itol.toolkit accelerates working with iTOL (Interactive Tree of Life) by an automated generation of annotation files. *Bioinformatics*. 2023;39(6):btad339. doi:10.1093/bioinformatics/btad339
9. Wu H, Lin A, Clarke KC, Shi W, Cardenas-Tristan A, Tu Z. A comprehensive quality assessment framework for linear features from Volunteered Geographic Information. *Int J Geog Inf Sci*. 2020;35(9):1826–1847. doi:10.1080/13658816.2020.1832228
10. Rocchini D, Wang B, Shi W, Miao Z. Confidence analysis of standard deviational ellipse and its extension into higher dimensional euclidean space. *PLoS One*. 2015;10(3):e0118537. doi:10.1371/journal.pone.0118537
11. Dixon P Ripley's K function; 2001.
12. Tang W, Feng W, Jia M. Massively parallel spatial point pattern analysis: ripley's K function accelerated using graphics processing units. *Int J Geog Inf Sci*. 2015;29(3):412–439. doi:10.1080/13658816.2014.976569
13. Wheeler DC. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996 – 2003. *Int J Health Geogr*. 2007;6(1):13. doi:10.1186/1476-072x-6-13
14. Baddeley A, Turner R. spatstat: anRPackage for analyzing spatial point patterns. *Journal of Statistical Software*. 2005;12(6):1–42. doi:10.18637/jss.v012.i06
15. Li M, Lu L, Jiang Q, 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. doi:10.1016/j.lanwpc.2023.100833
16. Yang T, Wang Y, Liu Q, 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. doi:10.1016/j.lanwpc.2021.100106
17. Asare P, Otchere ID, Bedeley E, et al. Whole genome sequencing and spatial analysis identifies recent tuberculosis transmission hotspots in Ghana. *Front Med*. 2020;7:161. doi:10.3389/fmed.2020.00161
18. Zetola NM, Moonan PK, Click E, et al. Population-based geospatial and molecular epidemiologic study of tuberculosis transmission dynamics, Botswana, 2012–2016. *Emerging Infect Diseases*. 2021;27(3):835–844. doi:10.3201/eid2703.203840
19. Sander J, Schupp J, Richter D. Getting together: social contact frequency across the life span. *Develop Psychology*. 2017;53(8):1571–1588. doi:10.1037/dev0000349
20. Yew WW, Yoshiyama T, Leung CC, Chan DP. Epidemiological, clinical and mechanistic perspectives of tuberculosis in older people. *Respirology*. 2018;23(6):567–575. doi:10.1111/resp.13303
21. Jiang Q, Liu Q, Ji L, et al. Citywide transmission of multidrug-resistant tuberculosis under china's rapid urbanization: a retrospective population-based genomic spatial epidemiological study. *Clin Infect Dis*. 2020;71(1):142–151. doi:10.1093/cid/ciz790
22. Faccini M, Codecasa LR, Ciconali G, et al. Tuberculosis outbreak in a primary school, Milan, Italy. *Emerging Infect Diseases*. 2013;19(3):485–487. doi:10.3201/eid1902.120527
23. You NN, Zhu LM, Li GL, et al. A tuberculosis school outbreak in China, 2018: reaching an often overlooked adolescent population. *Epidemiol Infect*. 2019;147:e303. doi:10.1017/s0950268819001882
24. Liu Q, Ma A, Wei L, et al. China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. *Nat Ecol Evol*. 2018;2(12):1982–1992. doi:10.1038/s41559-018-0680-6
25. Yang C, Lu L, Warren JL, et al. Internal migration and transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, genomic analysis. *Lancet Infect Dis*. 2018;18(7):788–795. doi:10.1016/s1473-3099(18)30218-4
26. Zelner JL, Murray MB, Becerra MC, et al. Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. *J Infect Dis*. 2016;213(2):287–294. doi:10.1093/infdis/jiv387
27. Merker M, Kohl TA, Niemann S, Supply P. The evolution of strain typing in the mycobacterium tuberculosis complex. *Adv Exp Med Biol*. 2017;1019:43–78. doi:10.1007/978-3-319-64371-7\_3
28. Bui DP, Chandran SS, Oren E, et al. Community transmission of multidrug-resistant tuberculosis is associated with activity space overlap in Lima, Peru. *BMC Infect Dis*. 2021;21(1):275. doi:10.1186/s12879-021-05953-8
29. Guerra-Assunção JA, Crampin AC, Houben R, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *eLife*. 2015;4(e05166). doi:10.7554/eLife.05166
30. Yin J, Zhang H, Gao Z, et al. Transmission of multidrug-resistant tuberculosis in Beijing, China: an epidemiological and genomic analysis. *Front Public Health*. 2022;10:1019198. doi:10.3389/fpubh.2022.1019198
31. Mijiti P, Liu C, Hong C, et al. Implications for TB control among migrants in large cities in China: a prospective population-based genomic epidemiology study in Shenzhen. *Emerging Microbes Infect* 2024;13(1):2287119. doi:10.1080/22221751.2023.2287119
32. Asare P, Asante-Poku A, Prah DA, et al. Reduced transmission of Mycobacterium africanum compared to Mycobacterium tuberculosis in urban West Africa. *Inter J Infect Dis*. 2018;73:30–42. doi:10.1016/j.ijid.2018.05.014

Infection and Drug Resistance

Dovepress

### Publish your work in this journal

Infection and Drug Resistance is an international, peer-reviewed open-access journal that focuses on the optimal treatment of infection (bacterial, fungal and viral) and the development and institution of preventive strategies to minimize the development and spread of resistance. The journal is specifically concerned with the epidemiology of antibiotic resistance and the mechanisms of resistance development and diffusion in both hospitals and the community. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/infection-and-drug-resistance-journal>