MAJOR ARTICLE

# Genomic Epidemiology of Multidrug-Resistant *Mycobacterium tuberculosis* During Transcontinental Spread

**Mireia Coscolla,[1,2] Pennan M. Barry,[3] John E. Oeltmann,[6] Heather Koshinsky,[4] Tambi Shaw,[3] Martin Cilnis,[3] Jamie Posey,[6] Jordan Rose,[5] Terry Weber,[3] Viacheslav Y. Fofanov,[4] Sebastien Gagneux,[1,2] Midori Kato-Maeda,[5] and John Z. Metcalfe[5]**

[1]Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, and [2]University of Basel, Switzerland; [3]Division of Communicable Disease Control, Center for Infectious Diseases, California Department of Public Health, Richmond, [4]Eureka Genomics, Hercules, and [5]Division of Pulmonary and Critical Care Medicine, Francis J. Curry International Tuberculosis Center, San Francisco General Hospital, University of California; and [6]Centers for Disease Control and Prevention, Atlanta, Georgia

**The transcontinental spread of multidrug-resistant (MDR) tuberculosis is poorly characterized in molecular epidemiologic studies. We used genomic sequencing to understand the establishment and dispersion of MDR *Mycobacterium tuberculosis* within a group of immigrants to the United States. We used a genomic epidemiology approach to study a genotypically matched (by spoligotype, IS6110 restriction fragment length polymorphism, and mycobacterial interspersed repetitive units–variable number of tandem repeat signature) lineage 2/Beijing MDR strain implicated in an outbreak of tuberculosis among refugees in Thailand and consecutive cases within California. All 46 MDR *M. tuberculosis* genomes from both Thailand and California were highly related, with a median difference of 10 single-nucleotide polymorphisms (SNPs). The Wat Tham Krabok (WTK) strain is a new sequence type distinguished from all known Beijing strains by 55 SNPs and a genomic deletion (Rv1267c) associated with increased fitness. Sequence data revealed a highly prevalent MDR strain that included several closely related but distinct allelic variants within Thailand, rather than the occurrence of a single outbreak. In California, sequencing data supported multiple independent introductions of WTK with subsequent transmission and reactivation within the state, as well as a potential super spreader with a prolonged infectious period. Twenty-seven drug resistance–conferring mutations and 4 putative compensatory mutations were found within WTK strains. Genomic sequencing has substantial epidemiologic value in both low- and high-burden settings in understanding transmission chains of highly prevalent MDR strains.**

***Keywords.*** *Mycobacterium tuberculosis*; drug resistance; genomics; epidemiology; EmbR.

*Mycobacterium tuberculosis* is an ancient human pathogen that continues to cause substantial morbidity and mortality, in part due to an expanding global epidemic of drug-resistant disease. In the United States, nearly 90% of multidrug-resistant (MDR) tuberculosis cases occur among foreign-born individuals [1], although the relative proportion occurring through reactivation of latent MDR strains, direct importation of active disease, and domestic transmission and reactivation, is not definitively known. Effective tuberculosis control strategies depend upon understanding these parameters among high-risk groups immigrating to the United States [2].

Analysis of data from next-generation whole-genome sequencing (WGS) allows detection of minute differences in genetic diversity and has contributed retrospectively to outbreak investigations [3–7] and population-based studies [8] in high-income settings. In the study of drug-resistant tuberculosis, WGS has improved understanding of causal mechanisms of drug resistance [9] and mutations compensatory for fitness costs associated with drug resistance [10]. Yet,

Received 23 September 2014; accepted 6 January 2015; electronically published 18 January 2015.

Correspondence: John Z. Metcalfe, MD, PhD, MPH, University of California, San Francisco, Division of Pulmonary and Critical Care Medicine, San Francisco General Hospital, 1001 Potrero Ave, Rm 5K1, San Francisco, CA 94110-0111 (john.metcalfe@ucsf.edu).

**The Journal of Infectious Diseases® 2015;212:302–10**



DOI: 10.1093/infdis/jiv025

transcontinental molecular epidemiology of drug-resistant tuberculosis, including data from both high- and low-burden settings, is poorly represented in existing molecular epidemiologic studies of tuberculosis [11].

During 2004–2005, high MDR tuberculosis case rates among refugees living at Wat Tham Krabok (WTK) in Thailand coincided with the final major resettlement of Hmong peoples to the United States [12]. Transcontinental importation and evidence for domestic transmission of a lineage 2/Beijing MDR *M. tuberculosis* strain led to major changes in Centers for Disease Control and Prevention (CDC) preimmigration tuberculosis screening protocols [13]. We generated and analyzed WGS data from *M. tuberculosis* genomes in both Thailand and the United States to clarify importation and establishment of the WTK strain within California.

## METHODS

### Study Population

Since the late 1970s, WTK (a Buddhist temple in Saraburi Province, Thailand) has been home to Hmong refugees fleeing political persecution in Laos. Following recognition of high MDR tuberculosis case rates among Hmong refugees seeking resettlement and those recently resettled in the United States, a CDC-coordinated outbreak investigation at WTK between April 2004 and July 2005 identified 272 tuberculosis cases among 15 455 refugees, with 24 of 57 culture-positive individuals (42%) found to have MDR tuberculosis. Twenty isolates had identical IS6110 restriction fragment length polymorphisms (IS6110-RFLPs), spoligotyping results, and mycobacterial interspersed repetitive units–variable numbers of tandem repeat (MIRU-VNTR) signatures; of these, 15 (75%) had contact investigation data [12], were available among CDC-banked specimens, and were included in our analysis [12]. Documented exposure among several patients who had MDR tuberculosis simultaneously, clustering of genotypes, concordant results of phenotypic drug susceptibility tests, and a high prevalence of tuberculin reactivity among household contacts were considered as evidence supporting an MDR tuberculosis outbreak within the camp.

