HHS Public Access
Author manuscript
Tuberculosis (Edinb). Author manuscript; available in PMC 2015 July 01.

Published in final edited form as:
Tuberculosis (Edinb). 2014 July ; 94(4): 434–440. doi:10.1016/j.tube.2014.04.005.

Whole-genome sequencing to detect recent transmission of *Mycobacterium tuberculosis* in settings with a high burden of tuberculosis

Tao Luo[1,*], **Chongguang Yang**[1,*], **Ying Peng**[2], **Liping Lu**[3], **Guomei Sun**[1], **Jie Wu**[4], **Xiaoping Jin**[3], **Jianjun Hong**[3], **Fabin Li**[2], **Jian Mei**[4], **Kathryn DeRiemer**[5,#], and **Qian Gao**[1,#]

Tao Luo: taoluo.fdu@gmail.com; Chongguang Yang: yangstoptb@gmail.com; Ying Peng: Lily_pengj@126.com; Liping Lu: luluyer-1194@163.com; Guomei Sun: sunguomei@gmail.com; Jie Wu: wu790429@sina.com; Xiaoping Jin: jxp654321@163.com; Jianjun Hong: hjj987654321@163.com; Fabin Li: hljlfb@163.com; Jian Mei: meijiansh@yahoo.com.cn; Kathryn DeRiemer: kderiemer@ucdavis.edu

[1]Key Laboratory of Medical Molecular Virology of Ministries of Education and Health, Institutes of Biomedical Sciences and Institute of Medical Microbiology, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China

[2]Tuberculosis (TB) Control Center of Heilongjiang Province, No. 40, Youfang Street, Harbin, Heilongjiang, 150030, China

[3]Department of TB Control, Songjiang District of Shanghai Municipal Center for Disease Control and Prevention, 1050 North Xi Lin Road, Shanghai 201620, China

[4]Department of TB Control, Shanghai Municipal Centers for Disease Control and Prevention, Shanghai, 200336, China

[5]University of California, Davis, School of Medicine, One Shields Avenue, Davis, California, 95616, USA

Abstract

Whole genome sequencing (WGS) of *Mycobacterium tuberculosis* has been used to trace the transmission of *Mycobacterium tuberculosis*, the causative agent of tuberculosis (TB). Previously published studies using WGS were conducted in developed countries with a low TB burden. We sought to evaluate the relative usefulness of traditional VNTR and SNP typing methods, WGS and epidemiological investigations to study the recent transmission of *M. tuberculosis* in a high TB burden country. We conducted epidemiological investigations of 42 TB patients whose *M. tuberculosis* isolates were classified into three clusters based on variable-number tandem repeat (VNTR) typing. We applied WGS to 32 (76.2%) of the 42 strains and calculated the pairwise genomic distances between strains within each cluster. Eighteen (56.3%) of the 32 strains had genomic differences ≥100 SNPs with every other strain, suggesting that direct transmission did not likely occurred. Ten strains were grouped into four WGS-based clusters with genomic

Correspondence: Qian Gao, Ph.D., Shanghai Medical College, Fudan University, 138 Yi Xue Yuan Road, Shanghai, China, 200032, Tel: (8621) 5423-7195, Fax: (8621) 5423-7971, qgao99@yahoo.com.
*These authors contributed equally for this work.
#These authors contributed equally for directing this work.

Conflict of interest statement
All authors declared no conflict of interest.

distances ≤5 SNPs within each cluster, and confirmed epidemiological links were identified in two of these clusters. Our results indicate that WGS provides reliable resolution for tracing the transmission of *M. tuberculosis* in high TB burden settings. The high resolution of WGS is particularly useful to confirm or exclude the possibility of direct transmission events defined by traditional typing methods.

Keywords

Tuberculosis; Transmission; Whole genome sequencing; Genotyping; Contact tracing

Introduction

Molecular typing enhanced our understanding of the epidemiology of tuberculosis (TB) during the past two decades. Three main typing methods, specifically *IS6110* restriction fragment length polymorphism (RFLP), spoligotyping, and variable-number tandem repeat (VNTR) analysis, are currently used for fingerprinting *Mycobacterium tuberculosis* strains to detect recent transmission.[1] Due to the intrinsic defects of these methods, such as their limited discriminatory power and homoplasy, the clustered strains defined by these methods may be both genetically and historically distantly related, and thus epidemiological investigations are usually needed to confirm recent transmission.[1,2] However, epidemiological investigations are time consuming, labor-intensive, and cannot consistently identify epidemiological links between cases with a missing source case or between TB cases that occur by short-term or casual contacts.[3]

