Published in final edited form as:

Lancet Infect Dis. 2018 July ; 18(7): 788–795. doi:10.1016/S1473-3099(18)30218-4.

Internal migration and the transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, and genomic analysis

Chongguang Yang, PhD^{1,2}, Liping Lu, MPH³, Joshua L. Warren, PhD⁴, Jie Wu, MPH⁵, Qi Jiang, BSc², Tianyu Zuo, BSc², Mingyu Gan, BSc², Mei Liu, BSc², Qingyun Liu, PhD², Kathryn DeRiemer, PhD⁶, Jianjun Hong, BSc³, Xin Shen, PhD⁵, Caroline Colijn, PhD⁷, Xiaoqin Guo, BMedSc³, Qian Gao, PhD², and Ted Cohen, DPH¹

¹Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, 60 College Street, New Haven, Connecticut, USA, 06510

²Key Laboratory of Medical Molecular Virology, Ministry of Education and Health, School of Basic Medical Sciences, Shanghai Public Health Clinical Center, Fudan University, 131 DongAn Road, Shanghai, China, 200032

³Department of Tuberculosis Control, Songjiang District Center for Disease Control and Prevention, 1050 Xi Lin Northern Road, Shanghai, China, 201620

⁴Department of Biostatistics, Yale School of Public Health, Yale University, 60 College Street, New Haven, Connecticut, USA, 06510

⁵Department of Tuberculosis Control, Shanghai Municipal Center for Disease Control and Prevention, 1380 West Zhong Shan Road, Shanghai, China, 200336

⁶School of Medicine, University of California, Davis, One Shields Avenue, Davis, CA, USA, 95616

⁷Department of Mathematics, Imperial College London, Kensington, London SW7 2AZ, UK

Summary

Background—Massive internal migration from rural to urban areas poses new challenges for tuberculosis control in China. We sought to combine genomic, spatial, and epidemiological data to describe the dynamics of tuberculosis in an urban setting receiving large numbers of migrants.

Methods—We conducted a population-based study of culture-positive *Mycobacterium tuberculosis* isolates in Songjiang, Shanghai between January 1, 2009 and December 31, 2015. We used whole-genome sequencing to discriminate apparent genetic clusters of *Mycobacterium*

Correspondence to: Qian Gao.

Conflict of interest

We declare that we have no conflict of interest.

Contributors

C.Y., X.G., T.C., and Q.G. contributed to the conception, design and management of the study. C.Y., L.L., Q.J., M.L., J.W., X.S., and K.D. contributed the microbiological and epidemiological data. L.L., Q.J., and J.H. contributed to the epidemiological investigations. T.Z., M.G., Q.L., and C.Y. contributed to the bioinformatics analyses. C.Y. and J. L.W. contributed to the geographical data analyses. C.Y. and C.C. contributed to the transmission inference analyses. C.Y., Q.G., and T.C. did the genomic, epidemiological, and statistical analyses, and prepared the manuscript. All authors contributed to and reviewed the final report.

tuberculosis sharing identical variable-number-tandem-repeat (VNTR) patterns and analyzed the relationship between proximity of residence and the risk of genomically-clustered *Mycobacterium tuberculosis*. Finally, we used genomic, spatial, and epidemiological data to estimate time of infection and transmission links among migrants and residents.

Findings—The majority (1211/1620, 75%) of culture-positive tuberculosis occurred among migrants. 150 of 218 (69%) individuals sharing identical VNTR patterns had isolates within ten single-nucleotide polymorphisms (SNPs) of at least one another strain, consistent with recent transmission of *Mycobacterium tuberculosis*. Pairs of strains collected from individuals living in closer proximity were more likely to be genetically similar. For every additional kilometer of distance between subjects' homes, the odds that genotypically-matched strains were within ten SNPs of each other decreased by about 10% (OR 0.89, 95% CI 0.87–0.91). We inferred that transmission from residents to migrants occurs as commonly as transmission from migrants to residents and we estimated that more than two-thirds of migrants in genomic clusters were infected locally after migration.

Interpretation—The primary mechanism driving local incidence in urban centers is local transmission involving both migrants and residents. Consideration of epidemiologic, genomic, and spatial data contributes to a richer understanding of local transmission dynamics and should inform the design of more effective interventions.

