Genomic Analysis of *Mycobacterium tuberculosis* Isolates and Construction of a Beijing Lineage Reference Genome

Woei-Fuh Wang[1,2], Mei-Yeh Jade Lu[1], Ting-Jen Rachel Cheng[3], Yi-Ching Tang[1], Yu-Chuan Teng[1], Teh-Yang Hwa[1], Yi-Hua Chen[1], Meng-Yun Li[1], Mei-Hua Wu[4], Pei-Chun Chuang[4], Ruwen Jou[4,*], Chi-Huey Wong[3,*], and Wen-Hsiung Li[1,5,*]

[1]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan 11529

[2]Center for Precision Medicine, China Medical University Hospital, Taichung, Taiwan 40447

[3]Genome Research Center, Academia Sinica, Taipei, Taiwan 11529

[4]Tuberculosis Research Center, Centers for Disease Control, Taipei, Taiwan 11561

[5]Department of Ecology and Evolution, University of Chicago, Illinois 60637

*Corresponding authors: E-mails: rwj@cdc.gov.tw; chwong@gate.sinica.edu.tw; whli@sinica.edu.tw.

Accepted: January 14, 2020

Data deposition: This project has been deposited in the NCBI SRA under the accession PRJNA505382.

Abstract

Tuberculosis (TB), an infectious disease caused by *Mycobacterium tuberculosis*, kills over 1 million people worldwide annually. Development of drug resistance (DR) in the pathogen is a major challenge for TB control. We conducted whole-genome analysis of seven Taiwan *M. tuberculosis* isolates: One drug susceptible (DS) and five DR Beijing lineage isolates and one DR Euro-American lineage isolate. Developing a new method for DR mutation identification and applying it to the next-generation sequencing (NGS) data from the 6 Beijing lineage isolates, we identified 13 known and 6 candidate DR mutations and provided experimental support for 4 of them. We assembled the genomes of one DS and two DR Beijing lineage isolates and the Euro-American lineage isolate using NGS data. Moreover, using both PacBio and NGS sequencing data, we obtained a high-quality assembly of an extensive DR Beijing lineage isolate. Comparative analysis of these five newly assembled genomes and two published complete genomes revealed a large number of genetic changes, including gene gains and losses, indels and translocations, suggesting rapid evolution of *M. tuberculosis*. We found the MazEF toxin–antitoxin system in all the seven isolates studied and several interesting mutations in MazEF proteins. Finally, we used the four assembled Beijing lineage genomes to construct a high-quality Beijing lineage reference genome that is DS and contains all the genes in the four genomes. It contains 212 genes not found in the standard reference H37Rv, which is Euro-American. It is therefore a better reference than H37Rv for the Beijing lineage, the predominant lineage in Asia.

Key words: *Mycobacterium tuberculosis*, drug resistance, TB genomes, Beijing lineage reference.

Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, is a major infectious killer worldwide. The World Health Organization (WHO) estimated that there were ~10 million new TB cases and ~1.3 million TB deaths (HIV-negative) in 2017 (World Health Organization 2018). A major challenge of TB control is the development of drug resistance (DR) in the pathogen. Multi-DR (MDR) is defined as resistance to at least isoniazid and rifampicin, whereas extensive DR (XDR) is MDR with additional resistance to any of fluoroquinolone (such as ofloxacin, levofloxacin, or moxifloxacin) and any of the injectable drugs (i.e., amikacin, kanamycin, or capreomycin) (World Health Organization 2018). According to WHO's estimate (World Health Organization 2018), globally ~460,000 people developed MDR-TB in 2017 and ~8.5% of these cases developed into XDR-TB. Mutations in the genes targeted by anti-TB drugs are the major causes of DR (Coll et al. 2015). Therefore, identification of the mutations associated with DR of TB is helpful for understanding the cause of DR, for adequate treatments, and for development of new diagnostic tests and anti-TB drugs.

