ERJ Express. Published on August 16, 2018 as doi: 10.1183/13993003.00246-2018

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

XDR tuberculosis in South Africa: genomic evidence supporting transmission in communities

Sara C. Auld, N. Sarita Shah, Barun Mathema, Tyler S. Brown, Nazir Ismail, Shaheed Vally Omar, James C.M. Brust, Kristin N. Nelson, Salim Allana, Angela Campbell, Koleka Mlisana, Pravi Moodley, Neel R. Gandhi

Please cite this article as: Auld SC, Sarita Shah N, Mathema B, *et al*. XDR tuberculosis in South Africa: genomic evidence supporting transmission in communities. *Eur Respir J* 2018; in press (https://doi.org/10.1183/13993003.00246-2018).

XDR tuberculosis in South Africa: genomic evidence supporting transmission in communities

Sara C Auld MD,[1,2] N Sarita Shah MD,[2,3] Barun Mathema PhD,[4] Tyler S Brown MD,[4,5] Nazir Ismail PhD,[6,7] Shaheed Vally Omar PhD,[6] James CM Brust MD,[8] Kristin N Nelson MPH,[2] Salim Allana MBBS,[2] Angela Campbell MA,[2] Koleka Mlisana MBChB,[9,10] Pravi Moodley PhD,[10] Neel R Gandhi MD[1,2]

1. Emory University School of Medicine, Atlanta, GA, 30322, USA
2. Emory University Rollins School of Public Health, Atlanta, GA, 30322, USA
3. Centers for Disease Control and Prevention, Atlanta, GA, 30333, USA
4. Columbia University Mailman School of Public Health, New York, NY, 10032, USA
5. Massachusetts General Hospital, Boston, MA, 02114, USA
6. National Institute for Communicable Diseases, Johannesburg, 2131, South Africa
7. Department of Medical Microbiology, University of Pretoria, Pretoria, 0002, South Africa
8. Albert Einstein College of Medicine, Bronx, NY, 10461, USA
9. School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal, Durban, 4041, South Africa
10. National Health Laboratory Service, Durban, 4001, South Africa

Corresponding Author:

Sara C Auld, MD, MSc
School of Medicine and Rollins School of Public Health
Emory University
1518 Clifton Rd NE, Claudia Nance Rollins Building, Room 3002

Atlanta, GA 30322, United States

Phone: +1 404-727-9140

Email: sauld@emory.edu

Social media summary: Much of XDR tuberculosis transmission may arise from casual contact between individuals not known to one another.

Abstract

Background: Despite evidence that transmission is driving an extensively drug-resistant (XDR) tuberculosis epidemic, our understanding of where and between whom transmission occurs is limited. We sought to determine whether there was genomic evidence of transmission between individuals without an epidemiologic connection.

Methods: We conducted a prospective study of XDR tuberculosis patients in KwaZulu-Natal, South Africa, during 2011–2014. We collected sociodemographic and clinical data, and identified epidemiologic links based on person-to-person or hospital-based connections. We performed whole-genome sequencing on the *Mycobacterium tuberculosis* isolates and determined pairwise single nucleotide polymorphism (SNP) differences.

Findings: Among 404 participants, 123 (30%) had person-to-person or hospital-based links, leaving 281 (70%) epidemiologically unlinked. The median SNP difference between participants with person-to-person and hospital-based links was 10 (IQR 8–24) and 16 (IQR 10–23), respectively. The median SNP difference between unlinked participants and their closest genomic link was 5 (IQR 3–9); half of unlinked participants were within 7 SNPs of at least five participants.

Conclusions: The majority of epidemiologically unlinked XDR tuberculosis patients had low pairwise SNP differences, consistent with transmission, with at least one other participant. These data suggest that much of transmission may result from casual contact in community settings between individuals not known to one another.

Introduction

Drug-resistant tuberculosis is increasing in many countries and threatens to reverse recent gains in tuberculosis control.[1, 2] Treatment for drug-resistant tuberculosis involves complex, toxic, and costly regimens and is associated with high mortality and poor outcomes.[2] In light of these challenges, prevention of drug-resistant disease is critical in order to reduce morbidity and mortality. Despite increasing evidence that transmission is driving the spread of drug-resistant tuberculosis in many parts of the world,[3-9] our understanding of where and between whom transmission is occurring remains limited. These gaps in our knowledge preclude our ability to design effective interventions to halt transmission and prevent new cases of drug-resistant tuberculosis.

Historically, contact investigations have been used to characterize tuberculosis outbreaks and chains of transmission. Following the advent of molecular genotyping, *Mycobacterium tuberculosis* (*Mtb*) strain-level data have supplemented and enhanced contact investigations, such that cases with similar strains are presumed to be part of a transmission network.[10] However, numerous studies of tuberculosis transmission have found that epidemiologic and genotypic data do not always align. For example, a connection to a close contact could be identified for only 9–30% of genotypically linked individuals across a range of settings,[11-14] suggesting that transmission may arise from casual contact between individuals not known to one another. Similarly, household contact studies in high-incidence settings have found that up to half of secondary tuberculosis cases are not genotypically linked to their presumed index case.[11, 15] These findings likely reflect the overwhelming burden of disease and multitude of transmission opportunities in areas where tuberculosis is prevalent. This discordance between epidemiologic and genotypic data among close contacts also suggests that additional tools, beyond contact investigations and conventional genotyping, may be required to identify the majority of transmission events, particularly in high-burden settings.

Whole-genome sequencing enables a more precise delineation of genetic differences between tuberculosis strains by examining greater than 90% of the *Mtb* genome, compared to less than 1% with traditional genotyping. Whole-genome sequencing identifies the sequential accumulation of single nucleotide polymorphism (SNP) differences and facilitates the construction of transmission chains, rather than simply clusters of related cases as identified through genotyping. Further, individuals with few SNP differences between their *Mtb* strains may represent a transmission link.[16, 17]

We recently demonstrated that 69–92% of XDR tuberculosis in KwaZulu-Natal province, South Africa can be attributed to person-to-person transmission of drug-resistant strains, rather than acquisition of resistance in the setting of prior tuberculosis treatment.[7] We found epidemiologic links through either close contact or hospitalizations for 30% of XDR tuberculosis cases; the epidemiologic source was not identified for the remaining 70% of cases. These findings are similar to estimates from previous population-based transmission studies, which have inferred community-based transmission from the absence of genotypic or genomic transmission links between close contacts.[11-15] However, prior studies have not closely examined genomic transmission links for individuals without an epidemiologic link. In this current study, we utilize whole-genome sequencing to identify potential transmission links between cases without close contact or overlapping hospitalization. By integrating epidemiologic and genomic data, we demonstrate how whole-genome sequencing can improve our ability to identify XDR tuberculosis transmission events occurring as a result of casual contact between individuals not known to one another. A better understanding of community-based transmission will inform targeted efforts to interrupt transmission and accelerate ongoing efforts to reduce the global burden of tuberculosis.