Funding—National Natural Science Foundation, National Science and Technology Major Project of China, and US National Institutes of Health

Keywords

Genomic epidemiology; Migration; Spatial analysis; Transmission inference; Tuberculosis

INTRODUCTION

Tuberculosis (TB) hits hardest in poor and vulnerable populations, and migrants are particularly affected. China has the third largest TB epidemic worldwide, with an estimated 0.9 million new cases and 52,000 deaths due to TB in 2016.^{1–2} Within the past two decades, migration within China (internal migration), primarily from rural to urban environments, has become very common and poses new challenges for TB control in cities.³

Today, most internal migrants in China are young adults who travel from the rural countryside to enter the wage economy in urban centers. In 2010, there were an estimated 260 million internal migrants (18% of the total national population), a number that is expected to increase in the coming decade.⁴

The notification rate of TB has been substantially higher in rural communities in China than in urban communities leading to predictions that internal migration from rural communities will result in increased notifications and transmission of TB in cities.^{3,5} This hypothesis is supported by the fact that rural-to-urban internal migrants, due at least in part to China's household registration (*Hukou*), face significant restrictions in accessing subsidized housing, social security, and medical care and insurance in urban centers.^{3,6–8} These migrants are also more likely than urban residents to share crowded living conditions, have low socio-

economic status, and minimal formal education, all of which are determinants often associated with greater risk of TB.^{7–8} Others have documented that such internal migrants have a greater delay between symptom onset and clinical presentation or TB diagnosis and experience less favorable treatment outcomes.^{9–11}

Here we combine epidemiological, molecular genetic, and spatial analysis to investigate the transmission dynamics of TB in an urban area experiencing rapid population growth as a result of internal migration.

METHODS

Study population

We conducted a population-based study of all culture-positive TB cases diagnosed in the Songjiang District of Shanghai between January 1, 2009 and December 31, 2015. Shanghai has experienced large population increases as a result of internal migration over the past two decades and in 2010 migrants constituted greater than 40% of the city's population. The majority of such internal migrants have settled in seven suburban districts within the city, of which Songjiang is typical in terms of the fraction of the population that is migrants, the demographic, educational, and job-seeking characteristics of migrants, and the fraction of notified TB that occurs among migrants (Table S1, Appendix, page 7 and Table S2, Appendix, pages 8–9). Currently, nearly two thirds of those living in Songjiang (62%, 1·10 of 1·77 million) are internal migrants; these migrants are defined as those without a Shanghai household registration status through Chinese *Hukou* system.

Beginning in 2004, the Shanghai Municipal Center for Disease Control and Prevention (Shanghai Municipal CDC) implemented a new policy extending free TB treatment to all migrants; such services were previously not freely available to individuals outside of their originally registered residence. During the period of our study, all individuals with TB symptoms including cough for at least two weeks, fever, chest pain, weight loss, night sweats, and abnormal chest radiograph had three sputum samples collected (spot, early morning, and night) and tested for *Mycobacterium tuberculosis* (*M. tuberculosis*) by microscopic exam and culture, consistent with national guidelines.² For individuals who were unable to produce sputum spontaneously, sputum induction was used. From each individual with culture-positive TB, a single pre-treatment sputum specimen was submitted to the Shanghai Municipal CDC TB reference laboratory for rifampin (RIF) and isoniazid (INH) drug susceptibility testing (DST) by the proportion method on Löwenstein-Jensen medium.

Epidemiological investigations of culture-positive TB patients

Demographic, clinical, and microbiological records were obtained from the routine TB surveillance system.¹² Epidemiological investigations were conducted within a week of TB diagnosis using a standardized questionnaire. Information collected included current residential address, the original household registration province and home address of migrants, identities of close contacts (i.e. family members, work colleagues and classmates),

known history of exposure to individuals with active TB disease, and specific locations visited where transmission may have plausibly occurred.