To identify DR mutations, we analyzed the genomes of six Beijing lineage *M. tuberculosis* isolates and one Euro-American lineage isolate collected in Taiwan with different DR profiles. These isolates were deposited at the Taiwan Centers for Disease Control (TCDC) and are here named as TCDC1, TCDC3, TCDC4, TCDC5, TCDC7, TCDC10, and TCDC11. We developed a simple method to identify DR mutations using NGS data from the six Beijing lineage isolates. Functional assays were then conducted to study four of the mutations identified.

The H37Rv genome has been used as a standard reference genome for detecting mutations. However, it is Euro-American, so it is not closely related to Beijing lineage isolates and may miss many genes that are Beijing lineage-specific. It is therefore desirable to construct a Beijing lineage reference. For this purpose, we sequenced and assembled the genomes of four Beijing lineage isolates (TCDC1, TCDC7, TCDC10, and TCDC11) and also one Euro-American isolate (TCDC3), using multiple next-generation sequencing (NGS) platforms. Moreover, for TCDC11, an XDR, we also used PacBio RS II to achieve a high-quality assembly. We used the assembled genomes of TCDC11 and the three other Beijing lineage isolates (TCDC1, TCDC7, and TCDC10) to construct a "Beijing lineage reference genome," denoted BLrg. Like H37Rv, BLrg is drug susceptible (DS). Moreover, it contains all of the genes in TCDC1, TCDC7, TCDC10, and TCDC11. This reference is likely to be very useful because the Beijing lineage is the most prevalent TB lineage in Asia (Reed et al. 2007; Rodríguez-Castillo et al. 2017; Zaychikova et al. 2018).

Finally, we conducted a comparative genomics analysis of the five newly assembled genomes, H37Rv and KZN605 to detect the genetic changes such as gene gains/losses, insertions and deletions (indels), and translocations among these genomes. This analysis provides information on how frequently such genetic changes occur during the evolution of *M. tuberculosis*.

Materials and Methods

Drug Susceptibility Testing and Spoligotyping

The *M. tuberculosis* strains studied were isolated from patients during 2010–2013. TCDC1 was DS (supplementary table S1, Supplementary Material online), which was used as a reference. TCDC11 was an XDR, which is rare in Taiwan. The other five isolates were MDRs. Spoligotyping was applied to genotype in all seven isolates (Jou et al. 2005). TCDC3 belonged to the Euro-American lineage, whereas the other six isolates belonged the Beijing lineage, which is the dominant lineage in Taiwan (58%). The six Beijing lineage isolates were again subject to the drug susceptibility test as in Chuang et al. (2016) (supplementary table S1, Supplementary Material online).

DNA Extraction and NGS Library Preparation

To obtain genomic DNA, fresh cultures of seven *M. tuberculosis* isolates were harvested at the Taiwan CDC and subjected to CTAB-based DNA extraction. Briefly, the cultures were harvested by centrifugation and heat inactivation. The cell pellets were homogenized by grinder, treated with lysozyme and RNaseA at 37 °C for 2 h, and subjected to lysis by adding SDS and proteinase K with overnight incubation at 56 °C. The DNA extraction was then carried out by adding CTAB/NaCl and incubating at 65 °C for 10 min with occasional agitation, followed by stepwise extraction with chloroform/isoamyl alcohol (24:1), phenol/chloroform/isoamyl alcohol (25:24:1), and finally with chloroform extraction. Nucleic acids were precipitated by adding isopropanol, and pellets were washed with 70% EtOH. The samples were resuspended and further treated with RNase A to degrade the remaining RNA, and purified by PCI extraction and EtOH precipitation. The DNA samples were then resuspended in 10 mM Tris (pH 8.0) and subjected to quantification and gel electrophoresis. The chromosomal DNA integrity was generally good, although different extents of smear were observed in some isolates due to the reduced fitness of these isolates.