Methods

Study setting and population

This study was conducted in KwaZulu-Natal province, which has a population of 10.3 million persons and the highest rates of tuberculosis (1,076 cases per 100,000 population) and HIV (16.9% prevalence) in South Africa.[18, 19] Patients with culture-confirmed XDR tuberculosis residing in KwaZulu-Natal were recruited into the Transmission of HIV-Associated XDR Tuberculosis (TRAX) study from 2011–2014.[7] The primary objective of the TRAX study was to determine the proportion of patients who had acquired (i.e., secondary to inadequate treatment) versus transmitted XDR tuberculosis.

XDR tuberculosis cases were identified through the single reference laboratory conducting drug-susceptibility testing for all public healthcare facilities in KwaZulu-Natal. Given the large number of XDR tuberculosis cases diagnosed annually in the province, a convenience sample of all diagnosed patients were screened for enrollment. Using the reference laboratory database, age, sex, and district were compared between enrolled and unenrolled participants to determine the representativeness of the study sample. Written informed consent was obtained from all participants, or from the next of kin of deceased or severely ill participants. Interviews were conducted to collect participant sociodemographics, tuberculosis and HIV history, and location and duration of all hospitalizations in the preceding five years. Social network interviews elicited information about close contacts from home, work, and other locations where participants spent at least two hours per week during the preceding five years. Full details of the clinical, social network, and laboratory methods have been previously described.[7]

Laboratory methods

The diagnostic XDR tuberculosis isolate was obtained for all participants and re-cultured on Löwenstein-Jensen slants. Sequencing material was obtained by performing population sweeps of culture plates; genomic DNA extraction and insertion sequence (IS)*6110*-based restriction fragment length polymorphism (RFLP) genotyping were performed according to standard methods.[20] Isolates underwent paired-end whole-genome sequencing and sequencing libraries were prepared using Nextera DNA kits (Illumina, San Diego, CA). Raw paired-end sequencing reads were generated on the Illumina-MiSeq platform and aligned to the H37Rv reference genome (NC_000962.2) using the Burrows-Wheeler Aligner.[21] All isolates had reads covering >99% of the reference genome and the lowest mean coverage depth for any isolate was 15X. SNPs were detected using standard pairwise resequencing techniques (Samtools v0.1.19) against the reference and filtered for quality, read consensus (>75% for the alternate allele) and proximity to indels (<50 base-pairs from any indel). SNPs in or within 50 base-pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were also excluded.[22]

Social network analysis

We analyzed social network interview data to identify epidemiologic links between participants. Person-to-person links were defined as two participants who named each other or named a common intermediary. We identified hospital-based epidemiologic links if there were overlapping dates of hospital admission when at least one participant was in a "vulnerable period," defined as at least 1 month prior to the collection of their XDR tuberculosis diagnostic specimen. Some participants had multiple person-to-person and/or hospital-based links. Participants without a person-to-person or hospital-based link were considered "unlinked." Sociodemographic, clinical, and social network data for linked and unlinked participants were compared using a chi-square or Kruskal-Wallis test.

Whole-genome sequence analysis

We next compared whole-genome sequencing data for linked and unlinked participants. Because we were interested in the ability of whole-genome sequencing to provide further discrimination beyond genotyping, we focused on SNP differences between participants with a matching RFLP pattern (defined as within 1 band). Thus, for linked participants, we identified those who had a matching RFLP pattern with their epidemiologic link, and then determined pairwise SNP differences between those individuals. If a participant had multiple epidemiologic links with a matching RFLP pattern, we selected the link with the fewest pairwise SNP differences (i.e., the "closest") in order to focus on the link most likely to reflect transmission.

For unlinked participants, we identified their closest genomic link (using pairwise SNP differences) within the study cohort. With this approach, unlinked participants had the opportunity to be connected to all other study participants, which would increase their probability of having a genomic link with a low SNP difference by chance alone. In order to account for this possibility, we conducted a sensitivity analysis to evaluate whether unlinked participants had multiple genomic connections at SNP thresholds consistent with transmission, which would support the likelihood of transmission with at least one other study participant with whom they did not have an epidemiologic link. For each unlinked participant we determined the number of other participants within 5, 7, or 10 SNPs.

Ethical considerations

The study was approved by the Institutional Review Boards of Emory University, Albert Einstein College of Medicine, and the University of KwaZulu-Natal, and by the US Centers for Disease Control and Prevention.

Results

Study population

Between May 2011 and August 2014, 1,027 patients were diagnosed with culture-confirmed XDR tuberculosis in KwaZulu-Natal (Figure 1). Study staff approached and screened 521 culture-confirmed XDR tuberculosis patients, and 404 (78%) were eligible and consented to study enrollment. Participants were enrolled from each of KwaZulu-Natal's 11 districts and were representative, by age, sex, and geographically, of all XDR tuberculosis cases diagnosed province-wide during the study period (p=0.52, 0.76, and 0.70, respectively). Over half of the participants were female (234, 58%), 311 (77%) were infected with HIV, and the median age was 34 years (interquartile range [IQR] 28–43) (Table 1).

Social interactions and mobility

We explored participants' social interactions and mobility in order to gain insight into their opportunities for exposure to and transmission of tuberculosis. Participants named a total of 2,901 close contacts from their homes, workplaces and other community locations, with a median of 7 contacts named per participant (IQR 4–10) (Table 2). 123 (30%) participants reported working outside the home, and 129 (32%) reported spending more than 2 hours per week in a community congregate location (e.g., churches, bars, hair salons). There were 46 (12%) participants who reported using public transport for at least 1 hour per day over the year prior to enrollment. In the five years prior to their XDR tuberculosis diagnosis, 298 (74%) participants reported at least one hospitalization, with 86 (29%) participants reporting two or more hospitalizations. These hospitalizations occurred at 53 different hospitals, with a median admission of 3 months (IQR 2–5).