Genotyping and whole genome sequencing of *M. tuberculosis* isolates

Genomic DNA was extracted from the culture from one sputum specimen per patient using the Cetyl Trimethyl Ammonium Bromide (CTAB) method. Beijing family strains (the most prevalent strains in China) were identified by a PCR-based assay and were confirmed by examination of specific single nucleotide polymorphisms (SNPs).¹³ We first performed variable number tandem repeat (VNTR) genotyping for each strain using a 9+3 loci set with hypervariable loci, which was developed specifically for discriminating strains within China where Beijing strains predominate.¹⁴ To further discriminate closely-related strains, we used whole-genome sequencing (WGS) for the subset of strains that shared identical VNTR genotype. Given resource limitations, we focused our sequencing efforts on the subset of strains that were in VNTR clusters of three or larger. WGS was done by Illumina HiSeq 2000 with an expected coverage of 100×. Paired-end reads were mapped to the reference genome H37Rv (GenBank AL123456) with Bowtie 2 (v2.3.1). The SAMtools (v1.6)/VarScan (v2.3.6) suite was used for calling SNPs (frequency ≥ 75%, further details provided in the Appendix, page 2). Drug-resistance associated genes and PE/PGRS and PE/PPE genes were excluded from our SNP analysis.¹⁵ For our primary analysis, we defined genomically-clustered strains as those within a threshold of ten or fewer SNPs;¹⁶ this SNP threshold was varied in subsequent sensitivity analyses. The sequencing data were deposited in the NCBI Sequence Read Archive (SRP124760).

Analysis of the relationship between spatial distance and genomic similarity of strains

The residential address of each TB patient at the time of diagnosis was geocoded by using ArcGIS (Esri, Redlands, CA, USA) and Google Maps tools (Google Inc., CA, USA). We also used a Chinese language based web geocoding tool from Baidu Maps to verify locations. We employed kernel density estimation to identify patterns of spatial aggregation of TB patients.¹⁷ We used multivariable logistic regression to test the hypothesis that spatial proximity was associated with greater genetic similarity (i.e. a smaller SNPs distance) among pairs of *M. tuberculosis* isolates (Appendix, pages 3–4).

Transmission inference based on epidemiological, genomic, and spatial analysis

We reconstructed plausible transmission networks based on consideration of (i) phylogenetic analysis, using a median-joining method with NETWORK (version 5.0); (ii) epidemiological linkages from questionnaire data;¹⁸ and (iii) spatial distribution of the residential locations of the genomically clustered TB patients. We also used a Bayesian evolutionary analysis to infer a time-labeled phylogeny by sampling trees (BEAST v1.8.4) and followed by using a Markov chain Monte Carlo (MCMC) based approach (TransPhylo, R package) to infer the estimated infection times and putative transmission directions between resident and migrant patients based on a consensus transmission tree that summarized the MCMC outputs (Appendix, pages 2–3).¹⁹

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

RESULTS

Study population

From 2009 to 2015, 1649 individuals were diagnosed with culture-positive TB in the Songjiang district. Of these, 29 were incarcerated prisoners or individuals without sufficient baseline information and were excluded from the study sample (Figure 1A). Of the remaining 1620 individuals with TB, 409 (25%) were residents and 1211 (75%) were internal migrants.

Eighty-nine percent of the migrants with incident TB (1073 of 1211) came to Songjiang from central and western regions of China, both of which have substantially higher notification rates of TB than Shanghai (Figure 1B and Figure S1, Appendix, page 13).

Migrants with TB were significantly younger than residents with TB (median age, 27 vs. 55 years; $p < 0.001$), were more often female (36% vs. 21%; $p < 0.001$) and more often first-time TB cases (93% vs. 87%, $p < 0.001$; Table 1). In addition, migrants with TB were less likely to have been diagnosed in TB-designated hospital compared with resident patients (14% vs. 25%, $p < 0.0001$). Seventy-five percent of migrant patients (776/1031) had migrated to Songjiang less than five years before being diagnosed with TB (Table S3, Appendix, page 10).

We display the crude notification rates of culture-positive TB by age (Figure S2, Appendix, page 14). Based on data from the 2010 national census,²⁰ the age-adjusted notification rate of culture-positive TB of migrants (18.20/100,000 per year) was higher than that of residents (8.35/100,000 per year).

Cluster identification based on VNTR genotyping and whole-genome sequencing

Of the 1607 culture-positive individuals with sufficient genotyping data (Figure 1A), isolates from 488 (30%) shared identical VNTR patterns; these isolates were grouped in 185 VNTR clusters with size ranging from two (131 clusters) to 11 (one cluster) patients (Table S4, Appendix, page 11). Beijing family strains were the most commonly observed isolates (79%, 1274/1607).

To further discriminate the apparent genetic clusters,²¹ we performed WGS for all of the isolates from VNTR clusters of size three and larger. In total, we obtained sufficient WGS data from 218 of 226 isolates from 52 different VNTR clusters (Figure 1A). Analysis of the pairwise nucleotide differences within each VNTR cluster revealed that the isolates from 150 individuals (102 migrants and 48 residents) were within ten SNPs of at least one other isolate from the same VNTR cluster, resulting in forty-four genomic clusters with size ranging from two to eight (Table S4 and Figure S3, Appendix, page 11 and 15).