Whole-Genome Sequencing Using NGS

To achieve high-quality genome assembly, the Roche 454 GS+ long-read platform was applied for de novo assembly, and the Illumina short-read platform for scaffolding and for error correction. For long-read sequencing, the *M. tuberculosis* genomic DNA was sheared by nebulization and ligated with barcoded 454 adaptors, followed by library construction using GS FLX Titanium Rapid Library Preparation Kit (Roche 454) and gel size selection to purify >1 kb fragments. Sequencing was done using GS FLX Titanium Sequencing Kit XL+.

For scaffolding and assembly error correction, genomic DNA was sheared using Covaris. Libraries were constructed using KAPA Library Prep Kits for Illumina NGS Platform (Roche) in conjunction with TruSeq single-index adaptors (Illumina). The paired-end libraries were subjected to gel size selection before being optimized with PCR amplification. All libraries were quantified by Qubit DNA HS DNA assay (ThermoFisher), profiles checked by BioAnalyzer (Agilent), and molar concentration normalized using KAPA NGS library qPCR kit (Roche). Paired-end sequencing was conducted on Illumina HiSeq2500.

Finally, in order to assemble a complete genome for the XDR strain TCDC11, a PacBio library of TCDC11 DNA was constructed with a gel size selection cutoff at 8 kb and sequenced using PacBio RS II.

Supplementary table S2*a*–*c*, Supplementary Material online shows the sequencing depth and read length of Illumina reads. Roche 454 data are shown in supplementary table S2*d*, Supplementary Material online, and PacBio data for

TCDC11 are shown in supplementary table S2e, Supplementary Material online.

Data Preprocessing and Reference Genome Mapping

The Illumina HiSeq2500 sequencing reads with Phred quality score lower than 15 and adapter sequences were clipped by Trimmomatic 0.32 (Bolger et al. 2014). The pruned sequencing reads shorter than 36 bp were discarded. Only the sequencing pairs with both reads that remained after the trimming were aligned to the reference genome H37Rv (NC_000962.2) (Cole et al. 1998) by aligner BWA 0.6.2 (Li and Durbin 2009). The mapped genome coverage of the sequenced *M. tuberculosis* isolates was between 94 and 99%. Those reads with mapping quality MAPQ <20 were discarded.

Single Nucleotide Polymorphism Calling

All mutations were identified by SAMtools 0.1.18 (Li et al. 2009). Only the mutations with the read depth above five reads and with mapping quality score ≥20 were considered true mutations. We then examined the functional effects of the mutations by SnpEff-3.3b (Cingolani et al. 2012).

Identification of DR Mutations

Genome-wide association study has been widely used to investigate disease-associated variants by analyzing whole-genome variations. It was applied to identify the genes and intergenic regions associated with resistance to anti-TB drugs (Zhang et al. 2013). However, the study focused on genomic regions associated with DR, but not specific DR mutations. Here, we propose two new rules to identify DR mutations (fig. 1): Rule 1: If a mutation is found in any DS isolate(s), it is not a DR mutation. This rule says that any mutation that is found in one or more drug-susceptible isolates cannot be a DR mutation even it is also found in a DR isolate. For example, mutations m_1 and m_2 in figure 1 are found in DR and/or DS isolates, so they are not candidate DR mutations. Rule 2: If the DR of an isolate can be explained by a known DR mutation, then no other mutation is considered a candidate DR mutation, unless it is also a known DR mutation. For example, mutations m_3 and m_5 are only found in drug-resistant isolates, so they are candidate DR mutations. On the other hand, although mutation m_4 is only found in isolate D, which is resistant to the drug under study, it is not considered a candidate DR mutation, because m_6^K, a known DR mutation, is also found in isolate D.

Note that the determination of DR mutations requires accurate drug susceptibility tests of all isolates under study and that the accuracy of our method depends on the number of isolates included in the study.

For simplicity, we focused on the known drug-resistant genes, although this approach may miss some DR mutations.