Recent developments in high-throughput whole genome sequencing (WGS) provide a powerful tool for studying the epidemiology of TB. The rate of change of *M. tuberculosis* was estimated as approximately 0.3–0.5 mutations per genome per year during its life cycle within the human host.[4–6] A threshold of five or fewer single nucleotide polymorphisms (SNPs) was suggested as the potential standard to define strains involved in chains of recent transmission within three years.[6] Compared to traditional typing methods, WGS provides higher resolution to investigate TB outbreaks by differentiating strains with identical VNTR and/or *IS6110* RFLP genotypes into smaller, more accurate clusters.[5,7] Since recombination and reverse mutation are rare in *M. tuberculosis*, the evolution of bacterial genome mostly represents the step-wise accumulation of mutations.[8,9] Therefore, phylogenetic networks of *M. tuberculosis* strains based on genomic mutations can be used to identify putative source cases, super-spreaders and transmissions patterns in the absence of extensive epidemiological data.[5,6,10]

However, since all previously published WGS-based epidemiological studies were conducted in developed countries with a low TB burden, the usefulness of WGS for tracing recent transmission in high TB burden settings remains unknown. In low TB burden regions, transmission of *M. tuberculosis* has been largely prevented by efficient control programs in the recent decades. Currently, most *M. tuberculosis* strains collected during a short time period (e.g., two years) in low TB burden regions are both historically and genetically distantly related. By contrast, the extensive transmission of *M. tuberculosis* in recent

decades has promoted the prevalence of a large number of genetically closely related strains in high TB burden regions, with an example of *M. tuberculosis* Beijing strains.[11] The homogeneity of Beijing strains has led to the low discriminatory power of VNTR typing in East Asia and South Africa.[12–14] In these settings, many strains with identical VNTR or *IS6110* RFLP genotypes were genetically distantly related and could be unambiguously differentiated by WGS. For example, two Beijing strains with identical *IS6110* RFLP pattern differed by as many as 130 SNPs by WGS in a high incidence area in Uzbekistan, excluding the possibility of recent transmission.[15] However, a large number of strains circulating in these areas were genetically extremely similar, and even WGS could not differentiate recent transmissions and reactivations caused by these strains. In a recent study from Russia, a number of *M. tuberculosis* Beijing strains with identical or very similar genomes were isolated from patients separated by large geographical distances, which were less likely to be epidemiologically linked.[16] Although Russia is not a high TB burden country,[17] it experienced a tuberculosis epidemic at the end of 20th century due to the disintegration of Soviet Union,[18] which gave rise to homogenous *M. tuberculosis* populations similar to those in the high TB burden areas.

Here, we applied WGS and epidemiological investigations to several *M. tuberculosis* clusters of Beijing strains defined by VNTR and SNP typing in two regions of China. We sought to evaluate the relative usefulness of VNTR typing, WGS and epidemiological investigations to study recent transmission of *M. tuberculosis* in high TB burden areas.

Methods

Bacterial samples and genotyping

A population-based molecular epidemiology study in Songjiang District, Shanghai, and Wuchang County, Heilongjiang province, was conducted from 1 June 2009 to 31 December 2010.[19] There were 396 and 184 strains collected in Songjiang District and Wuchang County, respectively, from TB patients that were culture positive for *M. tuberculosis*. All of the mycobacterial strains from the TB patients were typed with a 16-locus high-resolution VNTR (VNTR-16) set,[20] and 6 SNP sites (SNP-6) of 3R (DNA replication, repair and recombination) genes in our previous study.[19] Strains with an identical VNTR-16 and SNP-6 genotype were defined as a cluster and were assumed to represent recent transmission events. A total of 42 isolates of three large clusters from Wuchang County and Songjiang District were selected to recover with Lowenstein-Jensen (L-J) slants. For each recovered isolate, all colonies that grew on the L-J slants were scraped for DNA extraction. The genomic DNAs were extracted following the cetyltrimethyl-ammonium bromide-lysozyme (CTAB) method.[21] The VNTR-16 and SNP-6 genotypes of the recovered isolates were confirmed with the genomic DNAs according to the methods of our previous study.[19] Additionally, two hypervariable loci, VNTR 3232 and VNTR 4120, were typed in all isolates with primers and conditions previously described.[14]

Whole genome sequencing, SNP calling and phylogenetic analysis

The genomic DNAs of recovered strains were sent to Chinese National Human Genome Center (Shanghai, China) for whole genome sequencing. A 300-base-pair (bp) paired-end

library was constructed for each purified DNA sample. Sequencing was performed on the Illumina Hiseq 2000 with 100 or 115 cycles, with an expected coverage of 100. The sequencing data (FASTQ format) were deposited in the National Center for Biotechnology Information Sequence Read Archive (Accession No. **SRP029424**).