We examined the genetic distance among pairs of *M. tuberculosis* isolates from three different VNTR cluster types: those consisting of residents only, those consisting of migrants only, and those consisting of both residents and migrants (mixed clusters). After exclusion of the pairs separated by more than 100 SNPs (which were clearly not related through recent transmission by any reasonable SNP threshold), isolates from resident-only VNTR clusters had a significantly larger genetic distance (median difference of 19 SNPs) than migrant-only (median difference of two SNPs) and mixed VNTR clusters (median difference of three SNPs, $p < 0.0001$, Figure 2).

While the observed patterns of genotypic and genomic clustering suggest the possibility of local transmission between migrants and between migrants and residents within Songjiang, migrants may also have been infected prior to their arrival and thus such genetic similarity may be an artefact of infections imported from other parts of the country. Further inspection of genomic clusters involving at least two migrants reveals that most migrants (84%) in such genomic clusters traveled to Songjiang from different provinces, suggesting that such clustering likely reflects local transmission occurring after their arrival in Songjiang (Table S5, Appendix, page 12).

Spatial distribution of TB patients

Migrants living in Songjiang reside primarily within several sub-districts with industrial parks (Figure S4, Appendix, page 16). One of these industrial sub-districts has the highest TB notification rate (Figure 3A) and 94% (254/269) of TB notifications this sub-district were migrants. Kernel density analysis of the home residence reveals two distinct spatial distributions, with TB among migrants concentrated in multiple areas of eastern Songjiang (Figure 3B) and TB among residents concentrated in the center (i.e. downtown) of the district (Figure 3C). This pattern is consistent with the overall distribution of migrants and residents in Songjiang (Figure S4, Appendix, page 16).

To test the hypothesis that spatial proximity of home residences was associated with genomic similarity of isolates, we used a multivariable logistic regression model to evaluate the relationship between genomic clustering (as defined by a SNP threshold) and spatial proximity. The main model we use includes adjustment for the genetic sublineage of the strains and an indicator variable which tracks whether the pair of isolates was from two migrants, two residents, or one resident and one migrant; we note that a second model which does not adjust for these factors produces similar results but was not preferred because of its higher AIC value (Table 2). We found that the odds of two strains being genomically clustered decreased by about 10% for every additional kilometer between residences, and that this result was robust using different thresholds for defining these strains as genetically clustered (3–100 SNPs) (Table 2). Similarly, we found that the median geographic distance between pairs of isolates increased as the genetic distance threshold for a genomic cluster increased (Figure S5, Appendix, page 17).

Inference of TB transmission based on genomic, spatial and epidemiological data

Our epidemiological investigations, which occurred before the pathogen genetic data were available, identified possible links among only 26 (18%) of the 142 individuals within

genomic clusters. Nine percent (4/46) of the residents in genomic clusters reported epidemiological links to other cases (all among reported close contacts), compared with 23% (22/96) of the migrants in genomic clusters ($p=0.04$). Notably, not a single link between a migrant and resident was identified through routine epidemiological investigation.

Using epidemiologic and phylogenetic information, we reconstructed putative transmission networks among individuals sharing genomically-clustered strains; for ease of visualization, we focus on clusters of at least four individuals (Figure 4). Despite the absence of epidemiological links, the genetic distances between a migrant and resident in any of the mixed clusters were no greater than a single SNP. Three genomic clusters provide evidence that transmission of MDR-TB between residents and migrants occurred (Cluster-01 and 03 in Figure 4A and Cluster-24 in Figure S6, Appendix, page 18).

We observed spatial aggregation within the majority of genomic clusters, including many migrant-only (Figure 4B, Cluster-11, -12 -13 and -15), resident-only (Cluster-02), and all of the mixed clusters except Cluster-10. In some of the mixed clusters, the residents did not live within close proximity of migrants but the residents lived in proximity to one another, as did the migrants. For example, in mixed Cluster-03 and -09, all the residents lived close to one another in the downtown area, while the migrants lived in industrial areas elsewhere in the district. We observed similar patterns of spatial aggregation when considering genomic clusters of isolates from three individuals (Figure S6, Appendix, page 18).