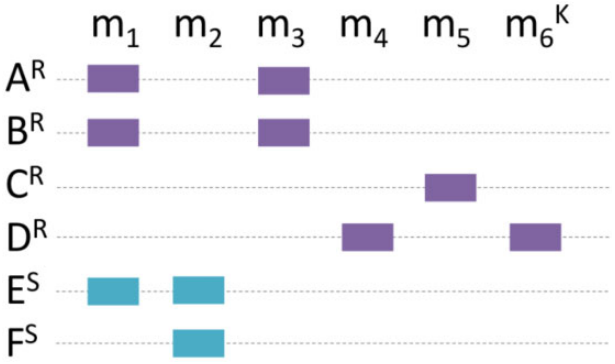

Fig. 1.—Hypothetical examples that include both drug-susceptible and -resistant isolates. Letters A–F denote the *M. tuberculosis* isolates studied. The superscripts R and S denote "resistant" and "susceptible," respectively. m_1, m_2, . . ., and m_6^K denote six different mutations; K denotes "known." A solid rectangle means the presence of the mutation. The purple and blue colors indicate that the mutation is found in a resistant or a susceptible isolate, respectively.

For each known DR gene, we considered mutations in the promoter regions and nonsynonymous mutations in the coding regions.

Functional Assays

We selected four identified DR mutations to conduct functional assays.

Binding of Aminoglycosides to RNA

Analyses were done using the BIAcore T200 instrument, and data were analyzed using BIA evaluation. Biotinylated RNA solutions (5′-biotin-GGCGUCA/GCGUUCGCGAAGUCGCC-3′) at 100 nM were renatured by heating to 85 °C and slowly cooled to ambient temperature. The solutions were adjusted to 1 M NaCl, 0.5× HBS before immobilization on the SA sensorchip (GE Healthcare) at a low rate of 10 µl/min for 5 min. The running buffer used for all experiments was HBS buffer, pH 7.4, containing 10 mM HEPES, 150 mM NaCl, 3 mM EDTA, and 0.005% surfactant P20; the buffer was filtered and degassed. Kanamycin was prepared by serial dilutions from stock solutions in RNase-free microfuge.

Measurement of Pyrazinamidase Activity

PncA gene was amplified from *M. tuberculosis* genomic DNA and the recombinant protein, pyrazinamidase (PZase) with an N-terminal $(His)_6$ tag was prepared from *Escherichia coli* cells following standard procedure. The purified PZase was incubated with pyrazinamide (0–100 mM) at 37 °C for 30 min. The kinetics parameters of the enzyme were determined by measuring the produced pyrazinoic acid (POA) using optical density at 480 nm after adding 0.05 volume of 500 mM ferrous ammonium sulfate.

Measurement of Gyrase Activity

GyrA and *gyrB* genes were amplified from *M. tuberculosis* genomic DNA and the recombinant proteins with an N-terminal $(His)_6$ tag were prepared from *E. coli* cells following standard procedures. The enzyme reactions contained 100 mM Tris–HCl (pH 7.6), 6 mM $MgCl_2$, 20 mM KCl, 1% dimethyl sulfoxide, 0.05 mg/ml of bovine serum albumin, 5 μg/ml PBR322, 700 μM ATP, and 2 μM Gyrase B or Topoisomerase ($GyrA_2GyrB_2$). ATP hydrolysis during the reaction was monitored by determining the increase in ADP using ADPGlo (Promega, Inc.).

Phylogenetic Analysis

A phylogenetic tree consisting of 85 global and 7 Taiwan *M. tuberculosis* isolates was constructed from the 58,161 single nucleotide mutations with respect to H37Rv, using the Maximum likelihood method with the Tamura-Nei model in Mega X (Kumar et al. 2018); 1,000 bootstraps were conducted. The 85 global isolates included 10 Japanese isolates and 10 Philippines isolates from NCBI and 65 isolates that were randomly selected from the PATRIC collection (Wattam et al. 2017). These 85 selected isolates include 50, 30, and 5 isolates in Asia, Eur-America, and Africa, respectively. The mutations in these and in the seven Taiwan isolates were identified by the same mutation-calling pipeline using the H37Rv genome as the reference and concatenated for phylogenetic analysis.