Epidemiologic links

There were 59 (15%) participants with a person-to-person link to at least one other study participant. The majority of links (84%) were to household members, although links also included co-workers (7%) and other individuals in the community (9%). Among participants hospitalized during the vulnerable period prior to their XDR diagnosis, 72 (18%) overlapped with another study participant and had a hospital-based link. Participants overlapped with a median of 3 other participants (IQR 1–18). In total, epidemiologic links were identified for 123 (30%) participants, of whom eight (2%) had both a person-to-person and a hospital-based link.

The remaining 281 (70%) participants were epidemiologically "unlinked." These unlinked participants were not significantly different from linked participants with regards to their sociodemographic and clinical characteristics, with the exception of being slightly older and less likely to report a cough (Table 1). Unlinked participants also did not differ in the number of close contacts they named, whether they lived in an urban or rural area, nor in the number of their hospitalizations (Table 2).

Genomic links

IS*6110* RFLP genotyping was completed on 386 (96%) participants' *Mtb* isolates and whole-genome sequencing was successful in 342 (85%) isolates (see Supplementary Table 1 for participant characteristics by SNPs and Supplementary Table 2 for RFLP clusters). These included 41 (69%) of 59 participants with a person-to-person link, 58 (81%) of 72 participants with a hospital-based link, and 243 (86%) of 281 unlinked participants (Figure 2).

Among participants with person-to-person links, 29 (71%) of 41 had a matching RFLP pattern with their epidemiologic link; their median pairwise SNP difference was 10 (IQR 8–24) (Figure 2). Among participants with hospital-based epidemiologic links, 37 (64%) had a matching RFLP patterns, and the

median pairwise SNP difference to their closest hospital-based link was 16 (IQR 10–23). Among epidemiologically unlinked participants, the median pairwise difference to their closest pair was 5 SNPs (IQR 3–9); thus, half of unlinked participants were within 5 SNPs of at least one other study participant.

The distribution of SNP differences provided further support of potential transmission events among unlinked participants (Figure 3). SNP differences for unlinked participants peaked at 1–4 SNPs (47%), with an additional 29% of participants with 5-9 SNP differences. The *Mtb* strain for five unlinked participants was genomically identical to another participant (i.e., 0 SNP differences) and 18 unlinked participants had an *Mtb* strain that was only 1 SNP different from another participant. In contrast, SNP differences for participants with person-to-person and hospital-based links had a bimodal distribution, with one peak at 5–9 SNPs and the other at >15 SNPs. This bimodal distribution precluded the identification of a distinct SNP threshold for transmission, as has been reported by other studies.[11, 23] Nonetheless, this distribution supports the likelihood that the majority of unlinked participants had a transmission link within the study sample, since 78% of them were within 10 SNPs of at least one other participant.

To further examine potential transmission events among unlinked participants, we examined the number of genomic connections each unlinked participant had below SNP thresholds previously put forth as suggestive of transmission.[23-26] A total of 192 (79%) unlinked participants were connected to at least one other participant by 10 or fewer SNPs, and the median number of connections at this threshold was with 29 participants (IQR 1–80) (Table 3). With a threshold of ≤7 SNPs, 173 (71%) unlinked participants were linked to at least one other participant and the median number of connections at this threshold was with 5 other participants (IQR 0–24). Finally, at a threshold of ≤5 SNPs, 143 (59%) unlinked participants were genomically linked with at least one other participant and the median

number of connections was with 1 other participant (IQR 0–9). Thus, nearly 60% of participants without an epidemiologic link had a genomic link consistent with transmission at the relatively stringent threshold of 5 SNPs, and many participants had multiple links at this threshold.

Discussion

South Africa is facing an epidemic of MDR and XDR tuberculosis driven by transmission of drug-resistant strains. The predominant role of transmission in drug-resistant tuberculosis epidemics has now been demonstrated in a number of settings, and modeling data suggest that transmission will fuel global increases in MDR and XDR tuberculosis over the coming decades.[3-6, 27] Unfortunately, our ability to design effective public health interventions to halt transmission is hindered by our limited understanding of transmission. In this study, we integrated epidemiologic and genomic data in the largest cohort of XDR tuberculosis patients to date and identified genomic links suggestive of transmission for the majority of epidemiologically unlinked participants.

Study participants had many opportunities for XDR tuberculosis transmission, with multiple social contacts, frequenting of community congregate locations, hospitalizations and use of public transport. In focusing on epidemiologically unlinked participants, we were surprised to find that the majority of them had SNP differences similar to, and even lower than, those between epidemiologically linked participants. In fact, we found that many unlinked participants were connected to several study participants by few SNPs. These findings suggest that many of these participants have transmission links that would not be identified by traditional contact investigations or hospital infection control programs—but rather, arise from casual, community-based contact. These links provide empiric

evidence for transmission between casual contacts, supporting the need for future studies to characterize and quantify tuberculosis transmission in community locations.

Our findings extend those of several previous studies that have hypothesized about the role of community transmission after finding that close contacts account for a minority of transmission in high-burden settings.[11, 15, 28-30] In Malawi, the use of population-based whole-genome sequencing revealed that only 9.4% of transmission occurred between close contacts.[11] Studies from China, and a small study from South Africa, also found that a high proportion of genomically linked cases did not have epidemiologic links.[23, 30, 31] Our study, however, provides genomic evidence for transmission between epidemiologically unlinked individuals—with multiple genomic links for the majority of these patients strengthening the likelihood of casual contact as a driver of transmission.

Nearly one-third of epidemiologically linked participants in our study had differing RFLP patterns from one another, which is consistent with previous studies where 39-62% of household contacts had different strains.[11, 12, 15] Furthermore, SNP differences among epidemiologically linked participants demonstrated a bimodal pattern, with half of pairs having a SNP difference below 10-12 SNPs and the other half with SNP differences in the 15-50 range, even when they shared an RFLP pattern. These findings, similar to those reported in other settings,[11, 15] suggest that only a small proportion of secondary cases are attributable to close contact with a known index case. The remaining secondary cases were likely infected with XDR tuberculosis from someone other than a close contact. Thus, in a high burden setting such as South Africa, the presence of an epidemiologic link may be more representative of shared risk factors for exposure than a true transmission link.

Whole-genome sequencing has been used in various settings to identify a SNP threshold indicative of transmission, [16, 17] with low SNP differences between individuals with an epidemiologic link considered a "gold standard" for transmission in several studies.[11, 23] Yet, in our study, many epidemiologically linked participants had SNP differences above previously proposed thresholds (e.g., 0–12 SNPs),[17] and neither epidemiologically linked nor unlinked participants had a SNP distribution with a clear transition point below which transmission could be deemed probable. The challenge of defining SNP thresholds is further highlighted by a recent study of a large tuberculosis outbreak in London where nearly 60% of strains over a 14-year period differed by zero or one SNP, and the maximum number of SNPs between 344 patients was five.[26] Further research is needed to elucidate how *Mtb* mutation rates vary according to pathogen, host, and epidemiologic factors.