California (2010 population, 37.2 million) state law requires reporting of all verified cases of tuberculosis (California Code of Regulations Title 17 §2500) to the California Department of Public Health (CDPH) Tuberculosis Registry. Routine genotyping of clinical *M. tuberculosis* isolates has occurred since 2004, with cases prior to 2004 undergoing genotyping upon request (eg, in the course of an outbreak investigation and for special projects). MDR *M. tuberculosis* isolates in California with a genotype matching that yielded by the WTK investigation in Thailand (based on criteria described in the Conventional Genotyping subsection, below) were identified by searching genotyping results in the national and CDPH tuberculosis genotyping database. Of 225 MDR tuberculosis cases diagnosed during 2004–2010, 22 (10%) occurred among persons of Hmong ethnicity. Five cases (23%) involved directly imported active MDR tuberculosis (symptomatic, culture-positive within 1 month of US arrival) and occurred concurrently with the Thailand outbreak. A systematic review of additional matching isolates that were identified during outbreak investigations and genotyped in California yielded 9 additional WTK MDR isolates collected during 1995–2003. Additional information from the epidemiologic investigation is available in the Supplementary Materials. All protocols were approved by the California Health and Human Services Agency Committee for the Protection of Human Subjects and the University of California, San Francisco, Committee for the Protection of Human Subjects.

### Conventional Genotyping

Extraction of genomic DNA from *M. tuberculosis* strains was performed during the log-phase growth of strains on culture medium. Spoligotyping, 24-locus MIRU-VNTR, and IS6110-RFLP genotyping were performed using standardized protocols. Isolates with an identical spoligotype (000000000003771), 24-locus MIRU-VNTR signature, and IS6110-RFLP (21-band pattern, ±1 band) were considered matching; for 12 of 31 California isolates (38%) and all Thai isolates, a 12-locus rather than 24-locus MIRU-VNTR genotype was available.

### Phenotypic Drug Susceptibility Testing

California state law (California Code of Regulations Title 17 §2505) requires submission of all *M. tuberculosis* isolates to local public health laboratories and submission of all MDR *M. tuberculosis* isolates to the California Department of Public Health Microbial Diseases Laboratory (MDL). Tests for first- and second-line antituberculosis drug susceptibilities were performed at local laboratories or at the MDL, using the BACTEC 460 (Becton Dickinson Diagnostic Instruments, Sparks, Maryland), the MGIT 960 system (Becton Dickinson), or the agar proportion method.

### Sequencing

Forty-six MDR strains (15 strains from Thailand, and 31 from California) were sequenced using HiSeq (Illumina; Supplementary Table 1). Burrows-Wheeler Aligner v0.5.8c (BWA) and SMALT (https://www.sanger.ac.uk/resources/software/smalt/) were used to map Illumina reads from these 46 genome sequences and 56 previously published lineage 2 genomes (Supplementary Table 1 and Supplementary Table 2) against an inferred common ancestor of all *M. tuberculosis* complex lineages. An inferred common ancestor, rather than a previously sequenced strain (eg, H37Rv), was used as reference to avoid recovering mutations present only in the previously sequenced strain. The average number of reads that covers each position in the reference

genome ranged from 40× to 350× in different strains, with a median of 110×. Only nonredundant single-nucleotide polymorphisms (SNPs) identified with BWA and SMALT mapping were retained. For each strain, we called SNPs with Phred-scaled probability scores of >20, read depths lower than double the average read depth of the genome, and a minimum of 5 reads. For filtering dense SNPs, a maximum of 2 SNPs were allowed within a window of size 10. Subclusters of strains were taken to represent putative transmission chains and were defined as genetically related *M. tuberculosis* isolates (≤4 SNP difference and proximal phylogenetic relationship according to our median joining network) from individuals with presumed or likely epidemiologic contact. Drug resistance–associated mutations identified in the Tuberculosis Drug Resistance Mutation Database [14] were retrieved from the SNP list (Table 1 and Supplementary Table 3). Drug resistance–conferring mutations (DRMs) in *rpoB* are noted in the text by use of *Escherichia coli* notation; compensatory mutations were identified as nonsynonymous SNPs in *rpoA* or *rpoC* (Supplementary Table 3).

### Phylogenetic Analysis

To examine the genetic diversity of the WTK strain, we sequenced and analyzed all 46 genomic sequences in conjunction with 56 widely diverse and geographically distributed lineage 2 strains selected from a global database (Supplementary Figure 1, Supplementary Table 1 and Supplementary Table 2). High-confidence DRMs were excluded from the diversity and phylogenetic analyses, since these are known to represent homoplastic events (ie, stereotyped evolution under the common selection pressure of antituberculosis medications).

Details about sequencing and phylogenetic analytic methods, including full references, are specified in the Supplementary Materials.