De Novo Genome Assembly

The whole-genome sequencing and assembly of TCDC1, TCDC3, TCDC7, TCDC10, and TCDC11 were performed using various NGS platforms available in-house and suitable algorithms. TCDC1, TCDC3, and TCDC10 were sequenced using Illumina HiSeq and MiSeq and Roche 454 (supplementary table S2, Supplementary Material online) and assembled with the workflow shown in supplementary figure S4*a*–*b*, Supplementary Material online. For each genome, Roche 454 reads were first used to assemble contigs using Newbler (Margulies et al. 2005), and Illumina pair-end sequencing reads were applied to obtain the scaffolds using SSPACE (Boetzer et al. 2011). The scaffolds representing the IS6110 insertion repeats in the TCDC3 and TCDC10 genomes were identified by aligning them to the reference IS6110 sequence. The scaffolds identified as IS6110 repeats were bridged by Illumina pair-end reads to the neighboring scaffolds grouped in the SSPACE contig graph. Next, Illumina Nextera mate-pair sequencing was applied to TCDC3 and TCDC10 to obtain long-jump data with designated gel size selection ranges for up to 10-15 kb (supplementary fig. S4*b*, Supplementary Material online). The mate-pair data sets with increasing jump distance were then used sequentially to concatenate scaffolds stepwise. Finally, Gapcloser (Luo et al. 2012) and Gapfiller (Nadalin et al. 2012) were used to close gaps within scaffolds.

TCDC7 was selected as the first target of a complete genome assembly to represent an MDR Beijing isolate. Roche 454 FLX reads (supplementary table S2*d*, Supplementary Material online) were used for contig assembly using Newbler (supplementary fig. S4*c*, Supplementary Material online), reaching a draft assembly of contigs broken by the repeat sequences, including the high occurrence of repeat IS6110. To overcome this problem, we used reference-guided assembly to resolve the Newbler contig graph. Based on the mapping coverage, the copy numbers of the repeat contigs were adjusted based on their read abundance relative to the average genome coverage. The set of new contigs were then aligned to the five published TB genomes (H37Ra, H37Rv, CDC1551, F11, and KZN1435) by Nucmer (Kurtz et al. 2004) to determine the contig syntenies. To guide the genome alignment, the syntenic consensus among all reference genomes was applied to order the TCDC7 contigs. If it was difficult to identify the consensus synteny among the reference genomes, F11 was used to guide contig orders because TCDC7 was found most similar to F11 based on the read mapping. The high GC content of the *M. tuberculosis* genome and the associated homopolymeric regions posed an additional challenge for genome assembly. Therefore, we used the great depths of Illumina HiSeq pair-end data (supplementary table S2*a*, Supplementary Material online) to conduct two rounds of homopolymeric error corrections of the TCDC7 assembly.

TCDC11 was sequenced using not only Illumina HiSeq2500, but also PacBio RSII (supplementary table S2*e*, Supplementary Material online). The draft assembly was obtained with the Hierarchical Genome Assembly Process (Chin et al. 2013) of the PacBio data, and subsequently the putative sequencing errors were corrected by a two-step base calling correction procedure. The first step was conducted using Quiver (Chin et al. 2013) with PacBio subreads. The second step used the Illumina paired-end reads that had been cleaned by removing adaptors and low quality bases and employed the genome assembly improvement tool, Pilon (Walker et al. 2014), to achieve the final assembly (supplementary fig. S4*d*, Supplementary Material online).

