

1 **Genotypic clustering does not imply recent tuberculosis transmission in a high**
2 **prevalence setting: A genomic epidemiology study in Lima, Peru**

3 Avika Dixit^{1,2}, Luca Freschi², Roger Vargas², Roger Calderon³, James Sacchettini⁴, Francis
4 Drobniewski⁵, Jerome T. Galea⁶, Carmen Contreras³, Rosa Yataco³, Zibiao Zhang^{2,7}, Leonid
5 Lecca^{2,3}, Sergios-Orestis Kolokotronis⁸, Barun Mathema⁹, Maha R. Farhat^{2,10}

6
7 ¹Boston Children's Hospital, Boston MA, ²Harvard Medical School, Boston MA, ³Socios En
8 Salud, Lima, Peru, ⁴Texas A&M University, College Station, TX, ⁵Imperial College, London, UK,
9 ⁶University of South Florida, Tampa FL, ⁷Brigham and Women's Hospital, Boston MA, ⁸SUNY
10 Downstate School of Public Health, Brooklyn NY, ⁹Mailman School of Public Health, Columbia
11 University, New York, NY, ¹⁰Massachusetts General Hospital, Boston, MA

12
13 Avika Dixit, MBBS
14 Affiliations: Boston Children's Hospital, Boston MA; Harvard Medical School, Boston MA
15 Address: 300 Longwood Ave, Mailstop 3103, Boston MA 02130, USA

16
17 Luca Freschi, PhD
18 Affiliations: Harvard Medical School, Boston MA
19 Address: 10 Shattuck St, Boston MA 02130, USA

20
21 Roger Vargas, BA
22 Affiliations: Harvard Medical School, Boston MA
23 Address: 10 Shattuck St, Boston MA 02130, USA

24
25 Roger Calderon, MSc
26 Affiliations: Socios En Salud
27 Address: Av. Chimpu Ocllo # 998 Urb. Residencial Lucyana. Carabayllo. Lima 06 – Peru

28
29 James Sacchettini, PhD
30 Affiliations: Texas A&M University, College Station, TX
31 Address: 300 Olsen Blvd, College Station, TX 77843-2128, USA

32
33 Francis Drobniewski, PhD
34 Affiliations: Imperial College London
35 Address: 8th Floor, Office 8.N10, Commonwealth Building, Hammersmith Campus, DuCane
36 Road, London W12 ONN, UK

37
38 Jerome T. Galea, PhD

1 Affiliations: University of South Florida School of Social Work, Tampa FL, USA
2 Address: 13301 Bruce B Downs Blvd, MHC 1416 A, Tampa, FL 33612-3807, USA

3
4 Carmen Contreras, BA
5 Affiliations: Socios En Salud
6 Address: Av. Chimpu Ocllo # 998 Urb. Residencial Lucyana. Carabayllo. Lima 06, Peru

7
8 Rosa Yataco
9 Affiliations: Socios En Salud
10 Address: Av. Chimpu Ocllo # 998 Urb. Residencial Lucyana. Carabayllo. Lima 06, Peru

11
12 Zibiao Zhang, MS
13 Affiliations: Brigham and Women's Hospital, Boston MA; Harvard Medical School, Boston MA
14 Address: 641 Huntington Avenue, Boston, MA 02115, USA

15
16 Leonid Lecca, MD
17 Affiliations: Socios En Salud; Harvard Medical School, Boston MA
18 Address: Av. Chimpu Ocllo # 998 Urb. Residencial Lucyana. Carabayllo. Lima 06, Peru

19
20 Sergios-Orestis Kolokotronis, PhD
21 Affiliations: School of Public Health, SUNY Downstate Medical Center, Brooklyn NY
22 Address: 450 Clarkson Ave, MSC43, Brooklyn, NY 11203-2098, USA

23
24 Barun Mathema, PhD
25 Affiliations: Mailman School of Public Health, Columbia University, New York, NY
26 Address: 722 W 168th St, Room 707, New York, NY 10032, USA

27
28 Maha R. Farhat, MD MSc
29 Affiliations: Massachusetts General Hospital, Boston MA; Harvard Medical School, Boston MA
30 Address: 10 Shattuck St, Boston MA 02130, USA

31
32 **Corresponding Authors:**

33
34 Avika Dixit
35 300 Longwood Ave, Mailstop 3103
36 Boston MA 02115, USA
37 Email: avika.dixit@childrens.harvard.edu
38 Telephone: 617-919-2900

39
40 Maha R. Farhat
41 10 Shattuck St, Boston MA 02130, USA
42 Email: maha@hms.harvard.edu

43

44

45

46

47

1 **Research in context**

2 **Evidence before this study**

3 Use of whole genome sequencing (WGS) to study tuberculosis (TB) transmission has proven to
4 have higher resolution than traditional typing methods in low-burden settings. The implications of
5 its use in high-burden settings are not well understood.

6

7 **Added value of this study**

8 Using WGS, we found that TB clusters defined by traditional typing methods may be circulating
9 for several decades. Genomic regions typically excluded from WGS analysis contain large
10 amount of genetic variation that may affect interpretation of transmission events. We also
11 identified five bacterial mutations that may contribute to transmission fitness.

12

13 **Implications of all the available evidence**

14 Added value of WGS for understanding TB transmission may be even higher in high-burden vs.
15 low-burden settings. Methods integrating variants found in polymorphic sites and insertions and
16 deletions are likely to have higher resolution. Several host and bacterial factors may be
17 responsible for higher transmissibility that can be targets of intervention to interrupt TB
18 transmission in communities.

19

20 **Abstract**

21 **Background:** Whole genome sequencing (WGS) can elucidate *Mycobacterium tuberculosis*
22 (Mtb) transmission patterns but more data is needed to guide its use in high-burden settings. In
23 a household-based transmissibility study of 4,000 TB patients in Lima, Peru, we identified a
24 large MIRU-VNTR Mtb cluster with a range of resistance phenotypes and studied host and
25 bacterial factors contributing to its spread.

1 **Methods:** WGS was performed on 61 of 148 isolates in the cluster. We compared transmission
2 link inference using epidemiological or genomic data with and without the inclusion of
3 controversial variants, and estimated the dates of emergence of the cluster and antimicrobial
4 drug resistance acquisition events by generating a time-calibrated phylogeny. We validated our
5 findings in genomic data from an outbreak of 325 TB cases in London. Using a larger set of
6 12,032 public Mtb genomes, we determined bacterial factors characterizing this cluster and
7 under positive selection in other Mtb lineages.

8 **Findings:** Four isolates were distantly related and the remaining 57 isolates diverged ca. 1968
9 (95% HPD: 1945-1985). Isoniazid resistance arose once, whereas rifampicin resistance
10 emerged subsequently at least three times. Amplification of other drug resistance occurred as
11 recently as within the last year of sampling. High quality PE/PPE variants and indels added
12 information for transmission inference. We identified five cluster-defining SNPs, including *esxV*
13 S23L to be potentially contributing to transmissibility.

14 **Interpretation:** Clusters defined by MIRU-VNTR typing, could be circulating for decades in a
15 high-burden setting. WGS allows for an improved understanding of transmission, as well as
16 bacterial resistance and fitness factors.

17 **Funding:** The study was funded by the National Institutes of Health (Peru Epi study U19-
18 AI076217 and K01-ES026835 to MRF). The funding sources had no role in any aspect of the
19 study, manuscript or decision to submit it for publication.