The low-quality bases at the 3′ ends were first trimmed using Sickle (https://github.com/ucdavis-bioinformatics/sickle) and were then mapped to the reference genome H37Rv (GenBank: AL123456) with Bowtie using non-gapped alignments.[22] Reads that mapped to more than one genomic position were discarded. The SAMtools/BCFtools suite was used for SNP calling.[23] Heterozygous calls and SNPs with coverage lower than three were filtered. SNPs in the PE/PPE, PE-PGRS and drug-resistance associated genes were also filtered,[24] using an in-house Perl script. SNP lists for individual strains were combined into a single non-redundant list, and corresponding base calls were recovered for each strain. The concatenated sequences of the 32 strains were used to generate a Maximum Likelihood (ML) phylogeny by MEGA5.[25] The sequences were also used to generate Median-joining (MJ) networks for each cluster with NETWORK (www.fluxus-engineering.com).

Strains with genomic difference(s) of ≤ 5 SNPs were defined as a WGS-based cluster according to previous study.[6] For SNPs within each WGS-based cluster, we further checked their homozygous/heterozygous status in the clustered strains. As *M. tuberculosis* is strictly clonal, a heterozygous call with both wild type and mutant alleles most likely indicates the origin of the mutant allele in the corresponding patient. We used the program LoFreq to determine the frequencies of different alleles for each position and to evaluate the reliability of the calls.[26] We kept only the calls in which the coverage was ≥10 and the less frequent allele was supported by at least five high-quality reads, as reliable calls.

Epidemiological investigation

Traditional epidemiological investigations were used to identify the persons, places, and behaviors that may contribute to the transmission of *M. tuberculosis* between TB patients within each cluster. TB patients were interviewed in their household, or in the local CDC facility if patients were not willing to be interviewed in their home. A social-network questionnaire was developed to identify shared social settings and to prioritize contact tracing. The questionnaire included information on the individual's place of residence, travel history, places of social aggregation, and identification of contacts with risk factors that might be associated with TB (for example, smoking or alcohol abuse) and specific locations. All interviews were performed retrospectively by trained interviewers. The study protocol was approved by the Ethics Committees of the Institutes of Biomedical Sciences in Fudan University. Epidemiological networks were drawn manually for visualization of the links identified by epidemiological investigation in each VNTR-16 based cluster. In each network, lines were drawn between geographically close patients (e.g., who lived in the same village or adjacent villages, or on the same street), family members or neighbors, and shared enclosed or open-air locations.

Results

Characteristics of the clustered cases

Based on the VNTR and SNP typing that was done in our previous study,[19] there were two large clusters (cluster A and B) with 16 and 7 cases, respectively, in Wuchang County during 1 June 2009 to 31 December 2010. The genotypes of clusters A and B were also identified in Songjiang District, and accounted for seven and five TB cases, respectively. In Songjiang District, there was another large cluster (cluster C) with five TB cases. The A, B and C genotypes accounted for the three largest clusters in Songjiang District. In total, the three clusters accounted for 42 TB cases, with 23 TB cases in Wuchang County and 19 TB cases in Songjiang District (Table 1). All clustered TB patients were adults ≥ 18 years old. The majority of the TB patients (28/42, 66.7%) in clusters were male, and most of them (38/42, 90.5%) were new TB cases. Among the 19 TB patients from Songjiang District, 12 (63.6%) were domestic migrants from other provinces of China.

VNTR and genomic similarities of the isolates

All 23 isolates from Wuchang County and 9 of the 19 strains from Songjiang District were successfully recovered. The VNTR-16 genotypes of all 32 recovered strains were confirmed and their genomes were sequenced (Table 1). The VNTR-16 genotypes of the three clusters were similar to each other (Figure 1). The genotypes of cluster A and B, which were identified both in Wuchang County and Songjiang District, differed in only one VNTR locus, QUB-11a. The genotype of cluster C in Songjiang District also had one locus (MIRU-10) that was different from the genotype of cluster A. By SNP typing, Clusters A and C belonged to Beijing family sublineage Bmyc10 and Cluster B belonged to Beijing family Bmyc210. Both sublineages belonged to the evolutionarily “modern” branch of Beijing family strains.[14,20]

The WGS-based analysis detected high genomic diversity for clusters A, B and C at both the intra- and inter-cluster level (Figure 1). Of the 32 strains with WGS results, 18 (56.3%) strains were different from every other strain by more than 100 SNPs. As the mutation rate of *M. tuberculosis* is as low as 0.3–0.5 SNPs per genome per year,[4–6] the large genetic differences observed here indicate that the strains are historically remotely related, rather than caused by recent transmission. The remaining 14 strains were divided into six groups with small or intermediate genomic variations within each group. Four groups of strains, three from cluster A and one from cluster C, had small genomic distances of ≤5 SNPs. Strains of these four groups were defined as WGS-based clusters as they may indicate recent transmissions according to previous study.[6] The other two groups of strains, which belong to cluster B, had intermediate genomics distances of 16 and 24 SNPs, respectively. Strains with identical genotypes of VNTR-16 and SNP-6 but from TB patients in different provinces were all genetically distantly related.