Gene Annotation

Gene annotation was done using Prokka (Seemann 2014) with the help of third-party tools: BLAST+ 2.2.8 (Camacho et al. 2009), HMMER (Eddy 1998), Aragorn (Laslett and Canback 2004), Prodigal (Hyatt et al. 2010), Infernal (Nawrocki and Eddy 2013), and RNAmmer (Lagesen et al. 2007). Most prokaryotic gene prediction tools use common prokaryotic gene features to annotate gene models, including adopting the bacterial, archaeal, and plant plastic genetic codes as the bacterial genetic code. In order to check the accuracy of Prokka annotation, we applied Prokka to predict

Gene	Isoniazid	Rifampicin	Pyrazinamide	Streptomycin	Ofloxacin	2nd-line injectables
inhA	S94A					
katG	R249H S315T A479E					
ndh	I68T					
Rv1592c	I332F					
rpoB		D435V S450L S450W				
rpoC		V483A P1040R				
pncA			G78V V139A			
rpsL				K43R		
gyrA					G88C A90V D94Y	
gyrB					K247N	
rrs						A1401G

Fig. 2.—Amino acid changes in identified DR mutations. Green and red denote, respectively, known DR mutations and candidate novel DR mutations identified in this study.

genes of the H37Rv genome. Supplementary table S3, Supplementary Material online shows that the usages of alternative start codons in 4% of the gene models were misannotated by PROKKA for the H37Rv genome. Alternative start codons (ATG, GTG, and TTG) affect the predicted gene length. For instance, *gyrB*, a known drug-resistant gene, has the start codon GTG in the H37Rv genome annotation, but it was predicted to start from an upstream ATG codon by Prokka, leading it to add extra 117 bp. To avoid this problem, we developed a reference-guided gene model reannotation pipeline to correct the misannotations by Prokka.

The reference-guided gene reannotation pipeline (supplementary fig. S4*e*, Supplementary Material online) first applied blastn to identify the TCDC genomic sequences similar to the H37Rv genes. We added the constraint that a gene model must be translatable. Moreover, we only considered the genes with the start codons ATG, GTG, or TTG because they were found in 99% of the *M. tuberculosis* genes.

Therefore, only those blastn hits that satisfied the following rules were considered reannotated gene models: 1) the alignment length between the query and the hit sequence was longer or equal to the query sequence length; 2) the percentage of identical matches was $\geq$90% between the two aligned sequences; 3) e-value was $\leq 10^{-15}$; 4) the alignment length between the query and the hit sequence was in multiples of three; 5) the start codon of the hit sequence was ATG, GTG, or TTG; and 6) the stop codon of the hit sequence was TAA, TAG, or TGA.

If a reannotated gene model had no overlap with any of the Prokka-annotated gene models, it was considered missed by Prokka and was added to the gene models. A Prokka-annotated gene model was replaced by a reannotated gene model if the overlap between the two gene models was longer than 90% of the reannotated gene model. If a reannotated gene model overlapped with more than one Prokka-annotated gene model, then the shorter overlapped Prokka-annotated gene models were discarded if they lay inside the reannotated gene model.

Results

Identification of Drug-Resistant Mutations

We conducted whole-genome sequencing of six Beijing lineage *M. tuberculosis* isolates collected in Taiwan, including one DS (TCDC1), four MDR (TCDC4, TCDC5, TCDC7, and TCDC10), and one XDR (TCDC11) isolates. The sequencing data statistics are presented in supplementary table S2*a*, Supplementary Material online. The mutations for each genome were identified by mapping the Illumina reads to the H37Rv genome (see Materials and Methods). Using a simple method we developed (see Materials and Methods), we identified DR mutations in the coding sequences and upstream regions of 42 known DR genes (Sandgren et al. 2009) in the 6 isolates. In total, we identified 13 known and 6 candidate novel DR mutations (fig. 2) (see Discussion).

Experimental Assays of Candidate Drug-Resistant Mutations

Among the mutations inferred, we selected four to do experimental assays. Compared with the wild-type PZase, the

Table 1
The Kinetics Parameters of the Wild-type PZase, PZase G78V, and V139A and the Wild-type GyrB and GyrB K247N