We aimed to further estimate the likely directionality of transmission occurring between residents and migrants in mixed clusters of at least three individuals based on the transmission inference and the consensus transmission tree (Figure S7 and S8, Appendix, pages 19–20). After excluding the initial transmission event leading to the putative source case of each cluster observed, we estimated 74 transmission events occurring among 72 genomically clustered patients, of which 37 were captured during this study. We did not find evidence that migrants were more likely to transmit to residents than residents were to migrants. 12 (54%) of 22 estimated transmission events initiated by a migrant had a secondary case that was a resident, and 9 (60%) of 15 estimated transmission events initiated by a resident had a secondary case that was migrant ($p=0.7$).

Using this Bayesian approach to estimate the most likely time of transmission among migrants with TB, we also aimed to quantify and compare the relative importance of importation of latent or incipient TB with the importance of local transmission in Songjiang. By comparing the estimated time of transmission with reported time of arrivals in Songjiang, 78% (47/60) of migrants with sufficient data to inform this estimate were likely infected after their arrival in Songjiang (Figure S9, Appendix, page 21). This is consistent with the observation that migrants sharing genomically-clustered strains usually originated from different home provinces (Table S5, Appendix, page 12).

DISCUSSION

In this seven-year population-based study involving WGS, geospatial analysis, and epidemiological investigation of TB in Songjiang, Shanghai, we found that immigration

from rural areas has been associated with a substantial increase in local TB notifications. The increase in the local burden of TB is not primarily attributable to importation and reactivation of latent infections acquired elsewhere. Rather, our analyses suggest that local transmission in Songjiang, occurring among migrants and between residents and migrants, is the dominant underlying mechanism driving increasing case notifications in this urban setting. These findings suggest both opportunities and challenges for improving TB control in cities.

The influx of internal migrants from rural areas has been suggested as an important driver of the increasing TB burden reported in urban settings in China,²² particularly in populous cities like Beijing, Guangzhou, Shenzhen, and Shanghai, where 40–70% of the total population consists of the internal migrants.^{8,23–26} In our study, migrants comprised 75% of all culture-positive TB cases in Songjiang and experienced substantially higher crude and age-adjusted TB notification rates than residents. Administrative and socio-economic barriers,⁷ as well as challenging working, and living conditions, characterized by overcrowding and poor sanitation,²⁷ likely contribute to the greater risk of TB disease experienced by migrants.⁷ Similar observed trends of rural to urban migration and comparable fractions of disease occurring amongst migrants suggest that our findings may be generalizable to other cities in China.

By leveraging genomic, epidemiological, and spatial data, we learned that transmission of TB occurs not only among migrants, but also between residents and migrants in Songjiang. We note that traditional epidemiological investigations failed to identify most estimated transmission links, and were especially poor at identifying links between migrants and residents. We uncovered previously unsuspected links and found that genomically-clustered strains infecting both migrants and residents were as likely to have been transmitted by residents to migrants and from migrants to residents.

We inferred that most *M. tuberculosis* infections that led to disease among migrants, occurred after their arrival in Songjiang. This finding emphasizes that the increasing case notifications in Songjiang are not primarily attributable to importation of incipient TB from rural areas. From a public health perspective, this is notable because it suggests that emphasis should be placed on interventions that can interrupt TB transmission within Songjiang.

We detected a significant association between geographic proximity of TB patient's residences and the genetic relatedness of the *M. tuberculosis* strains causing their disease. While previous studies have reported that geographic aggregation of TB cases may be associated with a genotypic clustering,^{28–30} none have formally examined the relatedness of spatial proximity and the genetic distance of *M. tuberculosis* transmission. We also determined a positive association between spatial proximity and genetic relatedness for almost all of the genomic clusters involving at least three individuals (Figure 4 and S6, Appendix, page 18), and some clusters showed extremely close geographical proximity (e.g., Cluster-11 consisting of six migrant patients, all of whom lived near a single industrial park).

The identification of spatial aggregation of closely related strains among migrant-only and resident-only clusters likely reflect contacts occurring in shared local social networks.²⁷ For mixed clusters, we found more dispersed spatial patterns, as migrants and residents tend to live in distinct neighborhoods (e.g. Cluster-01, 03, 04, and 09). Transmission events between migrants and residents suggest the possibility that transmission may occur through more casual contact in areas where populations intermix, such as the downtown area that serves as a center for transportation and entertainment.