20

21 **Keywords:** tuberculosis, drug resistance, whole genome sequencing

22

23

1 Introduction

2 Tuberculosis (TB) remains among the top ten causes of deaths globally, with 10.4 million new
3 cases in 2016 alone.¹ Peru remains a high burden country for multidrug (MDR) TB with 117 TB
4 cases were reported per 100,000 population in 2016 with approximately 9% being MDR or
5 rifampicin resistant (RR).¹ Molecular methods have been instrumental in tracing outbreaks and
6 single nucleotide polymorphism (SNPs) identified using whole genome sequencing (WGS) have
7 higher resolution in identifying transmission links when compared with traditional genotyping
8 methods like spoligotyping or *Mycobacterium* interspersed repetitive unit-variable number
9 tandem repeats (MIRU-VNTR).²⁻¹² Yet we don't yet fully understand how to use all the genetic
10 information generated by WGS, by convention as much as 10% of the genome is excluded^{2,3}
11 and in some instances too little remaining variation is found to aid resolution of transmission
12 chains.¹³ Resolving transmission events accurately is particularly challenging in high-burden
13 settings where multiple source case suspects are common. In addition to guiding public health
14 interventions including appropriate contact tracing, identifying the source case can inform
15 patient care in some cases such as in paediatric TB when the source case microbiological data
16 can inform treatment.^{14,15} Further, in high-burden countries the term 'outbreak' may not apply as
17 TB has been circulating continuously for decades or centuries.¹⁶ Given the renewed emphasis
18 on active case finding¹⁷ and widespread adoption of sequencing, a closer look at use of WGS in
19 high-burden settings is needed.

20
21 Control efforts against TB have been undermined by the emergence and spread of drug
22 resistant TB (DR-TB). Current evidence suggests that most cases of DR TB are a result of
23 transmission rather than de-novo evolution of the bacteria during treatment.¹⁸⁻²⁰ Factors known
24 to contribute to DR-TB transmission include delays in diagnosis and treatment,²¹ host factors
25 like age, and immune status,²²⁻²⁴ as well as bacterial factors such as fitness and
26 immunogenicity characteristics.²⁵⁻²⁷ It is well recognized that *Mycobacterium tuberculosis* (MTB)

1 strains with the same DR conferring mutations have a range of fitness.^{28–30} However, to date
2 few molecular fitness determinants have been characterized and seldom in the context of high
3 transmissibility.^{31,32} Improved knowledge of such bacterial factors can inform efforts for
4 transmission interruption by identifying targets for diagnosis, surveillance, and even potential
5 therapeutics targeting fitness mechanisms. Here we use WGS data of the largest TB MIRU-
6 VNTR cluster spanning pan-susceptible to MDR-TB isolates that was identified in 4,000 TB
7 patients enrolled in a household transmissibility study. We examine both host and TB genotypic
8 data to understand the evolution of isolates within this cluster, infer the timing of emergence of
9 antibiotic resistance, and identify genetic bacterial factors unique to this cluster that may have
10 contributed to its success.

11

12 **Methods**

13 *Study Design*

14 A TB household transmissibility and treatment outcome study was performed in northern Lima,
15 Peru from September 2009 to August 2012. The study procedures including patient enrollment
16 and consent have been previously described.^{33,34} Briefly, patients were enrolled if they were
17 diagnosed with pulmonary TB (PTB) at public health clinics and were followed through therapy.
18 Their household contacts were also followed with tuberculin skin testing and monitored for
19 development of TB for a period of 12 months. The following were collected at time of TB
20 diagnosis: clinical signs and symptoms, sociodemographic characteristics (e.g. age, gender,
21 occupation, household type), geographical coordinates of household and health center, co-
22 morbidities (HIV status, diabetes mellitus, renal disorder), as well as alcohol, tobacco and drug
23 use.

24 Approval was obtained from the Research Ethics Committee of the Peruvian National Institute
25 of Health (Lima, Peru) and the Committee on Human Studies at Harvard Medical School
26 (Boston, MA).

1 Please refer to the supplement for bacterial culture, drug susceptibility testing (DST), DNA
2 sequencing methodology and phylogenetic analysis.

3

4 **Results**

5 Patient and isolate characteristics

6 A large cluster of 148 strains with identical MIRU-VNTR pattern was identified. Majority of the
7 patients in this cluster were male, HIV-negative and had multidrug resistant (MDR) TB (Table 1).
8 About one-half were smear-positive (51.4%) and one-third used alcohol (35.8%) or other
9 intoxicating substances (27.9%). Two patients had isolates that did not meet our sequencing
10 quality criteria and were excluded. Of those patients with high quality sequence data (n=61) a
11 higher majority were male with MDR-TB but were otherwise comparable to the superset of 148
12 (Table 1). Sequencing data revealed that 58 isolates belonged to the Latin America-
13 Mediterranean LAM-4.3.3 sublineage, and three isolates were more distant and belonged to the
14 sublineages X -4.1.1, T-4.8, and LAM-4.3.2. In the LAM-4.3.3 group we found 371 SNPs and 81
15 indels in total, of these 24 substitutions and one indel occurred in DR regions, and 42
16 substitutions and 23 indels in PE/PPE genes. With the exclusion of DR and PE/PPE regions,
17 the average pairwise SNP difference between isolates was 21.69 (range: 0-84) and 22.7 (range:
18 1-100), excluding and including the PE/PPE regions, respectively.

19

20 The geographic distribution of strains based on household coordinates, colored by resistance
21 pattern is shown in Figure 1. Comparison between genetic and geographic distance did not
22 support that the cluster spread in a single geographic direction, even when three most distant
23 strains were excluded ($P = 0.2$, Supplementary Figure 1).

24

25 The SNP-based phylogenetic tree (Figure 2, Supplementary Figure 2) demonstrated that the
26 most genetically homogeneous group consisted of 57 isolates. These formed two main clades,

1 where the first contained isolates that were pan-susceptible or streptomycin mono-resistant, and
2 the second consisted of INH mono-resistant or MDR isolates. The origin of the 61 isolates was
3 estimated around the middle of the 14th century (1336 CE; 95% HPD 855-1680). The LAM-4.3
4 cluster of 58 isolates diverged ca 1923 (95% HPD 1856-1967) and innermost cluster of 57
5 isolates diverged ca 1968 (95% HPD 1945-1985).

6
7 Epidemiological vs. genotypic data: We compared the use of genetic and epidemiological data
8 for transmission inference. PE/PPE region differences and indels between closely related pairs
9 were visualized in Integrated Genome Viewer and confirmed to be high confidence
10 (Supplementary methods). Sequencing data was available for three of seven household contact
11 pairs in the cluster of 148 patients. Only one household link (index/parent M23 – contact/child
12 M12) was consistent with a recent transmission event on the tree and by genetic distance (SNP
13 difference = 1). When high-confidence PE/PPE regions were included, the genetic distance
14 between this pair increased to 3 variants. With the further addition of high confidence indels the
15 pairwise distance increased to 4, in the predicted 5 year interval (Supplementary Data 2). No
16 variation was observed in this pair in DR-related loci. The other two isolate pairs from household
17 contacts (index/parent M02 – contact/child M01, index/child M45 – contact/parent M58) were 17
18 and 421 SNPs apart, respectively. Isolates M01 and M02 were genetically closer to other
19 isolates on the tree (M01-M28 6 SNPs apart and M02-M42-M11-M37 all within 4 SNPs of each
20 other). A pair of isolates collected two months apart from a host who had not been on treatment
21 (MDR strains M06 and M10) was found to have no SNP differences outside of PE/PPE regions.
22 In PE/PPE regions, we found 5 high-quality SNPs; similarly, 2 indels were observed in other
23 regions.