**Table 1. Nonsynonymous Mutations According to Drug Resistance–Associated Locus**

| Drug Resistance Phenotype | Locus Name[a] | H37Rv Locus[a] | Nucleotide Position[a] | Amino Acid Change[a] | Reference Base | Mutant Base | Consequence of SNP |
|---|---|---|---|---|---|---|---|
| DR-flq | *gyrA* | Rv0006 | 7581 | D94N/D | G | R | Nonsynonymous |
| DR-emb-inh | *iniA* | Rv0342 | 412 280 | Q481H | G | T | Nonsynonymous |
| DR-rif | *rpoB* | Rv0667 | 761 139 | H445D[b] | C | G | Nonsynonymous |
| DR-rif | *rpoB* | Rv0667 | 761 139 | H445Y[b] | C | T | Nonsynonymous |
| DR-rif | *rpoB* | Rv0667 | 761 140 | H445R[b] | A | G | Nonsynonymous |
| DR-rif | *rpoB* | Rv0667 | 761 155 | S450L[c] | C | T | Nonsynonymous |
| DR-sm | *rpsL* | Rv0682 | 781 687 | K43R | A | G | Nonsynonymous |
| DR-sm | *rrs* | MTB000019 | 1 473 246 | . . . | A | G | . . . |
| DR-ami | *tlyA* | Rv1694 | 1 918 207 | G90S | G | A | Nonsynonymous |
| DR-ami | *tlyA* | Rv1694 | 1 918 664 | W242X | G | A | Stop-gain |
| DR-inh | *katG* | Rv1908c | 2 155 168 | S315T | C | G | Nonsynonymous |
| DR-pza | *pncA* | Rv2043c | 2 288 778 | V155G | A | C | Nonsynonymous |
| DR-pza | *pncA* | Rv2043c | 2 288 839 | T135A | T | C | Nonsynonymous |
| DR-pza | *pncA* | Rv2043c | 2 289 040 | W68R | A | G | Nonsynonymous |
| DR-pza | *pncA* | Rv2043c | 2 289 162 | L27P | A | G | Nonsynonymous |
| DR-pza | *pncA* | Rv2043c | 2 289 228 | I5T | A | G | Nonsynonymous |
| DR-inh | *accD6* | Rv2247 | 2 521 428 | D229G | A | G | Nonsynonymous |
| DR-emb | *manB* | Rv3264c | 3 645 731 | T83P | T | G | Nonsynonymous |
| DR-emb | *embB* | Rv3795 | 4 247 429 | M306L | A | C | Nonsynonymous |
| DR-emb | *embB* | Rv3795 | 4 247 429 | M306V | A | G | Nonsynonymous |
| DR-emb | *embB* | Rv3795 | 4 247 431 | M306I | G | A | Nonsynonymous |
| DR-emb | *embB* | Rv3795 | 4 247 574 | D354A | A | C | Nonsynonymous |
| DR-emb | *embB* | Rv3795 | 4 247 646 | A378E | C | A | Nonsynonymous |
| DR-emb | *embB* | Rv3795 | 4 247 730 | G406A | G | C | Nonsynonymous |
| DR-emb | . . . | Rv3806c | 4 269 387 | D149E | G | T | Nonsynonymous |
| DR-sm | gidB | Rv3919c | 4 407 927 | E92D | T | G | Nonsynonymous |
| DR-emb | *embA prom* | IG3858 | 4 243 222 | −11 | C | A | . . . |

Abbreviation: SNP, single-nucleotide polymorphism.

[a] Refer to the *Mycobacterium tuberculosis* H37Rv genome.

[b] Corresponds to H526 in *Escherichia coli*.

[c] Corresponds to S531 in *Escherichia coli*.