There are limitations to the interpretation of whole-genome sequencing data—many of which are not specific to this study, but represent broader challenges for genomic epidemiology. For example, although mutation rates have been characterized in laboratory settings, it remains difficult to estimate mutation rates from clinical and epidemiologic data given the variable latency period of tuberculosis and inherent uncertainty about when an individual may have been infected, particularly in a high-burden setting. Similarly, there is a growing literature demonstrating within-host variability of *Mtb* strains, both at a single time point and over time.[17, 32, 33] The clinical significance of this variability, how it is affected by host factors such as HIV co-infection, and any potential impact on transmission, remain unclear at present. The impact of clonal *Mtb* strains, such as the LAM4/KZN strain in KwaZulu-Natal,[34] on transmission dynamics is also not clear. Nevertheless, the diversity of SNP differences in this cohort indicates that whole-genome sequencing has adequate specificity to differentiate potential transmission events, even in the presence of an endemic strain.

Incomplete capture limits our ability to identify all transmission events. While our study was not designed to capture all XDR tuberculosis cases during the study period, it is possible that greater sampling would have increased the number of epidemiologic links and impacted our estimates of SNP differences between participants. However, studies with more complete sampling worldwide (e.g., United States, Spain, Netherlands, and Malawi) have still found a high proportion of cases without epidemiologic links.[11, 12, 14, 35] Moreover, we found that the majority of unlinked participants had multiple genomic links, even at stringent SNP thresholds. Thus, our findings likely represent a minimum estimate of the proportion of transmission attributable to casual contact.

The ongoing transmission of drug-resistant tuberculosis poses a grave threat to global tuberculosis control. Our ability to halt this growing epidemic will hinge upon the design of targeted interventions to interrupt transmission. We found many opportunities for casual contact and transmission among individuals with XDR tuberculosis. The majority of these individuals were epidemiologically unlinked, yet had genomic evidence of transmission with other study participants, highlighting the potentially substantial contribution of casual contact to the XDR tuberculosis epidemic in KwaZulu-Natal. While contact investigations and infection control programs to prevent nosocomial transmission have proven benefit as a fundamental pillar of tuberculosis control activities, further investigation of transmission through casual contact in community-based settings must be undertaken to determine how and where to intervene and augment our existing approaches for tuberculosis control.

Disclaimer: *The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the U.S. Department of Health and Human Services.*

Acknowledgments*: We are grateful to the study team at the University of KwaZulu-Natal for their tireless efforts in data collection, record abstraction, participant recruitment, and interviews. We thank the participants and their families who consented to participate in this study.*

Funding Source*: This study was primarily funded by a grant from the US National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH): R01AI089349 (PI Gandhi). It was also supported in part by NIH/NIAID grants: R01AI087465 (PI Gandhi), K23AI083088 (PI Brust), K23AI134182 (PI Auld), K24AI114444 (PI Gandhi), Emory CFAR P30AI050409 (PI Curran), Einstein CFAR P30AI051519 (PI Goldstein), by Einstein/Montefiore ICTR UL1 TR001073 (PI Shamoon), and by NIH/NHLBI T32 HL116271 (PI Guidot).*

References

1. Global Tuberculosis Report. WHO, 2017.
2. Dheda K, Gumbo T, Maartens G, Dooley KE, McNerney R, Murray M, Furin J, Nardell EA, London L, Lessem E, Theron G, van Helden P, Niemann S, Merker M, Dowdy D, Van Rie A, Siu GK, Pasipanodya JG, Rodrigues C, Clark TG, Sirgel FA, Esmail A, Lin HH, Atre SR, Schaaf HS, Chang KC, Lange C, Nahid P, Udwadia ZF, Horsburgh CR, Jr., Churchyard GJ, Menzies D, Hesseling AC, Nuermberger E, McIlleron H, Fennelly KP, Goemaere E, Jaramillo E, Low M, Jara CM, Padayatchi N, Warren RM. The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *The Lancet Respiratory medicine* 2017.
3. Gandhi NR, Weissman D, Moodley P, Ramathal M, Elson I, Kreiswirth BN, Mathema B, Shashkina E, Rothenberg R, Moll AP, Friedland G, Sturm AW, Shah NS. Nosocomial transmission of extensively drug-resistant tuberculosis in a rural hospital in South Africa. *J Infect Dis* 2013: 207(1): 9-17.
4. Becerra MC, Appleton SC, Franke MF, Chalco K, Arteaga F, Bayona J, Murray M, Atwood SS, Mitnick CD. Tuberculosis burden in households of patients with multidrug-resistant and extensively drug-resistant tuberculosis: a retrospective cohort study. *Lancet* 2011: 377(9760): 147-152.
5. Devaux I, Kremer K, Heersma H, Van Soolingen D. Clusters of multidrug-resistant Mycobacterium tuberculosis cases, Europe. *Emerg Infect Dis* 2009: 15(7): 1052-1060.
6. Yang C, Shen X, Peng Y, Lan R, Zhao Y, Long B, Luo T, Sun G, Li X, Qiao K, Gui X, Wu J, Xu J, Li F, Li D, Liu F, Shen M, Hong J, Mei J, DeRiemer K, Gao Q. Transmission of Mycobacterium tuberculosis in China: A Population-Based Molecular Epidemiologic Study. *Clinical Infectious Diseases* 2015: 61(2): 219-227.
7. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, Mlisana K, Allana S, Campbell A, Mthiyane T, Morris N, Mpangase P, van der Meulen H, Omar SV, Brown TS, Narechania A, Shaskina E, Kapwata T, Kreiswirth B, Gandhi NR. Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *New England Journal of Medicine* 2017: 376(3): 243-253.
8. Kendall EA, Fofana MO, Dowdy DW. Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. *The Lancet Respiratory medicine* 2015: 3(12): 963-972.
9. Marais BJ, Mlambo CK, Rastogi N, Zozio T, Duse AG, Victor TC, Marais E, Warren RM. Epidemic spread of multidrug-resistant tuberculosis in Johannesburg, South Africa. *J Clin Microbiol* 2013: 51(6): 1818-1825.
10. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. *Clinical microbiology reviews* 2006: 19(4): 658-685.
11. Glynn JR, Guerra-Assuncao JA, Houben RM, Sichali L, Mzembe T, Mwaungulu LK, Mwaungulu JN, McNerney R, Khan P, Parkhill J, Crampin AC, Clark TG. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLoS One* 2015: 10(7): e0132840.
12. Borrell S, Espanol M, Orcau A, Tudo G, March F, Cayla JA, Jansa JM, Alcaide F, Martin-Casabona N, Salvado M, Martinez JA, Vidal R, Sanchez F, Altet N, Coll P, Gonzalez-Martin J. Factors associated with differences between conventional contact tracing and molecular epidemiology in study of tuberculosis transmission and analysis in the city of Barcelona, Spain. *J Clin Microbiol* 2009: 47(1): 198-204.
13. Crampin AC, Glynn JR, Traore H, Yates MD, Mwaungulu L, Mwenebabu M, Chaguluka SD, Floyd S, Drobniewski F, Fine PE. Tuberculosis transmission attributable to close contacts and HIV status, Malawi. *Emerg Infect Dis* 2006: 12(5): 729-735.
14. Mathema B, Bifani PJ, Driscoll J, Steinlein L, Kurepina N, Moghazeh SL, Shashkina E, Marras SA, Campbell S, Mangura B, Shilkret K, Crawford JT, Frothingham R, Kreiswirth BN. Identification and evolution of an IS6110 low-copy-number Mycobacterium tuberculosis cluster. *J Infect Dis* 2002: 185(5): 641-649.