Concordance between WGS and hypervariable VNTR loci

Several previous studies demonstrated the importance of using hypervariable loci (VNTR 3232, VNTR 3820 and VNTR 4120) to enhance the resolution of VNTR typing in settings where Beijing family strains are prevalent.[14,27] To evaluate the concordance between VNTR

hypervariable loci and WGS data, we further typed VNTR 3232 and VNTR 4120 for all 32 isolates. All of the strains that were within each of the four WGS-based clusters shared identical profiles in these loci (Figure 1). For the two groups of strains within cluster B that had intermediate genomic distances, strains of one group showed identical genotypes in both loci, and strains of the other group showed a one-repeat difference at locus VNTR 3232 (Figure 1). Nevertheless, the results of the hypervariable loci typing and WGS were discordant. There were seven strains in cluster A that had identical alleles of 10 and 13 in these two hypervariable loci. By WGS, these strains were further discriminated into two distinct clusters and two distantly related singletons. In cluster B, there were five isolates grouped into two clusters based on the profiles of the hypervariable VNTR loci. According to WGS, two of the three strains that had identical alleles of 10 and 13 were relatively closely related and the other strain was distantly related with them. The two strains with identical alleles of 8 and 14 were distantly related.

Detection of recent transmission based on WGS and epidemiological investigation

No epidemiological link was identified between the 19 cases from Songjiang District. A total of 24 epidemiological links were identified between the 23 cases from Wuchang County. Epidemiological networks were drawn manually for visualization of these links in each VNTR-16 based cluster (Figure 2). Confirmed epidemiological links were identified in two of the four WGS-based clusters. Patients with strains H100135 and H100186 were family members, and their isolates had identical genomes, including mutations in the genes associated with drug resistance (supplementary table 1). The two patients with strains H100182 and H100200 were neighbors, and lived in the same village as the patient with strain H100110. The onset of symptoms in the patient with isolate H100182 was in September 2009, but TB disease was not diagnosed until 12 months later. Therefore, it is likely that the patient with isolate H100200 was directly infected by the patient with isolate H100182. Patients with isolates H100200 and H100110 were also indirectly linked, as they used go to the same street market in Wuchang City. No epidemiological links were found among patients with isolates H090071, H100090 and H090004. Based on WGS, strain H100090 acquired one additional SNP (45025, T→C) compared to the other two strains (Figure 1 and Figure 3). By checking the homozygous/heterozygous status of this SNP in all three isolates, the mutant allele C was also detected in strain H090071 at a frequency of 10.6% (95% CI, 5.6%–18.7%). It is likely that the mutation arose in the mycobacteria in the patient with isolate H090071 and the mutant clone was later transmitted to the patient with isolate H100090. For the two groups of strains with intermediate genetic distances in cluster B, the patients with isolates H100079 and H090009 lived on the same street, and no epidemiological link was identified between the patients with isolates H100198 and H100155.

There were six epidemiological links (two confirmed links and four possible links) that were identified between TB patients whose strains differed by small (≤5 SNPs) or intermediate genomic distances (16 or 24 SNPs). The remaining 18 epidemiological links were identified between TB patients whose strains differed by a large genomic distance (>100 SNPs). Three patients (with isolates H090041, H100200 and H100186) of cluster A used to go to the same Internet cafe. Another three patients of cluster A (with isolates H100182, H090041 and

H100119) used go to the same underground plaza for shopping or entertainment. Three patients (with strain H100200, H100110 and H090029) used go to the same street market in Wuchang City. In cluster B, patients of strains H100176 and H100146 lived on the same street, and the patients of H100079 and H100176 were familiar to each other and their work places were located on the same street. Patient of H100145 and H100155 used go to the same street market of their town. Finally, some patients of cluster A or B lived in the same villages or in adjacent villages (Figure 2).

Discussion

Our work evaluated the relative usefulness of traditional VNTR and SNP typing methods, WGS, and epidemiological investigations in settings with a high burden of TB. We demonstrated that the high resolution of WGS was very helpful to exclude transmission events that were identified by traditional VNTR typing. In addition, WGS provided other useful genetic information for defining direct transmission events, such as drug resistance-conferring mutations and the homozygous/heterozygous status of SNPs. WGS was more accurate than VNTR typing to identify likely transmission events in high TB burden settings.

Our results highlight the limitations of using VNTR typing and epidemiological investigations to detect transmission of *M. tuberculosis* in a high burden setting. In this study, the three clusters defined by VNTR-16 typing were divided into distantly related singletons or small clusters by WGS. In contrast, in low TB burden countries, most of the clustered strains that were identified by traditional genotyping methods were genetically closely related by WGS.[4,6] Epidemiological investigations are necessary and useful to identify putative transmission events between individual TB cases in VNTR-based clusters. However, most of the epidemiological links in the current study were between individuals whose isolates were distantly related, based on WGS. Epidemiological investigation can be difficult to conduct in certain populations, such as among the domestic migrants in Songjiang District in this study. Some of the migrants refused to be interviewed, and others had returned to their hometowns for treatment after their diagnosis and prior to their interview.