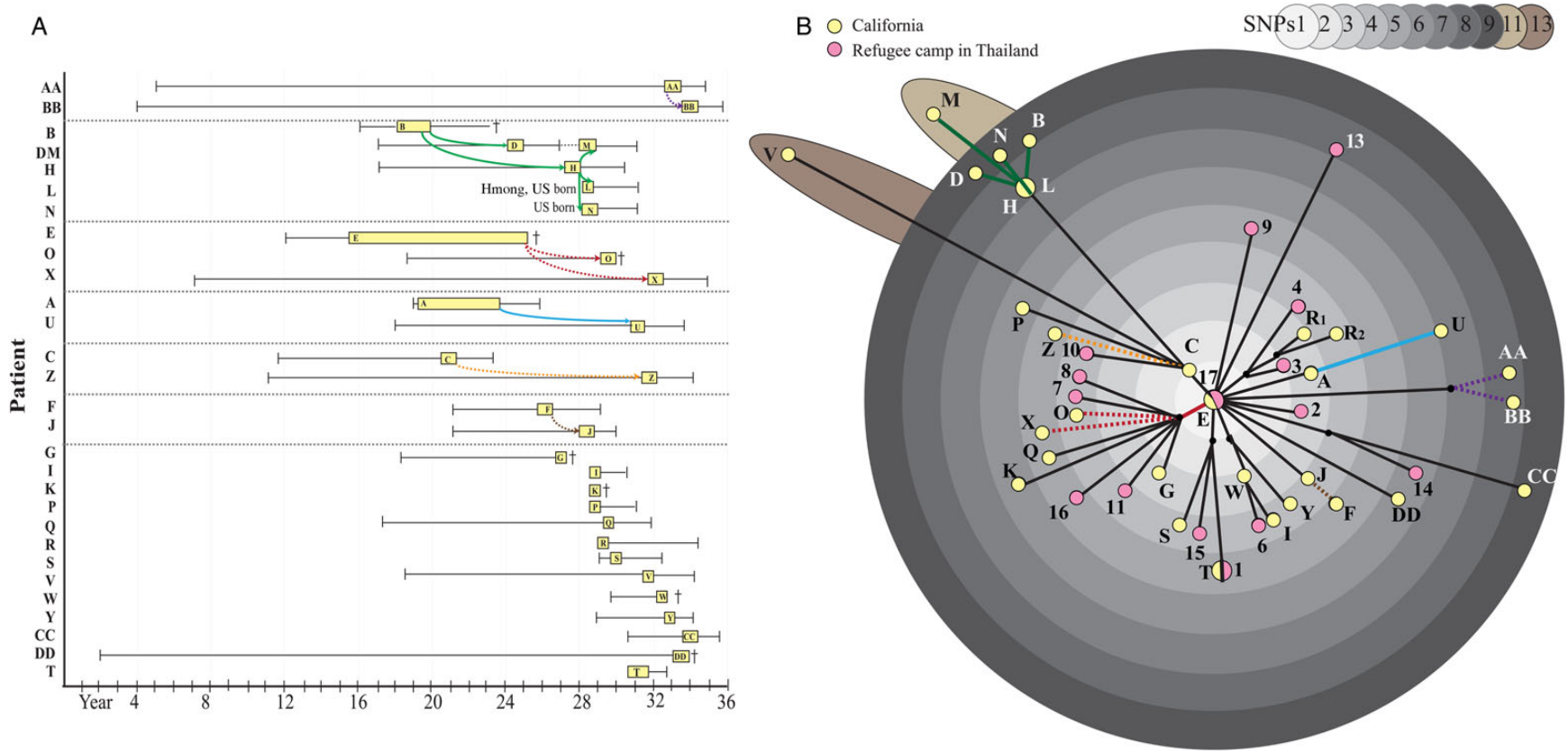


**Figure 1.** *A*, Epidemiologic linkages of multidrug-resistant tuberculosis cases in California and Thailand. California cases (yellow) are represented by the approximate date of US entry (left whisker), infectious period (box), and end of treatment (right whisker) or death (cross). Infectious periods were estimated as the interval from 3 months prior to symptom onset or abnormal chest radiograph through 2 weeks following initiation of appropriate antimicrobial therapy [15]. Epidemiologic links are indicated as colored arrows (solid lines indicate known epidemiologic links, and dashed lines indicate possible epidemiologic links). Twenty-nine patients for whom we had sequence data for the infecting strains are indicated in the *y*-axis. *B*, Median joining network. The relationships of 46 Wat Tham Krabok (WTK) isolates were determined using 150 variable single-nucleotide polymorphisms (SNPs), excluding high-confidence drug resistance–associated mutations (Supplementary Table 3). Isolates and epidemiologic links are colored as described for panel *A*. Inferred nodes (unsampled) are represented by black circles. Shading indicates the relative SNP difference with respect to the central node. Note that cases D and M and cases $R_1$ and $R_2$ are paired longitudinal cases corresponding to 1 patient each.

## RESULTS

### Genomic Epidemiologic Investigations

In total, we performed WGS and analysis of 46 genotypically matched isolates from 15 patients in WTK and 29 patients (2 of whom had MDR tuberculosis twice) in California over a 22-year interval (Figure 1 and Supplementary Table 1). The overall MDR tuberculosis incidence at WTK for 2004–2005 was 15 cases per 100 000. In California, the overall MDR tuberculosis incidence for 2004–2010 was 0.8 cases per 100 000 in the general population and 3.4 cases per 100 000 in the California Hmong population of 91 224 (in 2010). The median age of patients was in Thailand (35 years; interquartile range [IQR], 23–57 years) was similar to that in California (43 years; IQR, 20–66 years; $P = .4$). Approximately one-third of patients (n = 9) died during treatment in California, while mortality data were not available for Thailand. Three of 15 patients (20%) in Thailand and 6 of 29 (21%) in California were known to have previously received standard first-line tuberculosis treatment. Two individuals with a MDR WTK strain in California were born in the United States, one of whom was not of Hmong ethnicity. Among 7 individuals within household or community subclusters, the median time to reactivation following the end of the infectious period of an MDR tuberculosis putative source case was 6.2 years (IQR, 3.8–7.4 years). Among 13 individuals not within a well-supported transmission chain (genotypes of their isolates differed by >4 SNPs) and with tuberculosis not diagnosed upon arrival, the median time from US arrival to reactivation of MDR *M. tuberculosis* infection was 8.5 years (IQR, 3.8–12.0 years). During the study period in California, no tuberculosis due to a drug-susceptible WTK strain occurred (among any ethnic group), and all MDR tuberculosis affecting persons of Hmong ethnicity was caused by a WTK strain.

Genome sequencing resolved the WTK cluster defined by conventional genotyping into several subclusters (≤4 SNP difference, proximal phylogenetic relationships, and epidemiologic linkages) in both California and Thailand (Figure 1). In Thailand, only 3 of 12 cases (25%) regarded as epidemiologically linked in transmission chains were confirmed by genome sequencing. Moreover, we observed multiple branch points in the network, consistent with several closely related but distinct allelic variants (ie, a highly prevalent strain), rather than evidence of a single outbreak characterized by short genetic distances representing recent chains of transmission (eg, well-characterized in our study by the B-D-H-L-N group in California). The presence of a highly prevalent strain was further supported by multiple distinct combinations of DRMs, indicating drug resistance acquired independently on multiple occasions, rather than transmission of 1 drug-resistant strain from patient to patient (Figure 2).

In California, genomic data supported a single case (patient E) whose isolate occupied the central node in the WTK network within California. Patient E arrived in the United States in the mid-1980s; received a diagnosis of cavitary, smear-positive MDR tuberculosis 2 years later; withdrew from treatment within a year; and had sputum smear-positive MDR tuberculosis diagnosed at death, 10 years later, indicating a potentially prolonged infectious period. Seven cases (A, C, F, G, O, Q, and X) were contemporaneous with case E, and all had isolates that shared similar genotypes (≤4 SNP difference), suggesting that these cases may be in a chain of transmission. However, contact investigations could identify definitive epidemiologic links only among a subset of cases (O and X) that were extended family or household contacts (Figure 1). Interestingly, the isolate from patient 17 (who had MDR tuberculosis diagnosed in Thailand and had not been in contact with patient E for >15 years) co-occupied the central node of the network with a nearly identical genotype. Epidemiological data integrated with the genomic network also demonstrated multiple independent importation events from Thailand with reactivation and transmission within the state. Patients I, S, P, Y, and R arrived in the United States following the death of patient E and likely represent independent importation events. Patients D and $R_1$ had a second episode (M and $R_2$, respectively) of MDR tuberculosis, the former considered reinfection and the latter considered relapse following incomplete treatment. Public health contact investigation activities at the time of the most recent Hmong resettlement (2004) also documented an MDR tuberculosis outbreak. This outbreak involved two neighboring households and presumed transmission to a US-born person in a school setting. Genomic data demonstrated only minor differences in SNPs (≤4) between the isolate from patient B (the index case) and those from subsequent cases (H, L, M, and N) in this transmission chain. Overall, genomic data supported all known links (100%; 10 of 10) and 78% of possible links (7 of 9). In addition, 7 other links (AA-BB, E-A, E-C, E-F, E-G, E-Q, and I-W-Y) were suggested by genomic data but were not supported epidemiologically. Of note, directionality according to WGS violated the temporal sequence of linked cases within 2 subclusters (eg, J-F and B-L); this could be explained by the presence of missing cases (eg, in subcluster J-F, 3 additional nongenotyped cases occurred in the same family), timing of transmission relative to specimen collection, or mixed infection with multiple strains.