15. Verver S, Warren RM, Munch Z, Richardson M, van der Spuy GD, Borgdorff MW, Behr MA, Beyers N, van Helden PD. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet* 2004: 363(9404): 212-214.
16. Takiff HE, Feo O. Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. *Lancet Infect Dis* 2015: 15(9): 1077-1090.
17. Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Medicine* 2016: 14: 21.
18. Shisana O, Rehle T, Simbayi L, Zuma K, Jooste S, Zungu N, Labardios D, Onoya D, al. e. South African National HIV Prevalence, Incidence and Behavior Survey. HSRC Press, Cape Town, 2014.
19. Ndjeka N. Multi-Drug Resistant Tuberculosis: Strategic Overview on MDR-TB Care in South Africa: Department of Health, Republic of South Africa; 2014.
20. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993: 31(2): 406-409.
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009: 25(14): 1754-1760.
22. Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature communications* 2015: 6: 7119.
23. Yang C, Luo T, Shen X, Wu J, Gan M, Xu P, Wu Z, Lin S, Tian J, Liu Q, Yuan Z, Mei J, DeRiemer K, Gao Q. Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *The Lancet Infectious Diseases* 2016.
24. Guerra-Assuncao JA, Crampin AC, Houben RM, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RP, McNerney R, Fine PE, Parkhill J, Clark TG, Glynn JR. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *Elife* 2015: 4.
25. 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(2): 137-146.
26. 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 Medicine* 2016: 13(10): e1002137.
27. Sharma A, Hill A, Kurbatova E, van der Walt M, Kvasnovsky C, Tupasi TE, Caoili JC, Gler MT, Volchenkov GV, Kazennyy BY, Demikhova OV, Bayona J, Contreras C, Yagui M, Leimane V, Cho SN, Kim HJ, Kliiman K, Akksilp S, Jou R, Ershova J, Dalton T, Cegielski P. Estimating the future burden of multidrug-resistant and extensively drug-resistant tuberculosis in India, the Philippines, Russia, and South Africa: a mathematical modelling study. *The Lancet Infectious Diseases* 2017.
28. Andrews JR, Morrow C, Walensky RP, Wood R. Integrating social contact and environmental data in evaluating tuberculosis transmission in a South African township. *J Infect Dis* 2014: 210(4): 597-603.
29. Middelkoop K, Mathema B, Myer L, Shashkina E, Whitelaw A, Kaplan G, Kreiswirth B, Wood R, Bekker LG. Transmission of tuberculosis in a South African community with a high prevalence of HIV infection. *J Infect Dis* 2015: 211(1): 53-61.
30. Dheda K, Limberis JD, Pietersen E, Phelan J, Esmail A, Lesosky M, Fennelly KP, Te Riele J, Mastrapa B, Streicher EM, Dolby T, Abdallah AM, Ben-Rached F, Simpson J, Smith L, Gumbo T, van Helden P, Sirgel FA, McNerney R, Theron G, Pain A, Clark TG, Warren RM. Outcomes, infectiousness, and

transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. *The Lancet Respiratory medicine* 2017: 5(4): 269-281.
31. Wang W, Mathema B, Hu Y, Zhao Q, Jiang W, Xu B. Role of casual contacts in the recent transmission of tuberculosis in settings with high disease burden. *Clin Microbiol Infect* 2014: 20(11): 1140-1145.
32. Pérez-Lago L, Comas I, Navarro Y, González-Candelas F, Herranz M, Bouza E. Whole genome sequencing analysis of intrapatient microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. *J Infect Dis* 2014: 209.
33. Trauner A, Liu Q, Via LE, Liu X, Ruan X, Liang L, Shi H, Chen Y, Wang Z, Liang R, Zhang W, Wei W, Gao J, Sun G, Brites D, England K, Zhang G, Gagneux S, Barry CE, Gao Q. The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome biology* 2017: 18(1): 71.
34. Gandhi NR, Brust JC, Moodley P, Weissman D, Heo M, Ning Y, Moll AP, Friedland GH, Sturm AW, Shah NS. Minimal diversity of drug-resistant Mycobacterium tuberculosis strains, South Africa. *Emerg Infect Dis* 2014: 20(3): 426-433.
35. Borgdorff MW, van den Hof S, Kalisvaart N, Kremer K, van Soolingen D. Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. *American journal of epidemiology* 2011: 174(2): 243-251.

Table 1. Sociodemographic and clinical characteristics of participants by presence of epidemiologic links.