Our results indicate that a small genomic distance is necessary but not sufficient for defining recent transmission of *M. tuberculosis* in the studied settings. In a previous study, a genomic difference of ≤5 SNPs was suggested as the potential cut-off for defining strains of recent transmissions within three years in a low burden setting.[6] In our study, the genomic distances between strains with robust epidemiological links were all within five SNPs, suggesting this threshold is reliable in our setting. However, several TB patients whose strains differed by ≤5 SNPs lacked epidemiological links, which may indicate casual transmission or missing source cases. Considering the homogeneity of *M. tuberculosis* in our setting, it is also possible that these cases were resulted from reactivations that caused by remote infection of genetically closely related *M. tuberculosis* strains. According to previous studies, the number of mutations between epidemiologically linked cases could be unexpectedly high due to host factors and/or environmental factors.[4,8] In a recent study, as many as 14 SNPs were identified between strains isolated from patients with confirmed

epidemiological links.[28] Therefore, we cannot exclude the possibility of direct transmission cases that differ by 16 or 24 SNPs in this study.

WGS provides additional genetic information for exploring the details of transmission events. First, WGS provides the mutation profiles of all drug resistance-associated genes, which is important to detect the transmission of resistant strain or the acquisition of resistance. Second, the high depth of WGS enables one to detect the homozygous/heterozygous status of SNPs for strains within a transmission chain, providing valuable information to determine the direction of transmission events. Finally, the topology of the WGS-based phylogeny, such as the MJ network, can help to identify the source case, super spreaders and underlying transmission patterns.[5,6] In a previous study, a postulated node in the MJ network was usually treated as a missing case due to incomplete taxon sampling.[6] In this study, a postulated node was found between two cases (with isolates H100182 and H10020) with a confirmed relationship of direct transmission (Figure 3). Considering the diagnostic delay in the source case (with isolate H100182), new variants of *M. tuberculosis* may have evolved in the patient after transmission. Thus, the postulated node may represent the genotype of strains that were circulating when transmission occurred, rather than a missing case. Since diagnostic delays of TB are common in both high- and low-burden settings,[29,30] the proper interpretations of postulated nodes are important for defining transmission events using WGS.

Considering the ever-decreasing costs of WGS and the development of automatated data analysis tools, WGS is under consideration for routine TB public health practice in some developed countries.[6] Although it is not currently possible to conduct large-scale, high-throughput WGS-based programs, an appropriate combination of traditional typing methods and WGS could be cost-effective in those settings. Based on the present study, we recommend a set of robust, discriminatory VNTR loci, e.g., the standard 15-/24-locus set,[31] or the 9-locus set proposed by our lab more recently,[32] as a first-line typing method for large-scale typing that can be performed as soon as mycobacterial DNA is available. The hypervariable loci, such as VNTR 3820, VNTR 3232 and VNTR 4120, could be used for second-line typing of clustered strains, based on the first-line typing results.[32,33] WGS could be used as the third-line method for strains with clustered genotypes, based on the typing of hypervariable loci. Finally, targeted epidemiological investigations would be conducted only for WGS-based clustered cases. This algorithm would minimize the time and costs of WGS and epidemiological investigations. We believe WGS will be a useful tool for understanding TB epidemiology and improving public health practice in high burden settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Key Project of Chinese National Programs, China (2013ZX10003004-001, 2013ZX10004903-006) and the National Natural Science Foundation of China (91231115). This work was also partially supported by the National Institutes of Health (USA) through the Fogarty International Center (D43TW007887) and the National Institute of General Medical Sciences (DP2 OD006452). The sponsors had no other special involvement in this study.