### Phylogeny

WTK isolates from both Thailand and California were closely related (fixation index, 0.027) with a median of 10 SNPs (range, 0–20 SNPs) differentiating strains (Supplementary Figure 1 and Table 2). In a sensitivity analysis, the fixation index did not significantly differ according to whether the 5 imported MDR tuberculosis cases were considered in the California or Thailand group. Moreover, the California strains showed a higher genetic diversification, compared with the Thai strains ($\pi$ [±SD], 0.07 ± 0.005 vs 0.05 ± 0.005), suggesting multiple importation

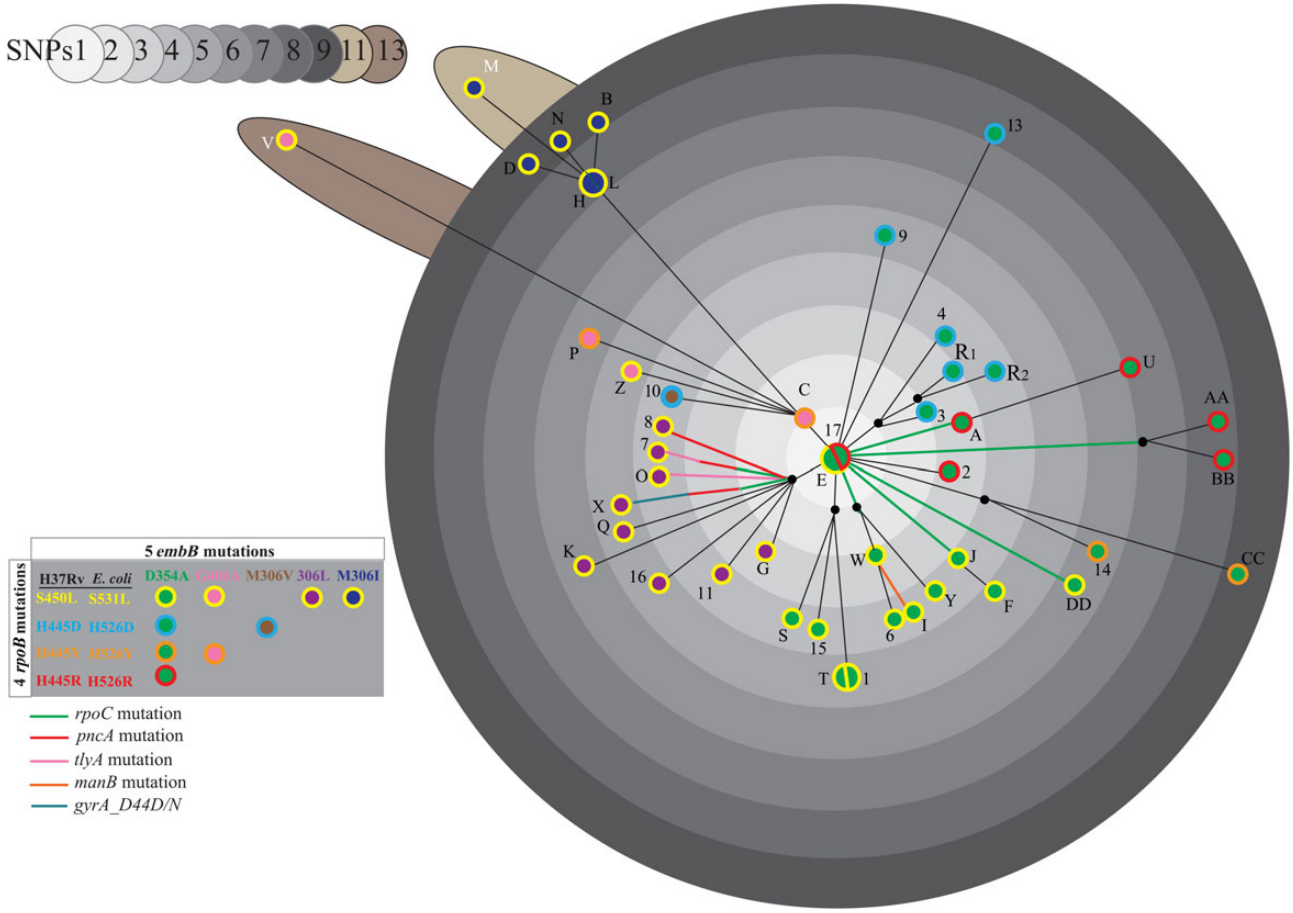


**Figure 2.** Median joining network with mapping of drug-resistance mutations. The relationships of 46 Wat Tham Krabok (WTK) isolates were determined using 150 variable single-nucleotide polymorphisms (SNPs), as described in Figure 1*B*. Isolates are coded according to *embB* (fill color) and *rpoB* (border color) mutation; other drug resistance and putative compensatory mutations are indicated in branches. Note that strains with matching drug resistance and/or compensatory mutations are clustered together. Shading indicates the relative SNP difference with respect to the central node.

events followed by the establishment and evolution of multiple WTK clones within California. This is further supported by the temporal appearance of the WTK strain in California and the finding that certain combinations of DRMs and compensatory mutations (eg, identical sets of *embB* resistance–conferring mutations) mapped exclusively to particular subclusters of the network (Figure 2).