Characteristic	Total (N = 404) N (%)	Name-based link (N = 59) N (%)	Hospital-based link (N = 72) N (%)[a]	Unlinked (N = 281) N (%)	p-value[b]
Female sex	234 (58)	34 (58)	39 (54)	165 (59)	0.62
Age (median, IQR)	34 (28–43)	33 (29–37)	32 (28–38)	35 (27–44)	**0.04**
Monthly household income ≤ 2500 ZAR[c]	325 (80)	52 (88)	55 (76)	225 (80)	0.77
Ever smoker	39 (10)	8 (14)	5 (7)	27 (10)	0.99
Healthcare worker	24 (6)	5 (8)	5 (7)	15 (5)	0.44
Mine worker	5 (1)	1 (2)	1 (1)	3 (1)	0.64
Diabetes	23 (6)	5 (8)	2 (3)	17 (6)	0.64
HIV-positive	311 (77)	47 (80)	53 (74)	218 (78)	0.67
Median CD4 cell count (IQR) (cells/mm^3)	255 (117–431)	239 (134–380)	239 (97–433)	257 (117–434)	0.54
Cough	333 (82)	52 (88)	63 (88)	224 (80)	**0.03**
Median duration – weeks (IQR)	8.5 (4–12)	8 (3.5–12)	11 (5–16)	8 (4–12)	0.15
Cavitation on CXR	70 (17)	11 (19)	18 (25)	42 (15)	0.06
Acid-fast bacilli (AFB) smear positive	270 (67)	36 (61)	54 (75)	185 (66)	0.58
Any previous TB treatment	291 (72)	38 (64)	58 (81)	202 (72)	0.92
Prior drug-susceptible TB treatment	260 (64)	36 (61)	51 (71)	179 (64)	0.68
Prior multidrug-resistant TB treatment	124 (31)	11 (19)	38 (53)	79 (28)	0.09

[a] There are 8 participants who had both name-based and hospital-based epidemiologic links. These participants were reported in both columns, but for the purposes of statistical analysis were categorized as name-based links and p-values comparing name-based vs hospital-based vs unlinked were reported accordingly. The analyses were repeated with these participants categorized as hospital-based links with no appreciable change in the p-values (data not shown).

[b] P-values are chi-square or Kruskal-Wallis and compare participants with epidemiologic links (either name-based or hospital-based) and unlinked participants.

[c] Currency conversion during study period approximately 1 U.S. Dollar (USD) = 8.4 South African Rand (ZAR); 2500 ZAR = $298 US.

Table 2. Social interactions and mobility of participants by presence of epidemiologic links.

Characteristic	Total (N = 404) N (%)	Name-based link (N = 59) N (%)	Hospital-based link (N = 72) N (%)[a]	Unlinked (N = 281) N (%)	p-value[b]
Number of named contacts (median, IQR)	7 (4-10)	7.5 (5-11)	6.5 (3.5-10)	6 (4-9.5)	0.16
Number of contacts at home (median, IQR)	5 (3-7)	6 (3-9)	5 (3-7)	5 (3-8)	0.30
Number of contacts at work (median, range)	0 (0-11)	0 (0-11)	0 (0-7)	0 (0-7)	0.69
Number of other contacts (median, range)	0 (0-9)	0 (0-9)	0 (0-8)	0 (0-7)	0.33
Rural home residence	204 (50)	24 (41)	39 (54)	145 (52)	0.50
≥2 home residences in previous 5 years	87 (22)	13 (22)	17 (24)	57 (20)	0.26
# working outside the home	123 (30)	19 (32)	22 (31)	82 (29)	0.27
Report spending >2 hours per week in congregate locations	129 (32)	24 (41)	21 (29)	84 (30)	0.27
Use of public transport >1 hour per day	46 (11)	3 (5)	9 (13)	34 (12)	0.72
Any previous hospitalization	298 (74)	42 (71)	72 (100)	192 (68)	**< 0.01**
# of hospitalizations (median, range)	1 (1–5)	1 (1–3)	1 (1–3)	1 (1–5)	0.89
≥2 hospitalizations	86 (29)	10 (24)	24 (33)	56 (29)	0.31
Months of hospitalizations (median, IQR)	3 (2-5)	2.5 (2-4)	5 (3-7)	2 (1-5)	**< 0.01**

[a] There are 8 participants who had both name-based and hospital-based epidemiologic links. These participants were reported in both columns, but for the purposes of statistical analysis were categorized as name-based links and p-values comparing name-based vs hospital-based vs unlinked were reported accordingly. The analyses were

repeated with these participants categorized as hospital-based links with no appreciable change in the p-values (data not shown).

[b] P-values are chi-square or Kruskal-Wallis and compare participants with epidemiologic links (either name-based or hospital-based) and unlinked participants.

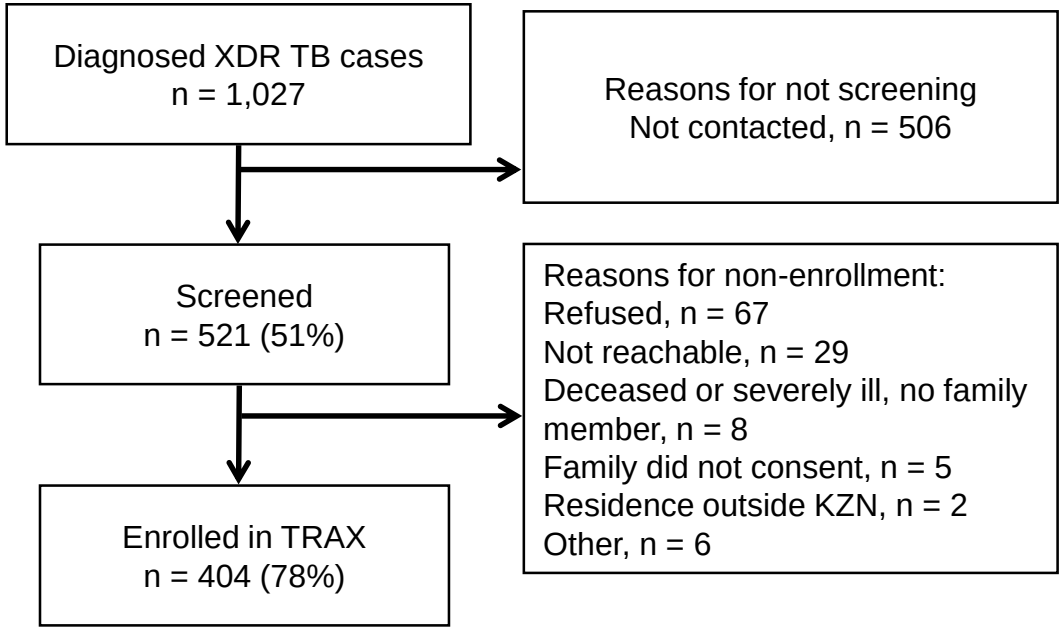
Table 3. Number of epidemiologically unlinked participants (n = 243) with genomic links to other study participants at thresholds of 5, 7, and 10 SNPs.