24
25 Looking at genetic evidence alone for recent transmission using a distance cutoff of ≤ 5 SNPs
26 (excluding DR and PE/PPE regions)³, 139 links among 38 patients were identified (Figure 3).

1 Other than the pair (index parent M23 and contact child M12), none of these belonged to the
2 same household. With the addition of high confidence indels and PE/PPE SNPs and using the
3 cutoff of ≤ 7 variants (distance between the two isolates belonging to the same patient when high
4 confidence indels and PE/PPE variants were included), there were 41 links among 30 patients,
5 *i.e.* 76% fewer links than when these variants were excluded. Phylogenetic trees built by
6 including indel variation also had notable differences within the cluster of 57 isolates
7 (Supplementary Figure 3).

8
9 Of the 375 isolates sequenced from a TB outbreak in London,¹³ 325 met our quality criteria and
10 were further examined. Using a genetic distance cut-off of ≤ 5 SNP (excluding DR and PE/PPE
11 regions), 309 of these isolates separated out into one large interconnected cluster consisting of
12 31,776 links. Among 38 serial isolates that were collected from 19 patients from that outbreak,
13 the largest difference between a pair with the inclusion of indels and PE/PPE regions was also
14 found to be 7 SNPs. When applying this as the threshold for identifying a genetic link, the
15 interconnected cluster was reduced to 294 strains with 28,230 links, *i.e.* 11% fewer links than
16 when PE/PPE variants and indels were excluded. .

17
18 Host factors: We measured host infectiousness in the cluster using the 'propensity to
19 propagate'³⁵ (PTP) method and identified five patients as having the highest possible score
20 (PTP > 4). This was related to patients being younger (20-29y) males with smear positive PTB
21 and a history of substance use (Supplementary Figure 4). Three of these were identified to be
22 within the network of patients with genetically close MDR isolates (Figure 3). The mean cluster
23 PTP was also high at 1.699.

24
25 Drug Resistance: First and second line drug resistance was acquired several times within the
26 core cluster of 57 isolates (Table 2, Supplementary Figure 5, and Supplementary Table 1).

1 *First line drugs:* Isoniazid (INH) resistance was acquired at least twice, and in both times with a
2 *katG* S315T mutation, ca 1968 (95% HPD 1945-1985) and ca 1972 (95% HPD 1952-1985).
3 Rifampicin resistance (RR) was acquired within the large INH resistant clade at least twice ce
4 1980 (95% HPD 1964-1990, *rpoB* D435V) and ce1986 (95% HPD 1973-1996, most frequent
5 variant *rpoB* S450L). Pyrazinamide (PZA) resistance followed RR and was acquired at least 4
6 times, the most frequent mutation was *pncA* Q10R acquired ce 1980 (95% HPD 1964-1990),
7 the most recent PZA resistance acquisition event was predicted to be within the last year of
8 isolation (95% HPD 2006-2012). Ethambutol resistance was acquired at least twice within the
9 MDR clade and contemporaneous with RR acquisition in both cases. The mutation Y319S was
10 the most common *embB* mutation observed.

11
12 *Second line drugs:* Of the 25 strains tested for ciprofloxacin, one MDR isolate (M43) acquired
13 resistance ce 1996 (95% HPD 1979-2009). Similarly only one (M30, also resistant to
14 capreomycin) of the 41 isolates tested for kanamycin acquired resistance ce 2009 (95% HPD
15 2006-2012). We were not able to identify a mutations in *gyr* or *rrs* to explain resistance to these
16 two drugs. For capreomycin, seven of forty tested isolates were resistant and carried the *tlyA*
17 G232D mutation estimated to have been acquired ce 2005 (95% HPD 2001-2008).

18
19 Resistance and Bacterial Fitness: As there were several isolates measured to be resistant by
20 the culture-based method that did not harbor any known resistance mutations, e.g. for EMB and
21 PZA, we attempted a phylogeny-based genome wide association within the group of 61 isolates
22 to identify new mutations associated with resistance (Supplementary Table 2). In addition to
23 identifying the known mutations that confer resistance to isoniazid and rifampicin, we found an
24 association between EMB resistance and a mutation (3778221AG) in the intergenic region
25 between *spoU* and *PE-PGRS51* genes (20 bp from *spoU* end and 347 bp before *PE-PGRS51*

1 start), corresponding to the acquisition of EMB resistance 32 years ago shown in
2 Supplementary Figure 5.

3
4 We identified 175 mutations that were unique to the core cluster of 57 isolates and were absent
5 from the 4 more distantly related isolates. Hypothesizing that a subset of these mutations may
6 have contributed to transmissibility of this cluster, we measured which are under positive
7 selection by looking at 12,032 other MTB isolates. Five mutations met our criteria for positive
8 selection i.e. were found to have a frequency of >5% in at least 3 other TB lineages (lineage 1,
9 2, 3, and non-LAM-4) (Figure 4). Of the five, two occurred in genes with known function *esxV*
10 which is an ESAT-6 like secreted protein and *cobD*, a cobalamin biosynthesis protein.

11 12 **Discussion**

13 In this detailed analysis of a MIRU-VNTR cluster with variable degrees of drug resistance from a
14 high prevalence setting, we show that traditional genotyping methods have a significantly lower
15 resolution in identifying transmission clusters as compared to WGS, particularly when variation
16 in PE/PPE regions, indels are incorporated into the analysis. Additionally, we found complex
17 evolutionary patterns within an otherwise identical MIRU cluster and identified the interplay of
18 host and epidemiological factors contributing to transmission potential of a cluster. The higher
19 resolution gained by WGS is consistent with prior reports^{2,3,36-38}, but the maximum genetic
20 distance we find between the clustered isolates is larger than previously seen and we further
21 estimate the group of LAM-4.3.3 sequenced isolates to have been circulating for over 8 decades
22 in our study community. Previous studies performing WGS of MIRU-VNTR clusters in low
23 prevalence settings have noted shorter genetic distance between isolates,^{2,3,13} and in one case
24 the distances were insufficient to reliably and consistently inform contact tracing interventions.¹³
25 It is possible, that certain features of our selected cluster have led to the observation of such
26 high levels of diversity. First our cluster spans the spectrum of pan-susceptible to resistant to 7

1 drugs, second is that our isolates all belong to lineage 4, a lineage that has been noted to be the
2 most phenotypically and genotypically diverse of the TB lineages.³⁹ However, the proportion of
3 diversity that could be linked directly to drug resistance was low. A parsimonious explanation of
4 the high degree of observed genomic diversity is that the rate of MIRU-VNTR pattern evolution
5 is on average slow and on the order of decades. Despite this, MIRU-VNTR likely offers sufficient
6 resolution in low prevalence settings as most TB cases there tend to be imported.^{2,3}