We found that the WTK strain was separated from all other lineage 2 strains sequenced to date and was defined by 55 specific SNPs present in all WTK isolates but absent in other lineage 2 strains (Supplementary Figure 1 and Table 2). Fourteen of these SNPs were found in intergenic regions, and 41 were found in coding regions, 24 (59%) of which were nonsynonymous (Table 2). Additionally, one intergenic SNP (between Rv0278c and Rv0279c) was homoplastic. All strains harbored genomic deletions previously described to be associated with lineage 2 strains (RD105, RD207, RD181, RD149, and RD152) [17], although only WTK strains harbored an additional deletion affecting the genetic locus Rv1267c.

### Drug Resistance and Compensatory Mutations

High-confidence DRMs corroborated phenotypic drug-susceptibility test results, indicating resistance to isoniazid (*katG* S315T), rifampin (*rpoB* H526D/Y/R and S531L in 1 subset each), ethambutol (*embB* [A378E in all and M306L/V/I, D354A, and G406A in 1 subset each); *Rv3806c* D149E]), and streptomycin (*rpsL* K43R) were found (Table 1 and Supplementary Tables 1 and 3) [14]. An identical *katG* mutation in conjunction with differing *rpoB* mutations indicates that a progenitor of the WTK strain was likely isoniazid resistant but not MDR. Additional pyrazinamide (*pncA*) and capreomycin (*tlyA*) DRMs were present in subsets of isolates. Following misdiagnosis and known fluoroquinolone exposure, patient X was found to have extensively drug-resistant tuberculosis with an isolate demonstrating heteroresistance to fluoroquinolones (both wild type and mutation *gyrA* D94N were detected). Four possible compensatory mutations in *rpoC* (Supplementary Table 3) were found in strains with *rpoB* S531L:V775M (patients 6, W, and I), F831L (patient X), W484G (patient 7), and P309S (patients F and J; Figure 2). In

**Table 2. Nonsynonymous Mutations Present in all Wat Tham Krabok (WTK) Strains but Not Present in Other Lineage 2 Strains From the Global Collection**

| Reference Base | Mutant Base | Locus[a] | Amino Acid Change[a] | Genomic Position[a] | Locus Name[a] |
|---|---|---|---|---|---|
| C | G | Rv0132c | A212P | 160 149 | fgd2 |
| C | G | Rv0592 | P464R | 691 891 | mce2D |
| A | G | Rv0614 | I168V | 709 857 | . . . |
| G | C | Rv0663 | R290P | 757 005 | atsD |
| C | G | Rv1166 | T402R | 1 297 356 | lpqW |
| T | C | Rv1524 | F66L | 1 718 921 | . . . |
| G | T | Rv1643 | A123S | 1 853 550 | rplT |
| G | C | Rv1784 | A140P | 2 021 051 | . . . |
| C | T | Rv1785c | A58T | 2 024 457 | cyp143 |
| C | G | Rv1813c | A124P | 2 055 743 | . . . |
| T | C | Rv1871c | N71S | 2 121 673 | . . . |
| T | C | Rv2520c | N74S | 2 837 395 | . . . |
| C | G | Rv2571c | A212P | 2 895 327 | . . . |
| A | G | Rv2601 | I72V | 2 928 601 | speE |
| G | A | Rv2700 | R24Q | 3 015 273 | . . . |
| A | G | Rv2702 | K197E | 3 017 446 | ppgK |
| C | G | Rv2763c | D70H | 3 073 402 | dfrA |
| G | A | Rv2834c | T269I | 3 140 509 | ugpE |
| C | G | Rv3201c | V399L | 3 575 842 | . . . |
| G | A | Rv3308 | G233D | 3 695 561 | pmmB |
| G | A | Rv3415c | A16V | 3 834 475 | . . . |
| G | T | Rv3447c | S1233R | 3 864 540 | . . . |
| G | C | Rv3596c | L473V | 4 039 288 | clpC1 |
| T | C | Rv3735 | F87S | 4 186 348 | . . . |

Data are from [16].

[a] Refer to the *Mycobacterium tuberculosis* H37Rv genome.

contrast, only a single *rpoC* mutation (S561P in patient 17, subclusters A-U and AA-BB) was associated with *rpoB* H526R.

In addition to 26 high-confidence DRMs (Table 1 and Supplementary Table 3), the WTK strain harbored 17 nonsynonymous or intergenic mutations recently proposed to be associated with multidrug resistance [18]. However, most of these mutations (12 of 17) have been previously identified as phylogenetic markers (6 SNPs are associated with lineage 2, 4 SNPs are associated with sublineage 2, and 2 SNPs are associated with lineages 2, 3, and 4) [16] and are therefore unlikely to have a causal association with drug resistance. Of the 5 remaining mutations, 1 each was distributed among 5 WTK strains, indicating a nonessential role in the propagation of the WTK strain.

## DISCUSSION

We used next-generation sequencing data to delineate the longitudinal clonal expansion of a lineage 2/Beijing MDR strain of *M. tuberculosis* among persons of Hmong ethnicity emigrating from a Southeast Asian setting with a high burden of MDR tuberculosis to the United States. We found that the domestic MDR tuberculosis rate among Hmong persons was >3 times that of the general population in California, a situation facilitated by poverty and social isolation following resettlement in the United States [19, 20]. Genomic data provided evidence against what was previously thought to be an MDR tuberculosis outbreak in Thailand, indicated a central role for specific individuals in the establishment of the WTK strain in California, and confirmed multiple importation and subsequent reactivation events over a 22-year period.

Contact investigations aim to identify cases of active and latent *M. tuberculosis* infection among contacts in order to institute effective preventive therapy. This effort has been supplemented by *M. tuberculosis* genotyping based on mobile and repetitive genetic elements for >2 decades. However, conventional genotyping techniques examine <1% of the *M. tuberculosis* genome and are often insufficiently specific in outbreak situations, owing to a rate of change (the so-called molecular clock) that is slower than the rate of ongoing transmission and pathogenesis [21]. In contrast, WGS provides high-resolution molecular mapping of *M. tuberculosis* that can identify short-term transmission events, even in the context of highly prevalent strains.