References

1. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev. 2006; 19:658–685.10.1128/CMR.00061-05 [PubMed: 17041139]
2. Portero JL, Rubio M. Molecular epidemiology of tuberculosis. N Engl J Med. 2003; 349:2364. [PubMed: 14674044]
3. Cronin WA, Golub JE, Lathan MJ, Mukasa LN, Hooper N, Razeq JH, Baruch NG, Mulcahy D, Benjamin WH, Magder LS, Strickland GT, Bishai WR. Molecular epidemiology of tuberculosis in a low- to moderate-incidence state: are contact investigations enough? Emerg Infect Dis. 2002; 8:1271–1279.10.3201/eid0811.020261 [PubMed: 12453355]
4. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SA, Siezen RJ, Borgdorff M, Bentley SD, Parkhill J, van Soolingen D. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC infectious diseases. 2013; 13:110.10.1186/1471-2334-13-110 [PubMed: 23446317]
5. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rusch-Gerdes S, Supply P, Kalinowski J, Niemann S. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. PLoS medicine. 2013; 10:e1001387.10.1371/journal.pmed.1001387 [PubMed: 23424287]
6. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis. 2013; 13:137–146.10.1016/S1473-3099(12)70277-3 [PubMed: 23158499]
7. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011; 364:730–739.10.1056/NEJMoa1003176 [PubMed: 21345102]
8. Schürch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, van Soolingen D. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol. 2010; 48:3403–3406.10.1128/JCM.00370-10 [PubMed: 20592143]
9. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol. 2006; 4:670–681.10.1038/nrmicro1472 [PubMed: 16912712]
10. Walker TM, Monk P, Smith EG, Peto TE. Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2013; 19:796–802.10.1111/1469-0691.12183
11. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. J Clin Microbiol. 1995; 33:3234–3238. [PubMed: 8586708]
12. Hanekom M, van der Spuy GD, Gey van Pittius NC, McEvoy CR, Hoek KG, Ndabambi SL, Jordaan AM, Victor TC, van Helden PD, Warren RM. Discordance between mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing and *IS6110* restriction fragment length polymorphism genotyping for analysis of *Mycobacterium tuberculosis* Beijing strains in a setting of high incidence of tuberculosis. J Clin Microbiol. 2008; 46:3338–3345.10.1128/JCM.00770-08 [PubMed: 18716230]
13. Mokrousov I, Narvskaya O, Vyazovaya A, Millet J, Otten T, Vishnevsky B, Rastogi N. *Mycobacterium tuberculosis* Beijing genotype in Russia: in search of informative variable-number tandem-repeat loci. J Clin Microbiol. 2008; 46:3576–3584.10.1128/JCM.00414-08 [PubMed: 18753356]
14. Luo T, Yang C, Gagneux S, Gicquel B, Mei J, Gao Q. Combination of single nucleotide polymorphism and variable-number tandem repeats for genotyping a homogenous population of

Mycobacterium tuberculosis Beijing strains in China. J Clin Microbiol. 2012; 50:633–639.10.1128/JCM.05539-11 [PubMed: 22205801]

15. Niemann S, Koser CU, Gagneux S, Plinke C, Homolka S, Bignell H, Carter RJ, Cheetham RK, Cox A, Gormley NA, Kokko-Gonzales P, Murray LJ, Rigatti R, Smith VP, Arends FP, Cox HS, Smith G, Archer JA. Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. PLoS One. 2009; 4:e7407.10.1371/journal.pone.0007407 [PubMed: 19823582]
16. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniewski F. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet. 2014; 46:279–286.10.1038/ng.2878 [PubMed: 24464101]
17. World Health Organization. Global tuberculosis report 2013. WHO/HTM/TB/2013.11; http://www.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf?ua=1
18. Toungoussova OS, Bjune G, Caugant DA. Epidemic of tuberculosis in the former Soviet Union: social and biological reasons. Tuberculosis (Edinb). 2006; 86:1–10.10.1016/j.tube.2005.04.001 [PubMed: 16256436]
19. Yang C, Luo T, Sun G, Qiao K, Sun G, DeRiemer K, Mei J, Gao Q. *Mycobacterium tuberculosis* Beijing strains favor transmission but not drug resistance in China. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2012; 55:1179–1187.10.1093/cid/cis670 [PubMed: 22865872]
20. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, Namouchi A, Jackson C, Rauzier J, Bifani P, Warren R, Rasolofo V, Mei J, Gao Q, Gicquel B. Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One. 2011; 6:e16020.10.1371/journal.pone.0016020 [PubMed: 21283803]
21. Larsen MH, Biermann K, Tandberg S, Hsu T, Jacobs WR Jr. Genetic Manipulation of *Mycobacterium tuberculosis*. Current protocols in microbiology. 2007; Chapter 10(Unit 10A): 12.10.1002/9780471729259.mc10a02s6
22. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009; 10:R25.10.1186/gb-2009-10-3-r25 [PubMed: 19261174]
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079.10.1093/bioinformatics/btp352 [PubMed: 19505943]
24. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med. 2009; 6:e2.10.1371/journal.pmed.1000002 [PubMed: 19209951]
25. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011; 28:2731–2739.10.1093/molbev/msr121 [PubMed: 21546353]
26. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic acids research. 2012; 40:11189–11201.10.1093/nar/gks918 [PubMed: 23066108]
27. Iwamoto T, Yoshida S, Suzuki K, Tomita M, Fujiyama R, Tanaka N, Kawakami Y, Ito M. Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on *Mycobacterium tuberculosis* strains predominated by the Beijing family. FEMS Microbiol Lett. 2007; 270:67–74.10.1111/j.1574-6968.2007.00658.x [PubMed: 17302938]
28. Pérez-Lago L, Comas I, Navarro Y, Gonzalez-Candelas F, Herranz M, Bouza E, Garcia-de-Viedma D. Whole Genome Sequencing Analysis of Intrapatient Microevolution in *Mycobacterium tuberculosis*: Potential Impact on the Inference of Tuberculosis Transmission. The Journal of infectious diseases. 201310.1093/infdis/jit439