7
8 The genes in the PE and PPE families constitute about 10% of the TB genome and have been
9 grouped together based on the proline-glutamate (PE) and proline-proline-glutamate (PPE)
10 signature motifs but members of this family are scattered throughout the genome and have
11 diverse functions.^{40,41} Because they carry a high GC content and contain repetitive areas they
12 have been typically excluded from analysis of sequencing data.⁴² However, recent advances in
13 sequencing technology allow for longer read lengths and increased throughput which combined
14 with more accurate bioinformatics pipelines, it is now possible to call variants in a proportion of
15 these genes with high confidence. We identified several high quality indels and variation in the
16 PE/PPE regions in our dataset. The commonly used cutoff of ≤ 5 SNPs to infer transmission
17 does not take into account the different evolutionary rate of these regions. In our study, when
18 identifying closely related strains with the inclusion of high confidence PE/PPE regions, the
19 number of possible links between strains decreased. An accepted standard that accounts for
20 variation in these regions would allow for improved resolution of transmission events. Although,
21 comparison of ancestral relationships in the phylogenetic trees without and with inclusion of
22 indels did not show significant differences, there were notable differences within the closely
23 related cluster highlighting that similarity measures that rely on SNPs alone could be
24 misleading. Inclusion of indels and PE/PPE regions in estimation of divergence dates is limited
25 by our current lack of knowledge regarding their evolutionary rates but these regions account for

1 an appreciable proportion of variation seen between closely related isolates and thus including
2 this information may affect interpretation of transmission links.⁴³

3
4 We identified many genomic links using the SNP distance threshold of ≤ 5 criterion³ that were
5 not discovered within household contact investigation, providing evidence that household
6 contact investigation is not sufficient to identify and treat secondary TB cases as transmission
7 can occur anywhere in the community. Additionally, 2 of 3 case pairs that belonged to the same
8 household were found to have large genetic distances making it more likely that transmission
9 occurred outside the household. Although the dataset used was relatively small, these findings
10 add to the current limited literature on the topic.⁴⁴⁻⁴⁶ Overall our study highlights the utility of
11 WGS in resolving transmission links particularly in high burden settings where several
12 transmission chains may occur simultaneously. WGS of a well characterized cluster through
13 MIRU-VNTR led to identification of several sub-clusters with further granularity achieved from
14 addition of variants in regions that are routinely excluded from these analyses. With decreasing
15 cost of WGS, sequencing data could be integrated with epidemiological investigation in lieu of
16 traditional fingerprinting methods to identify transmission clusters and for reconstruction of
17 contact networks, particularly given the increasing emphasis on active case finding for TB
18 elimination.¹⁷

19
20 Our phylogenetic dating procedures supports that acquisition of MDR is not recent in Lima, and
21 that MDR cases, given the observed phylogenic structure, are mostly related to transmission.
22 This finding is consistent with other studies performed in other countries including South
23 Africa.¹⁸ Within the MDR sub-cluster, ethambutol and pyrazinamide was acquired and
24 transmitted. This finding was similar to that of a study in Uganda,⁴⁷ where pyrazinamide
25 resistance typically arose in MDR strains with several different causative mutations. This

1 highlights the importance of testing for pyrazinamide resistance in order to determine benefit of
2 its use in MDR-TB treatment regimens.⁴⁸

3
4 Phenotypic resistance could not be explained by genotype in a few isolates including in our
5 study. To this effect, we undertook a GWAS procedure to identify drug resistant phenotype-
6 genotype associations. We identified the intergenic SNP 3778221AG 30bp downstream of the
7 portable tRNA/rRNA methylase gene *spoU* to be significantly associated with ethambutol
8 resistance. Although a causative mechanism for how this variant modulates EMB susceptibility
9 or fitness is not clear, this finding is supported further by a recent large genome-wide
10 association study of 1452 MTB isolates.⁴⁹

11
12 The cluster under study was the largest such cluster observed in the Lima household
13 transmission study. Its transmission success was likely due to both bacterial and host factors.
14 We quantified the host predilection to transmit TB with the PTP measure and found the cluster
15 to have a higher score than the median PTP measure reported by a study in Netherlands.³⁵ Our
16 study had five patients with a particularly high PTP above the highest reported value of 3-9,³⁵
17 potentially contributing to transmission in the population we studied. A few prior studies have
18 characterized bacterial genetic factors that contributed to increased transmissibility.^{24,27,50} We
19 add to this literature by identifying five cluster defining SNPs to be under positive selection in a
20 large TB genomic dataset. One of these SNPs (*esxV* S23L) is a member of the ESAT-6 family
21 of secreted proteins, some of which have been shown to be involved in host-pathogen
22 interactions and may thus have contributed to increased transmissibility.^{50,51}

23
24 Our study had several limitations. Contact tracing was done within household contacts and
25 hence epidemiological links in the community were possibly missed. Sequencing a subset of
26 isolates from the cluster may have led to missed links along the transmission chain. It may also

1 have led to an underestimation of the diversity. However, the sampled subset demonstrated a
2 substantial amount of diversity, more than would be expected within a cluster with identical
3 MIRU pattern.^{2,3} We also cannot exclude that the 2 outer most isolates were miss-assigned the
4 reported MIRU pattern and because of this we focused on the isolates confirmed to be of the
5 same lineage by in silico spoligotyping and the WGS SNP barcode. Finally, it is important to
6 note that our dating estimates are heavily reliant on the molecular clock rate that has been
7 previously reported in the literature.

8
9 In summary, our findings add to the evidence challenging the traditional interpretation of a
10 MIRU-VNTR cluster as indicating recent transmission and suggest that the benefits of WGS
11 over MIRU-VNTR may be even more prominent in high prevalence settings when TB
12 transmission has been ongoing without interruption, especially when high confidence PE/PPE
13 and indel genetic variants are included. WGS can also provide insights into biology of MTB to
14 improve our understanding of DR, transmission and host-pathogen interaction.

15

1 **Authors' contributions:**

2 Avika Dixit conducted the data analysis, drafted and revised the manuscript.

3 All authors provided key edits to the manuscript.

4 Additionally:

5 Luca Freschi contributed to the data analysis.

6 Roger Vargas contributed to the data analysis.

7 Roger Calderon did the DNA extraction and performed the DST testing

8 James Sacchettini conducting the sequencing of isolates included in the study.

9 Francis Drobniowski shared genomic and metadata from the London TB outbreak.

10 Jerome T. Galea helped conduct the household contact study in Peru.

11 Carmen Contreras helped conduct the household contact study in Peru

12 Rosa Yataco helped conduct the household contact study in Peru.

13 Zibiao Zhang managed the household contact study Peru data.

14 Leonid Lecca helped conduct the household contact study in Peru.

15 Sergios-Orestis Kolokotronis verified the phylogenetic analysis.

16 Barun Mathema contributed to study design and phylogenetic analysis.

17 Maha Farhat conceptualized the study, supervised the data analysis, reviewed and edited the
18 manuscript.

19

20 **Declaration of interests:**

21 Avika Dixit: None

22 Luca Freschi: None

23 Roger Vargas: None

24 Roger Calderon: None

25 James Sacchettini: None

26 Francis Drobniowski: None

- 1 Jerome T. Galea: None
- 2 Carmen Contreras: None
- 3 Rosa Yataco: None
- 4 Zibiao Zhang: None
- 5 Leonid Lecca: None
- 6 Sergios-Orestis Kolokotronis: None
- 7 Barun Mathema: None
- 8 Maha R. Farhat: None
- 9
- 10