In our study, genomic data were decisive in resolving a putative MDR tuberculosis outbreak in Thailand into multiple allelic variants of a highly prevalent strain. Despite extensive contact among cases and identical conventional genotypes [12], most cases were not related within recent transmission chains. Strains from lineage 2 (the East Asian lineage, which includes the Beijing family of strains) are associated with an increased risk of drug resistance [22–24] and have been found to account for the majority of MDR tuberculosis cases in monophyletic fashion in other settings [8]. In California, a setting with a lower tuberculosis burden examined over a longer interval, conventional genotyping was sufficient to distinguish the WTK strain from other MDR strains and to discern relatedness to the 2004 investigation in Thailand. High-resolution molecular techniques were necessary, however, to resolve short-term transmission chains, distinguish relapse from reinfection, and identify the central role of a potential super spreader in transmitting the WTK strain within California.

Interrogation of the complete *M. tuberculosis* genome is also advantageous in that it may identify genetic markers that explain phenotypic consequences. We identified a deletion of Rv1267c (*EmbR*) within the WTK strain that may simultaneously directly confer ethambutol resistance through mutations in the kinase-interacting domain of EmbR [25, 26] and alter the ability of the host to mount an efficient immune response through functional changes in the ratio of lipoarabinomannan (LAM) to lipomannan (LM) [26, 27]. The LAM/LM ratio has been associated with mycobacterial virulence, phagosome maturation, [28] apoptosis [29], and interferon signaling [30] in

macrophages and with interleukin 12 cytokine secretion by dendritic cells, all of which result in increased fitness [31].

Previous work in several bacteria, including *M. tuberculosis* complex, has shown that mutations in *rpoC* can compensate for the fitness defects associated with mutations in *rpoB* that confer resistance to rifampin [10, 32, 33]. In *M. tuberculosis*, these *rpoC* mutations have been strongly associated with the *rpoB* S531L mutation, which is the most frequent mutation in rifampin-resistant clinical strains [10] and associated with a minor fitness defect in *M. tuberculosis* [34]. Hence, many different *rpoC* mutations seem to be able to compensate for the fitness defect associated with *rpoB* S531L. Our study supports this view, since 4 of 5 *rpoC* mutations that we identified were found in strains carrying *rpoB* S531L. Interestingly, an alternate *rpoC* mutation (S561P) was strongly associated with *rpoB* H526R. *rpoB* H526R has been shown to have a much greater fitness cost than *rpoB* S531L [34]. The fact that we found only a single *rpoC* mutation linked to *rpoB* H526R suggests that compensation is somehow restricted in mutants carrying *rpoB* H526R, perhaps because the deleterious effect on fitness is stronger and therefore more difficult to compensate. More work is needed however to confirm this hypothesis.

Positive selection in strains exhibiting increasing levels of drug resistance map almost entirely to drug resistance–associated genes, and for drug-resistant *M. tuberculosis* (as in other microbes) [35], the molecular antibiogram (ie, the collection of detected DRMs) has been used to corroborate other genotypic information in inferring chains of transmission [36]. In our study, *embB* resistance–conferring mutations were strikingly congruent within network subclusters. However, our study was not powered to examine the utility of DRMs as phylogenetic markers, and homoplasy would likely be a limiting factor for analyses undertaken on a broader scale [37].

Our study has potential limitations. First, estimates of clustering are typically based on a nonrandom sample of cases in a given community, and inference of transmission chains is subject to the same limitations as conventional genotyping with regard to sampling fraction and cluster size [21]. In Thailand, and prior to 2004 in California, there was incomplete sampling of the population base. Thus, missed isolates identical to or highly similar to that infecting the putative super spreader (patient E) might have provided an alternate explanation for some of the transmission of the WTK strain in California. Second, information on epidemiologic links was abstracted retrospectively. Although detailed contact investigation data were often available, in particular for California cases, collecting additional information prospectively was not feasible. Thus, some epidemiologic linkages and subsequent relationships between patients might be missed, including the identification of alternate source cases. Third, we avoided application of strict SNP cut points in inferring direct transmission. Accurate estimation of transmission trees relevant to public health practice will continue to require a context of conventional epidemiologic information and a nuanced approach to SNP differences. This approach must account for within-person genetic variability (ie, pathogen variability during the same tuberculosis episode) and between-person genetic variability (ie, pathogen variability during sequential transmission events) due to microevolution [38], mixed-strain infection [39], and heteroresistance [40]. Further, because SNP variability is heterogeneous in tempo across the full genome [41], accurate measurement of a molecular clock will require calibration of SNP changes to the gene regions they occupy.

In conclusion, WGS has epidemiologic added value in low- and high-burden settings and aids our understanding of the transcontinental dispersion and transmission of MDR *M. tuberculosis*. Used in real time, WGS may have alerted public health authorities to the presence of missing cases in chains of ongoing transmission or unknown sites of transmission in both Thailand and California. However, overall improvements to tuberculosis control or patient-important outcomes, along with questions of cost-benefit in low-burden settings, remain to be determined.

## Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

## Notes

***Disclaimer.*** The authors alone are responsible for the views expressed in this publication, and they do not necessarily represent the decisions or policies of the Centers for Disease Control and Prevention.

***Financial support.*** This work was supported by the National Institutes of Health (grants K23 AI094251 [to J. Z. M.] and R01 AI090928 [to S. G.]), the Robert Wood Johnson Foundation (Amos Medical Faculty Development Program award to J. Z. M.), the Swiss National Science Foundation (PP00P3_150750 to S. G.), and the European Research Council (309540-EVODRTB to S. G.).

***Potential conflicts of interest.*** All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

## References

1. Centers for Disease Control and Prevention. Reported tuberculosis in the United States, **2012**. http://www.cdc.gov/tb/statistics/reports/2012/pdf/report2012.pdf. Accessed 11 December 2013.
2. Schwartzman K, Oxlade O, Barr RG, et al. Domestic returns from investment in the control of tuberculosis in other countries. N Engl J Med **2005**; 353:1008–20.
3. Schurch AC, Kremer K, Daviena O, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. J Clin Microbiol **2010**; 48:3403–6.
4. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med **2011**; 364:730–9.

5. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **2013**; 13:137–46.
6. Roetzer A, Diel R, Kohl TA, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **2013**; 10:e1001387.
7. Kato-Maeda M, Ho C, Passarelli B, et al. Use of whole genome sequencing to determine the microevolution of *Mycobacterium tuberculosis* during an outbreak. PLoS One **2013**; 8:e58235.
8. Casali N, Nikolayevskyy V, Balabanova Y, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet **2014**; 46:279–86.
9. Safi H, Lingaraju S, Amin A, et al. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes. Nat Genet **2013**; 45:1190–7.
10. Comas I, Borrell S, Roetzer A, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet **2012**; 44:106–10.
11. Long R, Nobert E, Chomyc S, et al. Transcontinental spread of multidrug-resistant Mycobacterium bovis. Am J Respir Crit Care Med **1999**; 159:2014–7.
12. Oeltmann JE, Varma JK, Ortega L, et al. Multidrug-resistant tuberculosis outbreak among US-bound Hmong refugees, Thailand, 2005. Emerg Infect Dis **2008**; 14:1715–21.
13. Lowenthal P, Westenhouse J, Moore M, Posey DL, Watt JP, Flood J. Reduced importation of tuberculosis after the implementation of an enhanced pre-immigration screening protocol. Int J Tuberc Lung Dis **2011**; 15:761–6.
14. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med **2009**; 6:e2.
15. National Tuberculosis Controllers Association, Centers for Disease Control and Prevention. Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. MMWR Recomm Rep **2005**; 54:1–47.
16. Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet **2013**; 45:1176–82.
17. Tsolaki AG, Hirsh AE, DeRiemer K, et al. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. Proc Natl Acad Sci U S A **2004**; 101:4865–70.
18. Zhang H, Li D, Zhao L, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet **2013**; 45:1255–60.
19. Leigh Brown P. A Hmong generation finds its voice in writing. New York Times 31 December 2011. http://www.nytimes.com/2012/01/01/us/a-hmong-generation-finds-its-voice-in-writing.html?adxnnl=1&pagewanted=all&adxnnlx=1388281090-lpvZQbK691t1oUtDgHMZwQ. Accessed 28 December 2013.
20. Fadiman A. The spirit catches you and you fall down: a Hmong child, her American doctors, and the collision of two cultures. New York: Farrar, Straus and Giroux, **1997**.
21. Glynn JR, Vynnycky E, Fine PE. Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. Am J Epidemiol **1999**; 149:366–71.
22. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis. Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance. Emerg Infect Dis **2006**; 12:736–43.
23. Borrell S, Gagneux S. Infectiousness, reproductive fitness and evolution of drug-resistant *Mycobacterium tuberculosis*. Int J Tuberc Lung Dis **2009**; 13:1456–66.
24. Ford CB, Shah RR, Maeda MK, et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet **2013**; 45:784–90.
25. Ramaswamy SV, Amin AG, Goksel S, et al. Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother **2000**; 44:326–36.
26. Sharma K, Gupta M, Pathak M, et al. Transcriptional control of the mycobacterial embCAB operon by PknH through a regulatory protein, EmbR, in vivo. J Bacteriol **2006**; 188:2936–44.
27. Zhang N, Torrelles JB, McNeil MR, et al. The Emb proteins of mycobacteria direct arabinosylation of lipoarabinomannan and arabinogalactan via an N-terminal recognition region and a C-terminal synthetic region. Mol Microbiol **2003**; 50:69–76.
28. Fratti RA, Chua J, Vergne I, Deretic V. *Mycobacterium tuberculosis* glycosylated phosphatidylinositol causes phagosome maturation arrest. Proc Natl Acad Sci U S A **2003**; 100:5437–42.
29. Rojas M, Garcia LF, Nigou J, Puzo G, Olivier M. Mannosylated lipoarabinomannan antagonizes *Mycobacterium tuberculosis*-induced macrophage apoptosis by altering Ca+2-dependent cell signaling. J Infect Dis **2000**; 182:240–51.
30. Hmama Z, Gabathuler R, Jefferies WA, de Jong G, Reiner NE. Attenuation of HLA-DR expression by mononuclear phagocytes infected with *Mycobacterium tuberculosis* is related to intracellular sequestration of immature class II heterodimers. J Immunol **1998**; 161: 4882–93.
31. Keane J, Remold HG, Kornfeld H. Virulent *Mycobacterium tuberculosis* strains evade apoptosis of infected alveolar macrophages. J Immunol **2000**; 164:2016–20.
32. Brandis G, Hughes D. Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. J Antimicrob Chemother **2013**; 68:2493–7.
33. Brandis G, Wrande M, Liljas L, Hughes D. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. Mol Microbiol **2012**; 85:142–51.
34. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJ. The competitive cost of antibiotic resistance in *Mycobacterium tuberculosis*. Science **2006**; 312:1944–6.
35. Harris SR, Cartwright EJ, Torok ME, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis **2013**; 13:130–6.
36. Clark TG, Mallard K, Coll F, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One **2013**; 8:e83012.
37. Ioerger TR, Feng Y, Chen X, et al. The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa. BMC Genomics **2010**; 11:670.
38. Perez-Lago L, Comas I, Navarro Y, et al. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. J Infect Dis **2014**; 209:98–108.
39. Merker M, Kohl TA, Roetzer A, et al. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. PLoS One **2013**; 8:e82551.
40. Rinder H. Hetero-resistance: an under-recognised confounder in diagnosis and therapy? J Med Microbiol **2001**; 50:1018–20.
41. Comas I, Chakravartti J, Small PM, et al. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat Genet **2010**; 42:498–503.