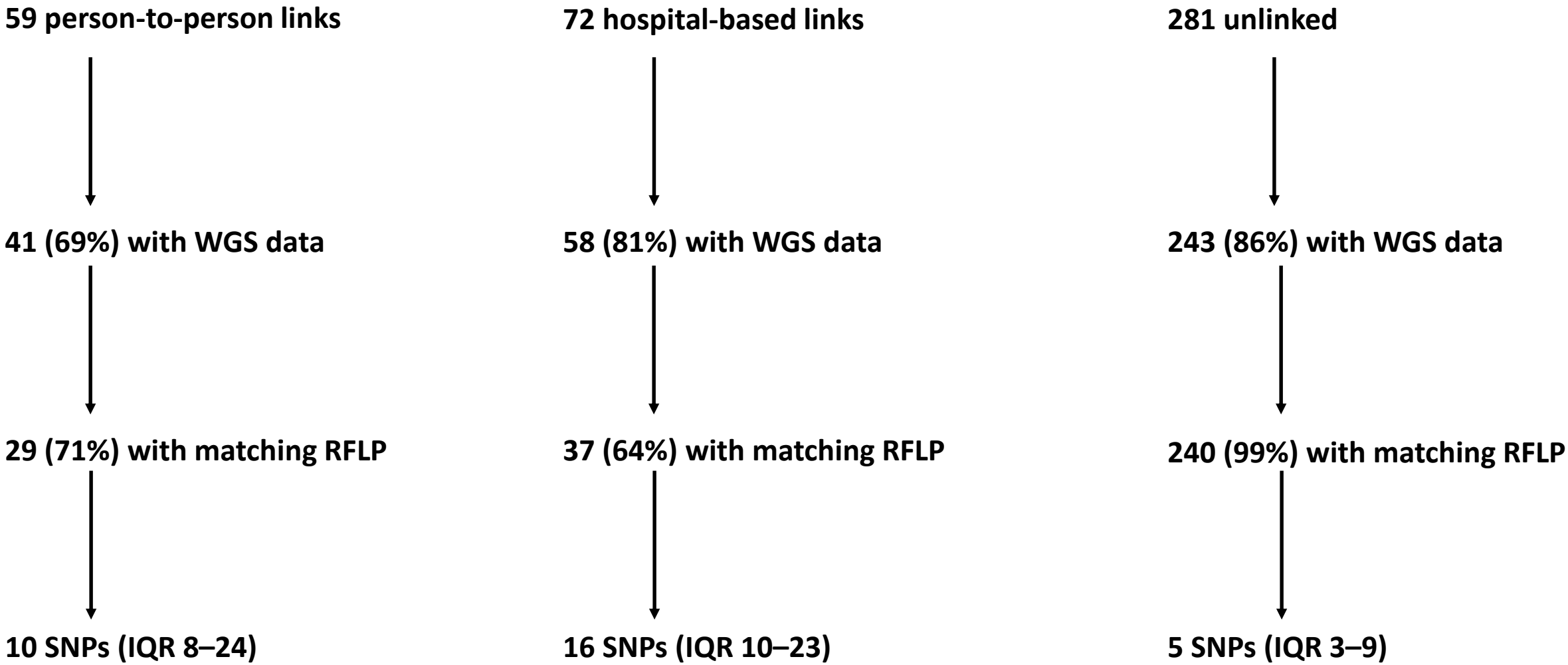
Number of connections to other study participants below threshold	**≤ 5 SNPs**	**≤ 7 SNPs**	**≤ 10 SNPs**
≥ 1	143 (59%)	173 (71%)	192 (79%)
≥ 2	109 (45%)	151 (62%)	177 (73%)
≥ 5	75 (31%)	122 (50%)	153 (63%)
≥ 10	53 (22%)	102 (42%)	148 (61%)

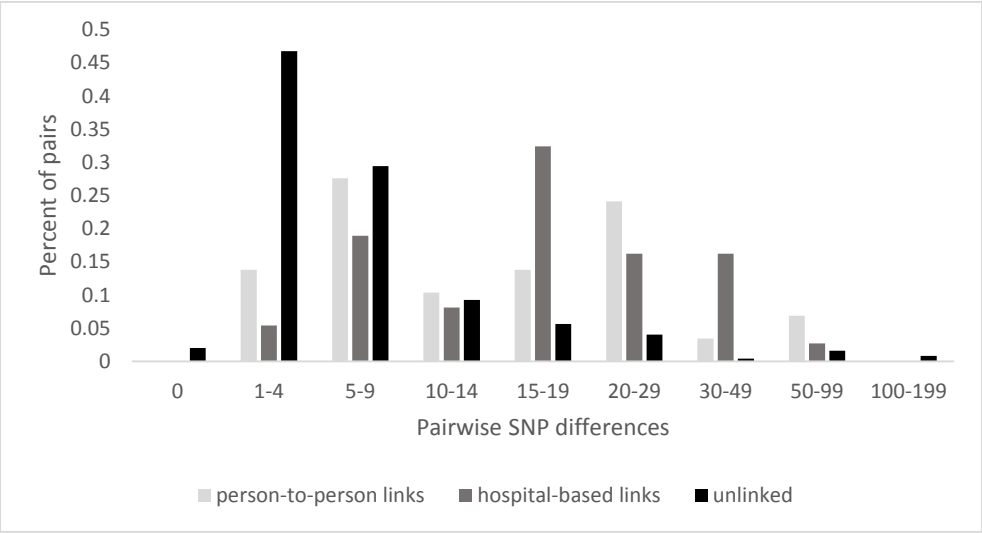
Figure 1. Flow diagram for participant enrollment.

Figure 2. Flow diagram for availability of whole-genome sequencing data and median SNP differences for participants with and without epidemiologic links.
SNP = single nucleotide polymorphism; WGS = whole-genome sequencing; RFLP = restriction fragment length polymorphism; IQR = interquartile range.

Figure 3. Distribution of pairwise SNP differences for participants with a matching RFLP pattern with their name-based epidemiologic link (n = 29), a hospital-based epidemiologic link (n = 37), and without an epidemiologic link (i.e., unlinked participants) (n = 240).