1 **References**

- 2 1 World Health Organization. Global tuberculosis report 2017. 2017.
- 3 2 Gardy JL, Johnston JC, Ho Sui SJ, *et al.* Whole-genome sequencing and social-network
4 analysis of a tuberculosis outbreak. *N Engl J Med* 2011; **364**: 730–9.
- 5 3 Walker TM, Ip CLC, Harrell RH, *et al.* Whole-genome sequencing to delineate
6 Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis*
7 2013; **13**: 137–46.
- 8 4 Roetzer A, Diel R, Kohl TA, *et al.* Whole Genome Sequencing versus Traditional Genotyping
9 for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular
10 Epidemiological Study. *PLOS Med* 2013; **10**: e1001387.
- 11 5 Kohl TA, Diel R, Harmsen D, *et al.* Whole-genome-based Mycobacterium tuberculosis
12 surveillance: a standardized, portable, and expandable approach. *J Clin Microbiol* 2014; **52**:
13 2479–86.
- 14 6 Gurjav U, Outhred AC, Jelfs P, *et al.* Whole Genome Sequencing Demonstrates Limited
15 Transmission within Identified Mycobacterium tuberculosis Clusters in New South Wales,
16 Australia. *PloS One* 2016; **11**: e0163612.
- 17 7 Outhred AC, Holmes N, Sadsad R, *et al.* Identifying Likely Transmission Pathways within a
18 10-Year Community Outbreak of Tuberculosis by High-Depth Whole Genome Sequencing.
19 *PloS One* 2016; **11**: e0150550.
- 20 8 Bryant JM, Harris SR, Parkhill J, *et al.* Whole-genome sequencing to establish relapse or re-
21 infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet Respir*
22 *Med* 2013; **1**: 786–92.
- 23 9 Luo T, Yang C, Peng Y, *et al.* Whole-genome sequencing to detect recent transmission of
24 Mycobacterium tuberculosis in settings with a high burden of tuberculosis. *Tuberc Edinb*
25 *Scotl* 2014; **94**: 434–40.
- 26 10 Guerra-Assunção JA, Crampin AC, Houben RMGJ, *et al.* Large-scale whole genome
27 sequencing of M. tuberculosis provides insights into transmission in a high prevalence area.
28 *eLife* 2015; **4**. DOI:10.7554/eLife.05166.
- 29 11 Clark TG, Mallard K, Coll F, *et al.* Elucidating emergence and transmission of multidrug-
30 resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PloS*
31 *One* 2013; **8**: e83012.
- 32 12 Ali A, Hasan Z, McNerney R, *et al.* Whole genome sequencing based characterization of
33 extensively drug-resistant Mycobacterium tuberculosis isolates from Pakistan. *PloS One* 2015;
34 **10**: e0117771.

- 1 13Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniowski F. Whole Genome Sequence
2 Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective
3 Observational Study. *PLoS Med* 2016; **13**: e1002137.
- 4 14Management of drug-resistant TB in children. In: Guidance for National Tuberculosis
5 Programmes on the Management of Tuberculosis in Children, 2nd edn. Bethesda, MD: World
6 Health Organization, 2014. <https://www.ncbi.nlm.nih.gov/books/NBK214453/> (accessed July
7 5, 2017).
- 8 15Devi NPPG, Swaminathan S. Drug-Resistant Tuberculosis: Pediatric Guidelines. *Curr Infect*
9 *Dis Rep* 2013; **15**: 356–63.
- 10 16Eldholm V, Monteserin J, Rieux A, *et al.* Four decades of transmission of a multidrug-
11 resistant Mycobacterium tuberculosis outbreak strain. *Nat Commun* 2015; **6**: 7119.
- 12 17Bloom BR. A Neglected Epidemic. *N Engl J Med* 2018; **378**: 291–3.
- 13 18Shah NS, Auld SC, Brust JCM, *et al.* Transmission of Extensively Drug-Resistant
14 Tuberculosis in South Africa. *N Engl J Med* 2017; **376**: 243–53.
- 15 19Dowdy DW, Theron G, Tornheim JA, Kendall EA. Drug-resistant tuberculosis in 2017: at a
16 crossroads. *Lancet Respir Med* 2017; **5**: 241–2.
- 17 20Dheda K, Limberis JD, Pietersen E, *et al.* Outcomes, infectiousness, and transmission
18 dynamics of patients with extensively drug-resistant tuberculosis and home-discharged
19 patients with programmatically incurable tuberculosis: a prospective cohort study. *Lancet*
20 *Respir Med* 2017; **5**: 269–81.
- 21 21WHO Treatment Guidelines for Drug-Resistant Tuberculosis, 2016 Update. Geneva: World
22 Health Organization, 2016 <http://www.ncbi.nlm.nih.gov/books/NBK390455/> (accessed Feb
23 13, 2017).
- 24 22Schaaf HS, Van Rie A, Gie RP, *et al.* Transmission of multidrug-resistant tuberculosis.
25 *Pediatr Infect Dis J* 2000; **19**: 695–9.
- 26 23Borgdorff MW, Hof SVD, Kremer K, *et al.* Progress towards tuberculosis elimination: secular
27 trend, immigration and transmission. *Eur Respir J* 2010; **36**: 339–47.
- 28 24Li W-B, Zhang Y-Q, Xing J, Ma Z-Y, Qu Y-H, Li X-X. Factors associated with primary
29 transmission of multidrug-resistant tuberculosis compared with healthy controls in Henan
30 Province, China. *Infect Dis Poverty* 2015; **4**: 14.
- 31 25Albanna AS, Menzies D. Drug-resistant tuberculosis: what are the treatment options? *Drugs*
32 2011; **71**: 815–25.
- 33 26Verhagen LM, van den Hof S, van Deutekom H, *et al.* Mycobacterial factors relevant for
34 transmission of tuberculosis. *J Infect Dis* 2011; **203**: 1249–55.

- 1 27Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, *et al.* Transmissible Mycobacterium
2 tuberculosis Strains Share Genetic Markers and Immune Phenotypes. *Am J Respir Crit Care*
3 *Med* 2016; published online Dec 20. DOI:10.1164/rccm.201605-1042OC.
- 4 28Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM. The epidemiological fitness cost of
5 drug resistance in Mycobacterium tuberculosis. *Proc Natl Acad Sci* 2009; **106**: 14711–5.
- 6 29Salvatore PP, Becerra MC, Abel zur Wiesch P, *et al.* Fitness Costs of Drug Resistance
7 Mutations in Multidrug-Resistant Mycobacterium tuberculosis: A Household-Based Case-
8 Control Study. *J Infect Dis* 2016; **213**: 149–55.
- 9 30Morcillo NS, Imperiale BR, Di Giulio Á, Zumárraga MJ, Takiff H, Cataldi ÁA. Fitness of
10 drug resistant Mycobacterium tuberculosis and the impact on the transmission among
11 household contacts. *Tuberc Edinb Scotl* 2014; **94**: 672–7.
- 12 31Comas I, Borrell S, Roetzer A, *et al.* Whole-genome sequencing of rifampicin-resistant
13 Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase
14 genes. *Nat Genet* 2012; **44**: 106–10.
- 15 32Farhat MR, Shapiro BJ, Kieser KJ, *et al.* Genomic Analysis Identifies Targets of Convergent
16 Positive Selection in Drug Resistant Mycobacterium tuberculosis. *Nat Genet* 2013; **45**: 1183–
17 9.
- 18 33Huang C-C, Tchetgen ET, Becerra MC, *et al.* The Effect of HIV-Related Immunosuppression
19 on the Risk of Tuberculosis Transmission to Household Contacts. *Clin Infect Dis Off Publ*
20 *Infect Dis Soc Am* 2014; **58**: 765–74.
- 21 34Velásquez GE, Calderon RI, Mitnick CD, *et al.* Pyrazinamide Resistance Assays and Two-
22 Month Sputum Culture Status in Patients with Multidrug-Resistant Tuberculosis. *Antimicrob*
23 *Agents Chemother* 2016; **60**: 6766–73.
- 24 35Nebenzahl-Guimaraes H, Borgdorff MW, Murray MB, van Soolingen D. A novel approach -
25 the propensity to propagate (PTP) method for controlling for host factors in studying the
26 transmission of Mycobacterium tuberculosis. *PLoS One* 2014; **9**: e97816.
- 27 36Tyler AD, Randell E, Baikie M, *et al.* Application of whole genome sequence analysis to the
28 study of Mycobacterium tuberculosis in Nunavut, Canada. *PLoS One* 2017; **12**: e0185656.
- 29 37Fiebig L, Kohl TA, Popovici O, *et al.* A joint cross-border investigation of a cluster of
30 multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic,
31 genotyping and whole genome sequencing methods: lessons learnt. *Euro Surveill Bull Eur Sur*
32 *Mal Transm Eur Commun Dis Bull* 2017; **22**. DOI:10.2807/1560-7917.ES.2017.22.2.30439.
- 33 38Norheim G, Seterelv S, Arnesen TM, *et al.* Tuberculosis Outbreak in an Educational
34 Institution in Norway. *J Clin Microbiol* 2017; **55**: 1327–33.
- 35 39Stucki D, Brites D, Jeljeli L, *et al.* Mycobacterium tuberculosis lineage 4 comprises globally
36 distributed and geographically restricted sublineages. *Nat Genet* 2016; **48**: 1535–43.