There are notable limitations to our study that are worth emphasizing. First, and most importantly, our ability to reliably infer patterns of TB transmission in Songjiang is dependent on how complete our sampling of *M. tuberculosis* isolates was during the study period. We know that not all notified TB cases had culture-positive disease, and we were not able to consider those without isolates available for genotyping and sequencing in our analyses. Furthermore, it is likely that we have missed cases occurring among migrants who returned to their home provinces for diagnosis or treatment, and the potential bias associated with these missing cases is difficult to quantify. However, because migrants had better healthcare access and resources in Shanghai than in other cities in China during our study period, we believe that the impact of missing migrant cases is minimized because we conducted this study in this setting. The completeness of sampling of *M. tuberculosis* isolates also affects the quality of our inference about transmission and estimation of most likely times of infection. Second, we are unable to infer the overall proportion of strains that were genomically-clustered (e.g. within a particular SNP distance from another strain) since WGS was restricted to VNTR clusters of size three or larger. Third, our spatial analysis was restricted to residential addresses at the time of diagnosis. Although the routine investigation included questions about frequently visited locations, this information was not uniformly collected and we did not attempt to use it in our analysis beyond noting when multiple cases in a genomic cluster reporting a shared location.

In conclusion, consideration of genomic, spatial and epidemiological data helped us to better understand the complex dynamics of TB transmission in an urban center experiencing important changes in migration. We have shown that the major driver of TB infections occurring among migrants was due to local transmission not attributable to imported disease, a finding that should motivate TB control programs to increase access to diagnosis and care and more intensive case finding activities to halt the transmission of TB among both migrants and residents.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the health workers at the Tuberculosis control department of Songjiang Center for Disease Control and Prevention for their support. We thank Dr. Leonid Chindelevitch from Simon Fraser University, Canada for his valuable suggestion for genotypic analysis.

Funding

This study was funded by National Science and Technology Major Project of China [2017ZX10201302 and 2018ZX10715012 to Q.G.], Natural Science Foundation of China [81402727 to C.Y.], National Institutes of Health (NIH) [R01 AI112438-01 to T.C.], and MIDAS Center for Communicable Disease Dynamics [U54 GM088558 to T.C. and C.Y.]; this work was also supported by Natural Science Foundation of China [91631301 and 81661128043 to Q.G.] and Sanming project of Medicine in Shenzhen [SZSM201611030 to Q.G.]; the International Postdoctoral Fellowship Program from China Postdoctoral Council [20150058 to C.Y.]; CTSA Grant from the National Center for Advancing Translational Science [UL1 TR001863 and KL2 TR001862 to J.L.W.] and NIH grants [DP2OD006452 to K.D. and R25 TW009343 to C.Y.].

References

1. WHO. Global tuberculosis report 2017. World Health Organization; 2017. http://www.who.int/tb/publications/global_report/en/
2. Wang L, Liu J, Chin DP. Progress in tuberculosis control and the evolving public-health system in China. *Lancet*. 2007; 369(9562):691–6. [PubMed: 17321314]
3. Tobe RG, Xu L, Song P, Huang Y. The rural-to-urban migrant population in China: gloomy prospects for tuberculosis control. *Biosci Trends*. 2011; 5(6):226–30. [PubMed: 22281535]
4. Peng X. China's demographic history and future challenges. *Science*. 2011; 333(6042):581–7. [PubMed: 21798939]
5. Dhavan P, Dias HM, Creswell J, Weil D. An overview of tuberculosis and migration. *Int J Tuberc Lung Dis*. 2017; 21(6):610–23. [PubMed: 28482955]
6. Wei X, Chen J, Chen P, et al. Barriers to TB care for rural-to-urban migrant TB patients in Shanghai: a qualitative study. *Trop Med Int Health*. 2009; 14(7):754–60. [PubMed: 19392747]
7. Hu X, Cook S, Salazar MA. Internal migration and health in China. *Lancet*. 2008; 372(9651):1717–9. [PubMed: 18930533]
8. Jia ZW, Jia XW, Liu YX, et al. Spatial analysis of tuberculosis cases in migrants and permanent residents, Beijing, 2000–2006. *Emerg Infect Dis*. 2008; 14(9):1413–9. [PubMed: 18760008]
9. Zhou C, Tobe RG, Chu J, Gen H, Wang X, Xu L. Detection delay of pulmonary tuberculosis patients among migrants in China: a cross-sectional study. *Int J Tuberc Lung Dis*. 2012; 16(12):1630–6. [PubMed: 23131261]
10. Wang W, Jiang Q, Abdullah AS, Xu B. Barriers in accessing to tuberculosis care among non-residents in Shanghai: a descriptive study of delays in diagnosis. *Eur J Public Health*. 2007; 17(5):419–23. [PubMed: 17412714]
11. Tobe RG, Xu L, Zhou C, Yuan Q, Geng H, Wang X. Factors affecting patient delay of diagnosis and completion of Direct Observation Therapy, Short-course (DOTS) among the migrant population in Shandong, China. *Biosci Trends*. 2013; 7(3):122–8. [PubMed: 23836035]
12. Shen X, Xia Z, Li X, et al. Tuberculosis in an urban area in China: differences between urban migrants and local residents. *PLoS One*. 2012; 7(11):e51133. [PubMed: 23226479]
13. Yang C, Luo T, Sun G, et al. *Mycobacterium tuberculosis* Beijing strains favor transmission but not drug resistance in China. *Clin Infect Dis*. 2012; 55(9):1179–87. [PubMed: 22865872]
14. Yang C, Shen X, Peng Y, et al. Transmission of *Mycobacterium tuberculosis* in China: a population-based molecular epidemiologic study. *Clin Infect Dis*. 2015; 61(2):219–27. [PubMed: 25829000]
15. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLoS Med*. 2016; 13(10):e1002137. [PubMed: 27701423]
16. Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F. Whole genome sequencing of *Mycobacterium tuberculosis* for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis (Edinb)*. 2016; 98:77–85. [PubMed: 27156621]
17. Ribeiro FK, Pan W, Bertolde A, et al. Genotypic and Spatial Analysis of *Mycobacterium tuberculosis* Transmission in a High-Incidence Urban Setting. *Clin Infect Dis*. 2015; 61(5):758–66. [PubMed: 25948063]
18. Yang C, Luo T, Shen X, et al. Transmission of multidrug-resistant *Mycobacterium tuberculosis* in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *Lancet Infect Dis*. 2017; 17(3):275–84. [PubMed: 27919643]

19. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol.* 2014; 31(7):1869–79. [PubMed: 24714079]
20. National Bureau of Statistics of China. The 2010 Population Census of The People's Republic of China. China Statistics Press; Apr, 2013. <http://www.stats.gov.cn/english/statisticaldata/censusdata/>
21. Luo T, Yang C, Peng Y, et al. Whole-genome sequencing to detect recent transmission of *Mycobacterium tuberculosis* in settings with a high burden of tuberculosis. *Tuberculosis (Edinb).* 2014; 94(4):434–40. [PubMed: 24888866]
22. Sun YX, Zhu L, Lu ZH, Jia ZW. Notification Rate of Tuberculosis among Migrants in China 2005–2014: A Systematic Review and Meta-analysis. *Chinese Medical Journal.* 2016; 129(15): 1856–60. [PubMed: 27453237]
23. Zhang LX, Tu DH, An YS, Enarson DA. The impact of migrants on the epidemiology of tuberculosis in Beijing, China. *Int J Tuberc Lung Dis.* 2006; 10(9):959–62. [PubMed: 16964784]
24. Li X, Yang Q, Feng B, et al. Tuberculosis infection in rural labor migrants in Shenzhen, China: Emerging challenge to tuberculosis control during urbanization. *Sci Rep.* 2017; 7(1):4457. [PubMed: 28667275]
25. Zeng J, Shi L, Zou X, Chen W, Ling L. Rural-to-Urban Migrants' Experiences with Primary Care under Different Types of Medical Institutions in Guangzhou, China. *PLoS One.* 2015; 10(10):e0140922. [PubMed: 26474161]
26. Chen J, Qi L, Xia Z, et al. Which urban migrants default from tuberculosis treatment in Shanghai, China? *PLoS One.* 2013; 8(11):e81351. [PubMed: 24312292]
27. Shen J, Huang Y. The working and living space of the 'floating population' in China. *Asia Pac Viewpoint.* 2003; 44(1):11.
28. Smith CM, Maguire H, Anderson C, Macdonald N, Hayward AC. Multiple large clusters of tuberculosis in London: a cross-sectional analysis of molecular and spatial data. *ERJ Open Res.* 2017; 3(1)
29. 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): 287–94. [PubMed: 26175455]
30. Bishai WR, Graham NM, Harrington S, et al. Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. *JAMA.* 1998; 280(19):1679–84. [PubMed: 9831999]

Figure 1. Study flowchart (A) and the distribution of provinces of origin for migrant TB patients in Songjiang (B), Shanghai, 2009–2015

TB, tuberculosis; VNTR, variable number of tandem repeats; WGS, whole-genome sequencing; The size of circle in panel B represents the number of culture positive tuberculosis cases, stratified by Beijing strains (orange) and non-Beijing strains (blue).