Diagnosed XDR TB cases
n = 1,027
Reasons for not screening
Not contacted, n = 506
Screened
n = 521 (51%)
Reasons for non-enrollment:
Refused, n = 67
Not reachable, n = 29
Deceased or severely ill, no family member, n = 8
Family did not consent, n = 5
Residence outside KZN, n = 2
Other, n = 6
Enrolled in TRAX
n = 404 (78%)


59 person-to-person links
41 (69%) with WGS data
29 (71%) with matching RFLP
10 SNPs (IQR 8–24)
72 hospital-based links
58 (81%) with WGS data
37 (64%) with matching RFLP
16 SNPs (IQR 10–23)
281 unlinked
243 (86%) with WGS data
240 (99%) with matching RFLP
5 SNPs (IQR 3–9)


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Percent of pairs
0
1-4
5-9
10-14
15-19
20-29
30-49
50-99
100-199
Pairwise SNP differences
person-to-person links
hospital-based links
unlinked

Supplementary Table 1. Sociodemographic and clinical characteristics of participants with WGS available (N = 344) according to SNP differences to the closest other study participant.

Characteristic	Total (N = 344) N (%)	0-2 SNPs (N = 95) N (%)	3-5 SNPs (N = 97) N (%)	6-10 SNPs (N = 67) N (%)	> 10 SNPs (N = 85)	p-value[a]
Female sex	202 (59)	52 (55)	66 (68)	39 (58)	45 (53)	0.15
Age (median, IQR)	34 (28-43)	35 (29-43)	34 (26-47)	35 (30-45)	32 (28-39)	0.20
Monthly household income ≤ 2500 ZAR[b]	273 (79)	72 (76)	85 (88)	54 (81)	62 (73)	0.07
Ever smoker	35 (10)	11 (12)	9 (9)	4 (6)	11 (13)	0.51
Healthcare worker	24 (7)	8 (8)	3 (3)	8 (12)	5 (6)	0.15
Mine worker	5 (1)	1 (1)	0 (0)	3 (4)	1 (1)	0.12
Diabetes	22 (6)	3 (3)	6 (6)	10 (15)	3 (4)	**0.01**
HIV-positive	266 (77)	76 (80)	74 (76)	55 (82)	61 (72)	0.42
Median CD4 cell count (IQR) (cells/mm^3)	240 (111-420)	212 (87-375)	296 (120-461)	194 (97-327)	311 (167-434)	**0.03**
Viral load undetectable	133 (50)	37 (49)	38 (51)	27 (49)	31 (51)	0.99
Receiving ART at study enrollment[c]	153 (74)	44 (76)	35 (65)	30 (71)	44 (85)	0.13
Cough	284 (83)	83 (87)	78 (80)	57 (85)	66 (78)	0.31
Median duration – weeks (IQR)	8 (4-12)	8 (4-12)	8 (4-12)	10 (5-12)	9 (4-12)	0.75
Cavitation on CXR	60 (17)	22 (23)	15 (15)	6 (9)	17 (20)	0.10
Acid-fast bacilli (AFB) smear positive	235 (70)	62 (67)	65 (69)	47 (73)	61 (73)	0.81
Any previous TB treatment	247 (72)	72 (76)	63 (65)	44 (66)	68 (80)	0.07
Prior drug-susceptible TB treatment	219 (64)	64 (67)	54 (56)	41 (61)	60 (71)	0.16
Median duration treatment – months (IQR)	6 (6-12)	7 (6-12)	7 (6-13)	6 (6-10)	6 (6-12)	0.79

Prior multidrug-resistant TB treatment	105 (31)	32 (34)	23(24)	10 (15)	40 (47)	**< 0.0001**
Median duration treatment – months (IQR)	6 (4-14)	6 (5-14)	10 (6-15)	3.5 (3-4)	6.5 (4-9)	0.26
Prior cure or treatment completed[d]	7 (7)	3 (10)	2 (8)	0 (0)	2 (7)	0.22
Prior treatment failure[d]	80 (82)	23 (79)	21 (91)	10 (100)	21 (70)	
Prior lost to follow-up or transferred[d]	11 (11)	3 (10)	1 (4)	0 (0)	7 (23)	

[a] P-values are chi-square, Fisher's exact or Kruskal-Wallis.

[b] Currency conversion during study period approximately 1 U.S. Dollar (USD) = 8.4 South African Rand (ZAR); 2500 ZAR = $298 US.

[c] ART start date available for 206 (77%) of participants.

[d] Treatment outcome available for 98 (93%) of the 105 who reported prior MDR TB treatment.

Supplementary Table 2. Pairwise SNP distances between isolates within the same RLFP group (AH-W) and within the same cluster group (1001-2300). Cluster groups containing fewer than three isolates are excluded. Median pairwise distance is shown for the closest participant within a cluster and for the overall cluster.

					SNP distance	
Cluster	# Isolates	# WGS	Lineage	Family	Median closest	Median overall
10(AH)	**15**	**8**	**4**	**X**		**14**
1003	10	6			6	11.5
11(BF)	**3**	**1**	**4**	**T**	**.**	**.**
12(BH)	**6**	**4**	**4**	**S**		**13**
13(BM)	**2**	**1**	**4**	**S**	**.**	**.**
14(BW)	**9**	**5**	**4**	**Harlem**		**18.5**
1401	6	3			6	24
15(CC)	**7**	**4**	**4**	**LAM**		**6.5**
16(GD)	**2**	**0**	**3**	**CAS**	**.**	**.**
17(GY)	**15**	**11**	**4**	**T**		**20**
1701	4	3			10	20
1702	4	4			9	16
18(HP)	**285**	**230**	**4**	**LAM**		**18**
1801	4	4			7	23
1802	3	3			23	23
1805	4	4			19.5	23.5
1810	14	13			8	21.5

1817	5	5			10	15.5
1818	208	163			5	16
1823	8	8			10.5	17.5
19(KO)	**1**	**1**	**4**	**S**	**.**	**.**
20(KR)	**1**	**1**	**4**	**T**	**.**	**.**
21(M)	**1**	**1**	**4**	**T**	**.**	**.**
22(MH)	**28**	**20**	**2**	**Beijing**		**18**
2201	5	3			14	21
2207	4	3			6	10
23(W)	**7**	**4**	**2**	**Beijing**		**211.5**

WGS: Whole genome sequencing

Supplementary Table 3. SNP differences by epidemiologic link and history of prior MDR TB for participants with available WGS and matching RFLP pattern.

	Person-to-person links (N = 29)		**Hospital-based links (N = 37)**		**Unlinked (N = 240)**	
	Prior MDR	No MDR	Prior MDR	No MDR	Prior MDR	No MDR
N (%)	4 (14)	25 (86)	19 (54)	16 (46)	68 (28)	172 (72)
SNPs (median [IQR])	44 (31-50)	10 (6-19)	18 (15-27)	12 (9-18)	5 (2-14)	5 (3-8)