- 1 40 Musser JM, Amin A, Ramaswamy S. Negligible Genetic Diversity of Mycobacterium
2 tuberculosis Host Immune System Protein Targets: Evidence of Limited Selective Pressure.
3 *Genetics* 2000; **155**: 7–16.
- 4 41 Copin R, Coscollá M, Seiffert SN, *et al.* Sequence Diversity in the *pe_pgrs* Genes of
5 Mycobacterium tuberculosis Is Independent of Human T Cell Recognition. *mBio* 2014; **5**:
6 e00960-13.
- 7 42 Galagan JE. Genomic insights into tuberculosis. *Nat Rev Genet* 2014; **15**: 307–20.
- 8 43 Stimson J, Gardy JL, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP threshold:
9 identifying outbreak clusters using inferred transmissions. *bioRxiv* 2018; : 319707.
- 10 44 Kompala T, Shenoj SV, Friedland G. Transmission of tuberculosis in resource-limited
11 settings. *Curr HIV/AIDS Rep* 2013; **10**: 264–72.
- 12 45 Middelkoop K, Mathema B, Myer L, *et al.* Transmission of Tuberculosis in a South African
13 Community With a High Prevalence of HIV Infection. *J Infect Dis* 2015; **211**: 53–61.
- 14 46 Mathema B, Andrews JR, Cohen T, *et al.* Drivers of Tuberculosis Transmission. *J Infect Dis*
15 2017; **216**: S644–53.
- 16 47 Ssengooba W, Meehan CJ, Lukoye D, *et al.* Whole genome sequencing to complement
17 tuberculosis drug resistance surveys in Uganda. *Infect Genet Evol J Mol Epidemiol Evol*
18 *Genet Infect Dis* 2016; **40**: 8–16.
- 19 48 Miotto P, Cirillo DM, Migliori GB. Drug Resistance in Mycobacterium tuberculosis:
20 Molecular Mechanisms Challenging Fluoroquinolones and Pyrazinamide Effectiveness.
21 *Chest* 2015; **147**: 1135–43.
- 22 49 Farhat M r., Calderon R, Ioerger T, *et al.* Genome Wide Association Study of Mycobacterium
23 Tuberculosis Reveals Multiple Novel Genes Associated with Large Increase in Drug
24 Minimum Inhibitory Concentrations. In: A25. TUBERCULOSIS MANAGEMENT: NEW
25 INSIGHTS. American Thoracic Society, 2018: A1153–A1153.
- 26 50 Holt KE, McAdam P, Thai PVK, *et al.* Frequent transmission of the Mycobacterium
27 tuberculosis Beijing lineage and positive selection for the *EsxW* Beijing variant in Vietnam.
28 *Nat Genet* 2018; **50**: 849–56.
- 29 51 Pallen MJ. The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system?
30 *Trends Microbiol* 2002; **10**: 209–12.

31

32

33

34

1 **Acknowledgement**

2

3 We thank Megan Murray, the co-PI of the Peru Epi study for helpful input on the manuscript. We

4 thank the patients and their families who contributed to this study. We also thank the Partners in

5 Health healthcare personnel at participating health centers in Lima, Peru. We wish to thank Dr.

6 Robert Husson for his invaluable feedback on an initial draft of the manuscript.

7

8

1 **Tables and Figures**

2 **Table 1:** Patient characteristics. None of the variables were significantly different between the
 3 patient with and without sequencing data using a t-test or a Chi-squared test.

	Cluster*	With WGS**
Age***	32.06 (17-86)	34.8 (17-86)
Female	47 (31.8%)	17 (27.9%)
Positive sputum smear	76 (51.4%)	30 (49.1%)
Heavy drinker in previous year	53 (35.8%)	24 (39.3%)
Drug use	41 (27.9%) (Total n = 147)	20 (33.3%) (Total n = 60)
HIV positive	3 (2.1%) (Total n = 140)	3 (5.3%) (Total n = 57)
Resistance Pattern [#]	N (%)	N (%)
Susceptible	26 (17.6)	9 (14.8)
Isoniazid Resistant	18 (12.2)	13 (21.3)
Multi-drug Resistant	56 (37.8)	31 (50.8)
Other	36 (24.3)	8 (13.1)
Not available	12 (8.1)	0 (0)

4 *N = 148 unless specified

5 **N = 61 unless specified

6 ***Mean and range

7 [#]Resistance pattern for three of the sequenced strains was inferred based on mutations