29. Meyssonnier V, Li X, Shen X, Wang H, Li DY, Liu ZM, Liu G, Mei J, Gao Q. Factors associated with delayed tuberculosis diagnosis in China. European journal of public health. 2013; 23:253–257.10.1093/eurpub/cks037 [PubMed: 22874738]

30. Sreeramareddy CT, Panduru KV, Menten J, Van den Ende J. Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature. BMC infectious diseases. 2009; 9:91.10.1186/1471-2334-9-91 [PubMed: 19519917]

31. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol. 2006; 44:4498–4510.10.1128/JCM.01392-06 [PubMed: 17005759]

32. Luo T, Yang C, Pang Y, Zhao Y, Mei J, Gao Q. Development of a Hierarchical Variable-Number Tandem Repeat Typing Scheme for Mycobacterium tuberculosis in China. PloS one. 2014; 9:e89726.10.1371/journal.pone.0089726 [PubMed: 24586989]

33. Allix-Beguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, Campos-Herrero I, Mokrousov I, Niemann S, Kontsevaya I, Rastogi N, Samper S, Sng LH, Warren RM, Supply P. Proposal of a Consensus Set of Hypervariable Mycobacterial Interspersed Repetitive-Unit-Variable-Number Tandem-Repeat loci for Subtyping of Mycobacterium tuberculosis Beijing Isolates. J Clin Microbiol. 2014; 52:164–172.10.1128/JCM.02519-13 [PubMed: 24172154]

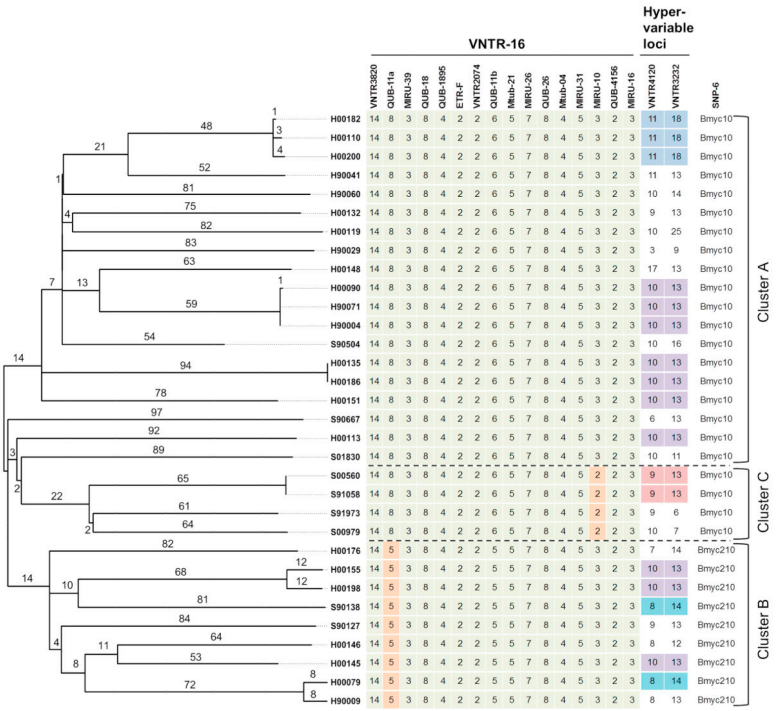

Figure 1.

Whole genome sequencing (WGS) based Maximum Likelihood (ML) phylogeny of the 32 strains in clusters A, B and C and their corresponding genotypes in 16 variable number tandem repeat (VNTR-16) loci and 6 single nucleotide polymorphism (SNP) loci. Branch labels indicate the number of SNPs per branch. The consensus genotype of each locus of VNTR-16 is colored in light green, and the discordant genotypes in loci QUB-11a and MIRU-10 are colored in orange. For two hypervariable loci, colors that are the same indicate genotypes that were shared by at least two strains within a cluster, and uncolored genotypes indicate unique genotypes.