Figure 2. The distribution of pair-wise genetic distances (number of SNPs) within VNTR-based clusters

The color of the bar represents the cluster type: blue, resident-only clusters; green, migrant-only clusters; pink, mixed clusters. The vertical dashed lines indicate a ten SNP threshold.

Figure 3. Spatial distribution of notification rate of culture-positive tuberculosis. (A) Overall TB notification rate by sub-district and kernel density estimation (KDE) for migrant TB (B) and resident TB (C) patients in Songjiang, Shanghai, 2009–2015.

Figure 4. Genetic distance and geographic distribution of genomic clusters

(A) Clusters with at least four patients are presented with genetic distances and epidemiologic links. Each circle (migrant) or triangle (resident) represents an individual within a genomic cluster. Individuals with epidemiological links in a cluster are shown with pink outlines and annotated. *M. tuberculosis* isolates are separated by lines with length (as suggested by the dots) representing genetic distance. Arrows indicate the next closest isolate in the sequenced collection. (B) Spatial distribution of clusters presented in panel A. Orange, migrant patients; Blue, resident patients.

Table 1

Characteristics of migrant and resident patients in Songjiang, Shanghai, 2009–2015

Characteristics	Migrant patients (n=1211)	(%)	Resident patients (n=409)	(%)	P value
Female	441	(36.4)	87	(21.2)	< 0.001
Age, median years (IQR)	27	(22–38)	55	(37–71)	< 0.001
15–24	471	(38.9)	48	(11.7)	Reference
25–44	567	(46.8)	83	(20.3)	0.05
45–64	150	(12.4)	130	(31.8)	< 0.001
65+	23	(1.9)	148	(36.2)	< 0.001
Occupations					
Commercial service	245	(20.2)	70	(17.1)	Reference
Labor worker	755	(62.4)	85	(20.8)	< 0.001
Farmers	37	(3.1)	115	(28.1)	< 0.001
Students	45	(3.7)	27	(6.6)	< 0.01
Retirement	10	(0.8)	62	(15.2)	< 0.001
Others	119	(9.8)	50	(12.2)	0.08
Case detection					
Referral	610	(50.4)	160	(39.1)	Reference
Follow-up after referral	347	(28.7)	111	(27.2)	0.15
Clinical consultation (TB-designated)	166	(13.7)	104	(25.4)	< 0.0001
Physical examination	84	(3.9)	31	(7.6)	0.13
Others	4	(0.3)	3	(0.7)	0.15
Previously treated for TB	87	(7.2)	54	(13.2)	< 0.001
Median diagnostic delay, days (IQR)	26	(15–47)	30	(18–46)	0.45
Cavitary	509	(42.0)	180	(44.0)	0.48
Treatment outcomes					
Cured/completion	1006	(83.1)	317	(77.5)	Reference
Failure	9	(0.7)	4	(1.0)	0.56
Death	8	(0.7)	18	(4.4)	< 0.001
Lost to follow-up	7	(0.6)	0	(0)	0.36
On treatment	156	(12.9)	56	(13.7)	0.43

Characteristics	Migrant patients (n=1211)	Resident patients (n=409)	P value
Others	25 (2.0)	14 (3.4)	0.08
Laboratory testing			
Beijing strains	928 (76.6)	359 (87.8)	< 0.001
Sputum smear positive	762 (63.0)	264 (64.7)	0.53
Drug susceptible	1072 (88.5)	356 (87.0)	Reference
Mono-drug resistance	92 (7.6)	36 (8.8)	0.42
MDR-TB	47 (3.9)	17 (4.2)	0.81

Abbreviation: TB, tuberculosis; IQR, the interquartile range; MDR, multidrug resistance, resistant to at least isoniazid and rifampin.

Table 2
Summary of logistic regression model of the association between spatial distance and genetic relatedness.

Genetic distance of a genomic cluster, SNPs	Geographic distance between homes (per each additional kilometer)		P value	AIC
	Odds ratio*	95%CI		
3	0.90	0.87–0.93	<0.0001	1516
5	0.89	0.87–0.92	<0.0001	1792
10	0.89	0.87–0.91	<0.0001	2020
50	0.90	0.87–0.92	<0.0001	2189
100	0.93	0.91–0.95	<0.0001	3188

* Adjusted for covariates including household registration status (migrant or resident) and genetic sub lineages; AIC, Akaike information criterion.