8

9

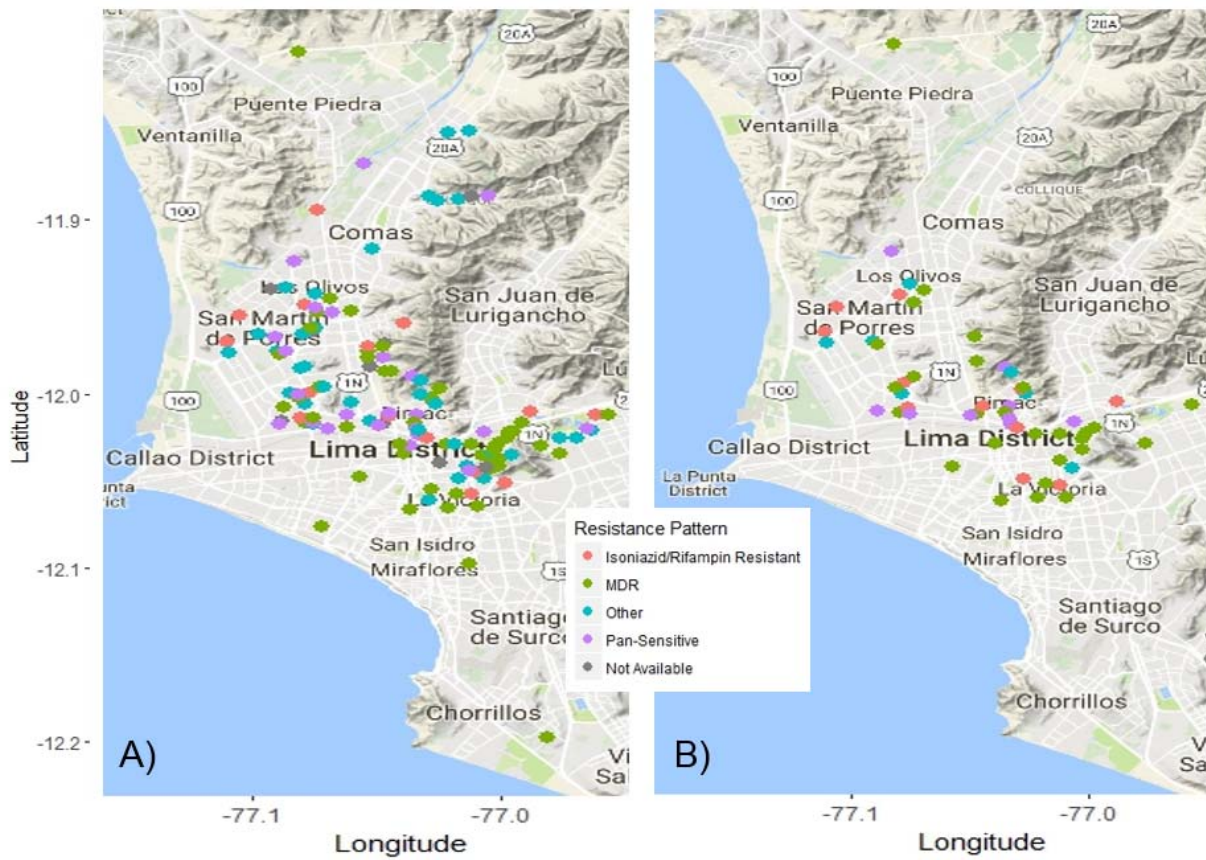
10

1 **Table 2:** Drug resistance acquisition in MTB strains based on phenotype

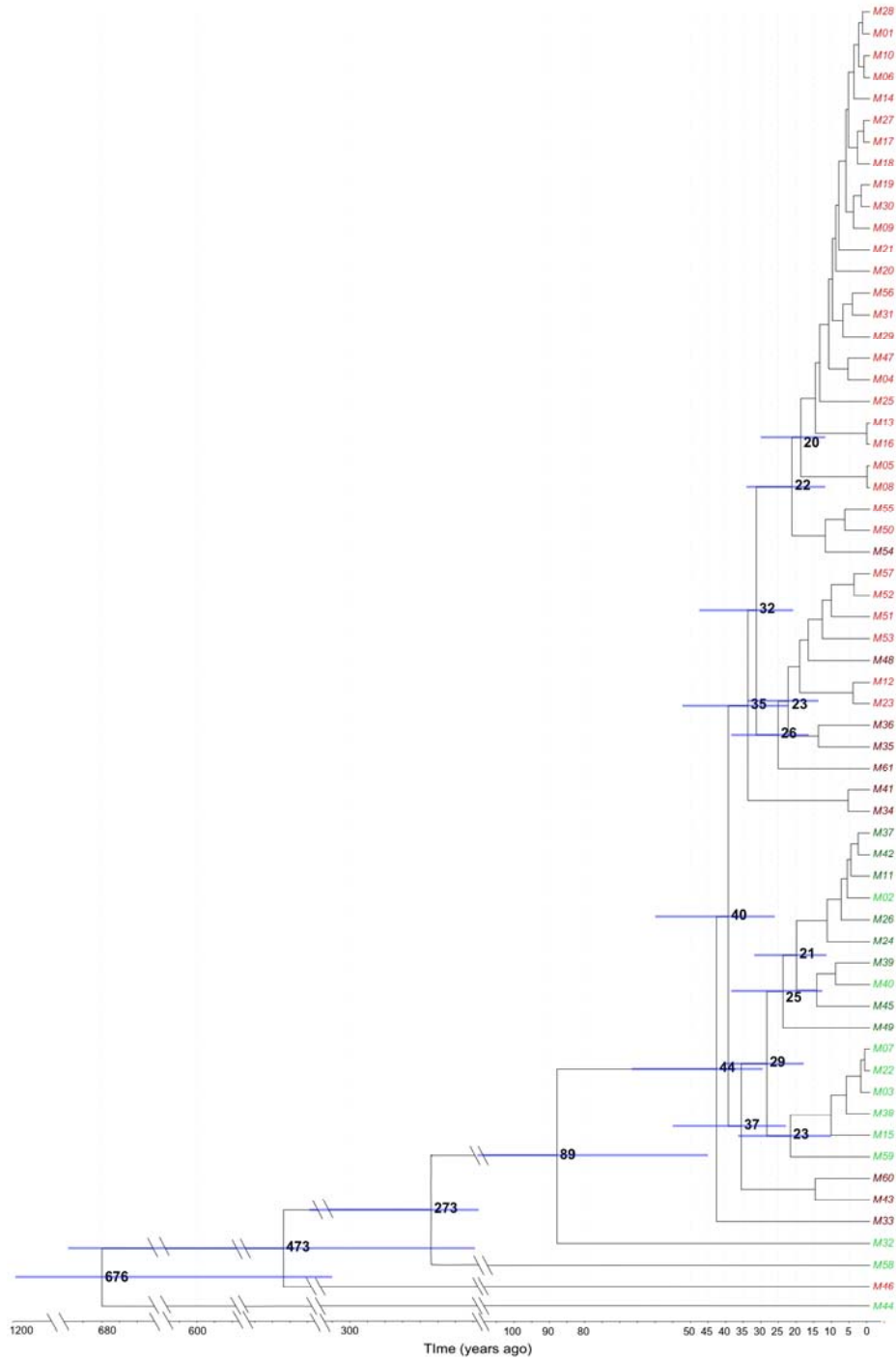
Antibiotic	Minimum DR acquisition instances	Most recent DR acquisition (year)
Isoniazid	2	1972
Rifampicin	2	2007
Pyrazinamide	5	2011
Ethambutol	3	1996
Ciprofloxacin	1	1996
Kanamycin	1	2009
Capreomycin	1	2003