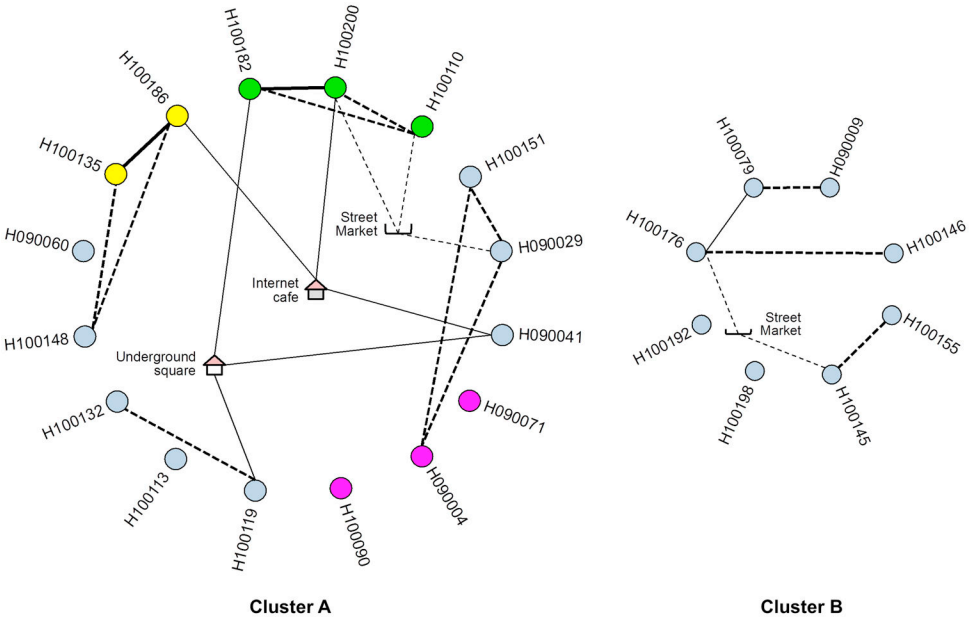

Figure 2.

Epidemiologic links that were identified among TB patients of cluster A and cluster B from Wuchang County, derived from the epidemiological investigations. Circles in light blue, isolates of singletons based on WGS; circles in other colors, isolates of WGS-based clusters; thick solid lines, confirmed epidemiological links; thin solid lines, corresponding TB patients who spent time in an enclosed public place, or TB patients who were familiar and lived or worked in geographically close places; thick dotted lines, corresponding TB patients who lived or worked in geographically close places (e.g., in the same village or in adjacent villages, or on the same street in a town); thin dotted lines, corresponding patients who spent time in an open public place.

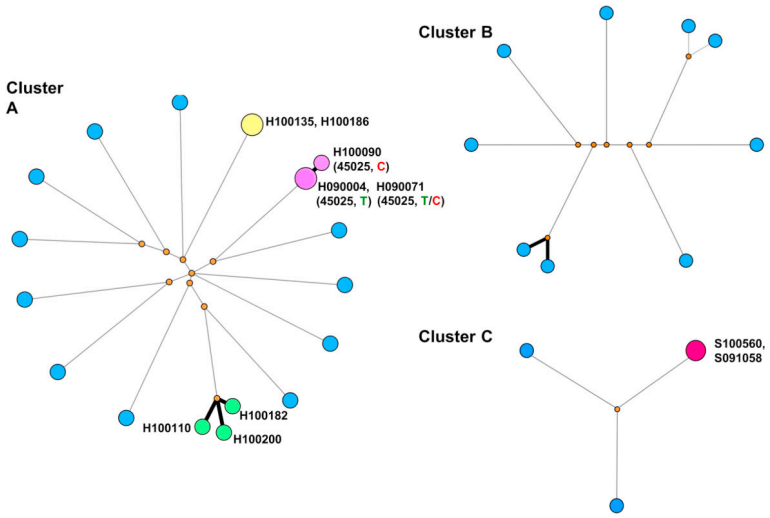

Figure 3.

Median-Joining (MJ) networks of clusters A, B and C. Branches in thin lines, genetic distances ≥ 10 SNPs; branches in thick lines, genetic distances ≤ 10 SNPs; small orange circles, postulated nodes; blue nodes, WGS-based singletons; nodes with colors other than blue or orange, WGS-based clustered isolates (with genomic distance ≤5 SNPs between isolates).

Table 1

Characteristics of clustered cases in Songjiang District, Shanghai, and Wuchang County, Heilongjiang province, from June 1, 2009 to December 2010.

Cluster/sublineage	No. of cases	Cases with WGS data (%)	Male cases no. (%)	Age median (range)	New cases no. (%)	Delay in diagnosis, median (range)[a]	AFB positive	Cavity[b]
Wuchang County								
Cluster A/Bmyc10	16	16 (100)	9 (56.3)	44 (17–79)	16 (100)	92 (9–367)	10 (62.5)	3 (18.8)
Cluster B/Bmyc210	7	7 (100)	5 (71.4)	52 (30–70)	5 (71.4)	33 (22–106)	4 (57.1)	3 (42.9)
Songjiang District								
Cluster A/Bmyc10	7	3 (42.9)	6 (85.7)	44 (29–75)	7 (100)	28 (6–372)	5 (71.4)	3 (42.9)
Cluster B/Bmyc210	5	2 (40.0)	4 (80.0)	31 (21–73)	5 (100)	14 (6–62)	2 (50.0)	2 (50.0)
Cluster C/Bmyc10	7	4 (57.1)	4 (57.1)	38 (16–62)	5 (71.4)	32 (19–44)	1 (33.3)	0

Abbreviations: No., number; WGS, whole genome sequencing; AFB, acid-fast bacilli

[a]Delay in diagnosis, days

[b]Cavity present in the initial chest radiograph at time of diagnosis