2

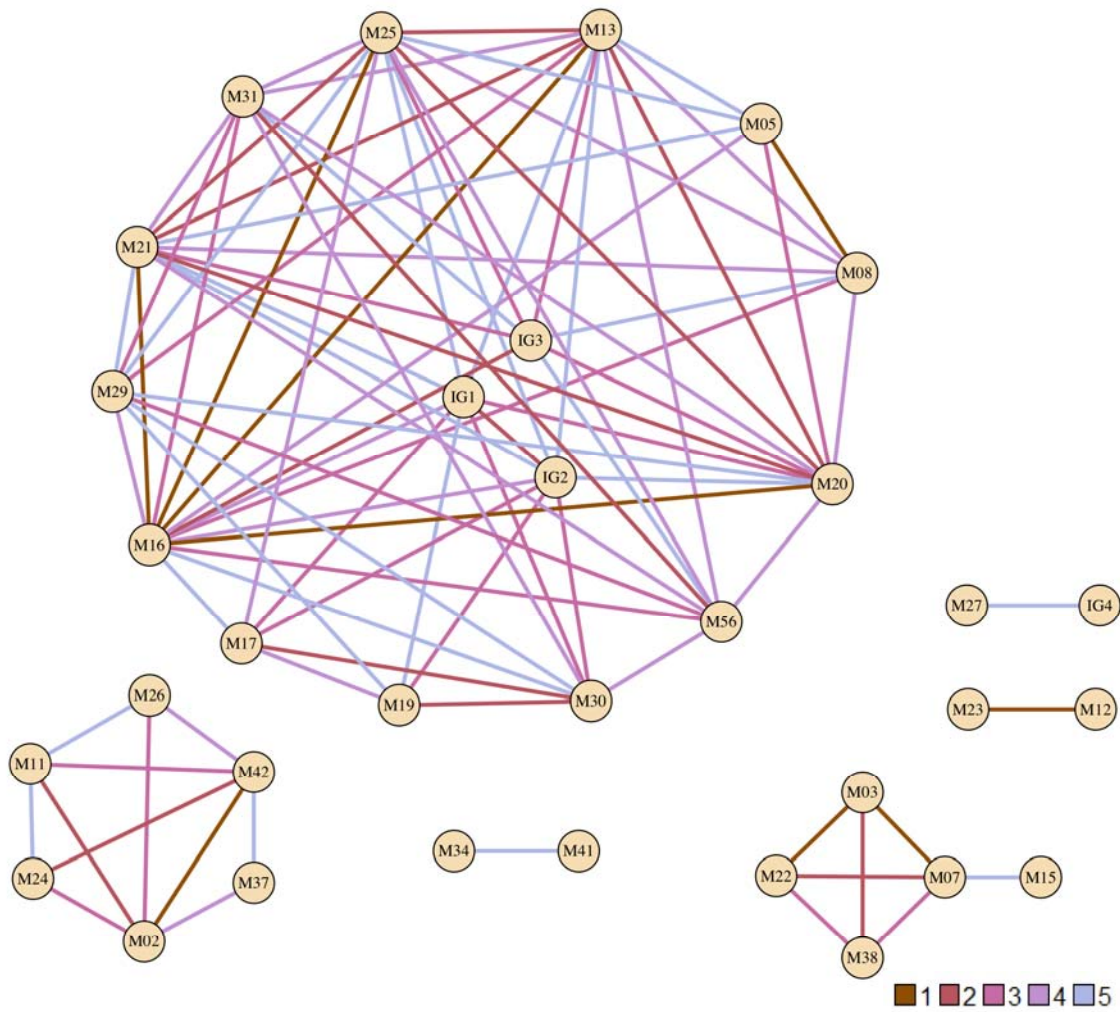
3



- 1
- 2 Figure 1: Geographic location of strains with resistance pattern. A) All 148 strains with identical MIRU-VNTR patterns,
- 3 B) 61 strains that were sequenced.



1
2 Figure 2: Bayesian maximum clade credibility phylogenetic tree created via BEAST (using single nucleotide
3 polymorphisms) of 61 strains with nodes in increasing order of age. Node ages < 15 years are not shown for clarity.
4 Bars represent 95% HPD interval for node age. Color of tip represents drug susceptibility - Green: pan-susceptible,
5 Dark Red: Resistant only to Isoniazid or Rifampicin, Dark Green: Resistant to a drug other than Isoniazid or
6 Rifampicin, Red: multi-drug resistant.



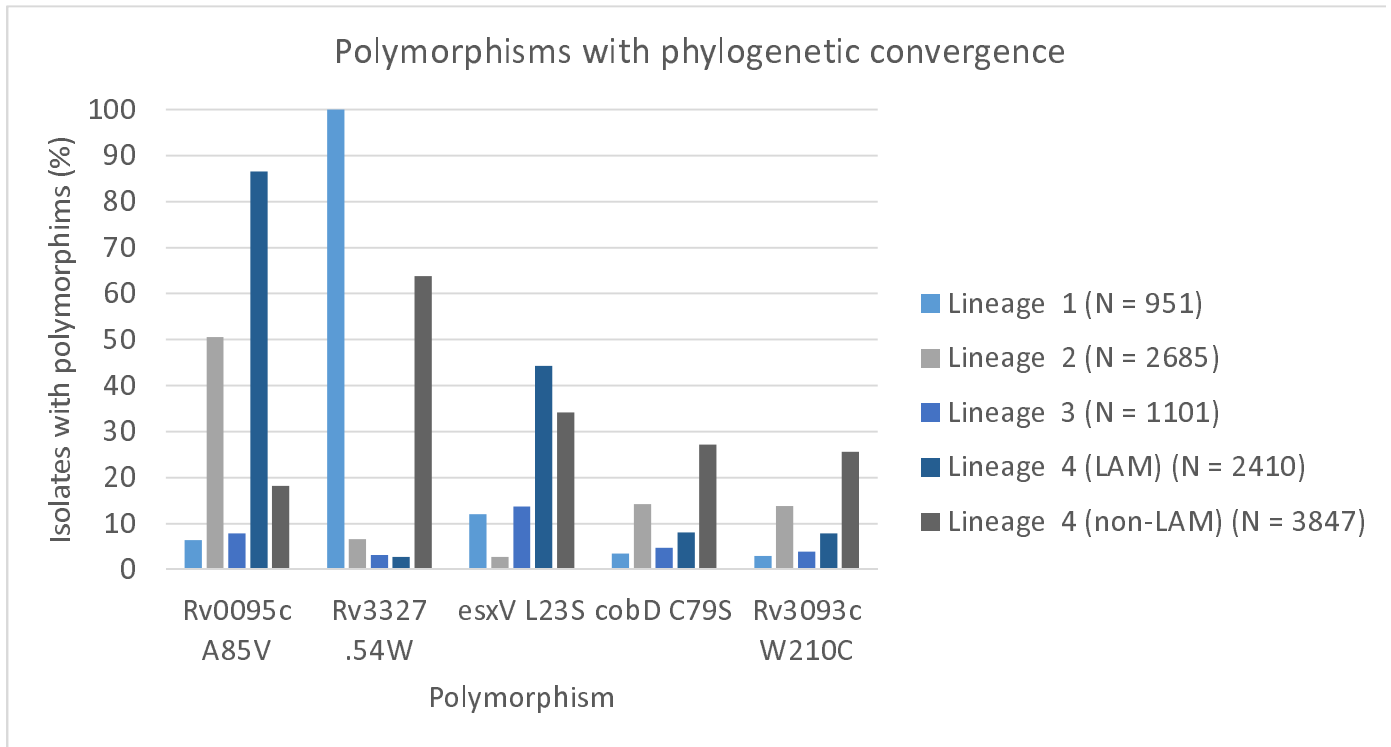
1

2

3 *Figure 3: Network of strains that were less than or equal to 5 single nucleotide polymorphisms (SNPs) apart (with the*
4 *exclusion of PE/PPE regions and regions coding for drug resistance). IG = Identical Group – consists of strains that*
5 *had zero SNP difference, IG1 = M09, M14, M01; IG2 = M06, M10; IG3 = M04, M07; IG4 = M18, M28. Legend shows*
6 *number of different SNPs and corresponding color of the edges. Plot generated using R package igraph.*

7

1



2
3

4 *Figure 4: Single nucleotide polymorphisms in high-transmission cluster of 57 strains showing phylogenetic*
5 *convergence and their percent frequency among Mycobacterium tuberculosis lineages.*

6