	PZase			GyrB	
	Wild-type	G78V (TCDC4 and TCDC10)	V139A (TCDC11)	Wild-type	K247N (TCDC10)
V_{max} (mM/min)	1.5±0.3	0.5±0.2	0.4±0.07	8.13±0.84	3.2±0.7
K_M (mM)	1.9±1.2	5.2±3.3	2.9±2.6	68.9±66.7	2.3±2.5
k_{cat} (/min)	148.3±22.1	62.1±22.9	8.9±1.8	162.7±16.8	3.2±0.7
Sp. Activity (μmol/mg/min)	7.4	3.1±1.1	0.4±0.1	2.2±0.2	0.04±0.01
k_{cat}/K_M(/min/mM)	57.0±18.4	13.8±5.3	4.5±3.0	5.4±5.5	2.7±2.5

mutant PZase with G78V, which was found in TCDC4 and TCDC10, showed a higher K_M and a lower k_{cat} (table 1). As a result, the mutant PZase had a higher enzyme efficiency and its specific activity was about 2-fold lower than that of the wild-type. PZase V139A, which was found in TCDC11, had a 10-fold lower activity than the wild-type PZase. These enzymatic studies indicated that PZase G78V and V139A had a lower efficiency in metabolizing pyrazinamide to its active form POA, which may explain the observed DR in the corresponding isolated *M. tuberculosis* strains.

GyrB (gyrase B) activity was initially measured by determining its adenosine triphosphate synthase (ATPase) activity and the K247N mutation in GyrB of TCDC10 was found to cause a significant decrease in k_{cat}, resulting in a ~2-fold decrease in enzyme efficiency (table 1).

The binding of kanamycin to *E. coli* RNA (Woodcock et al. 1991) is well studied, yet no experimental support of kanamycin binding to *M. tuberculosis* RNA has been reported. We thus used synthetic RNA to measure the binding of kanamycin to the target RNA with either A (wild-type) or G (mutant) at position 1401. The surface plasmon resonance analysis revealed the dissociation constant K_D of kanamycin binding to the RNA was 5.0×10^{-5} for 1401A ,whereas 3.8×10^{-4} for 1401G, confirming that the A1401G mutation results in lower binding of kanamycin to RNA (Salian et al. 2012).

Phylogenetic Reconstruction

A phylogenetic tree was reconstructed for the Beijing, CAS-Delhi, Euro-American and Indo-oceanic lineages (fig. 3). Over half of the isolates in the tree belong to the Beijing lineage (blue color), which is large and spreads all over the world. The phylogenetic tree confirms the spoligotypes of the seven Taiwan *M. tuberculosis* isolates, which are marked by blue donuts. TCDC3, the Euro-American lineage isolate (red), is distantly related to the other six Taiwan isolates but clustered with a Thailand isolate (Mtb.Thailand.1700938), several Netherlands isolates, two UK isolates and several Asian isolates. Using the SpolPred software (Coll et al. 2012) and the SITVITWEB database (Demay et al. 2012), we found that TCDC3 and Mtb.Thailand.1700938 belong to two different sublineages, the T2 and T1 lineages, which are common in Europe. The other six isolates form a clade well separated from all other Beijing lineage isolates, although they have different DR profiles. TCDC11, an XDR, is closely related to TCDC4, TDCD5, and TCDC10, which are MDR. TCDC1, which is DS, is likely the common ancestor of TCDC4, TCDC5, TCDC10, and TCDC11.

The DR mutation K43R in the *rpsL* gene is shared by TCDC11, TCDC5, TCDC4, and TCDC10 (fig. 4), so that it apparently arose only once in the common ancestor of these four isolates. However, the S450L mutation in *rpoB* is shared by TCDC4, TCDC10, and TCDC11 but not by TCDC5. Thus, it apparently arose twice, once in the ancestor of TCDC11 and the other time in the common ancestor of TCDC4 and TCDC10. In addition, TCDC4 and TCDC10 share the mutation A479E in *katG*, the mutation I68T in *ndh*, the mutation G78V in *pncA*, and the mutation G88C in *gyrA*, supporting the clustering of TCDC4 and TCDC10 in one clade. All the other mutations are found in only one of the six isolates. TCDC7 shares no DR mutation with the other five isolates, so it is clearly an outgroup to the other five isolates and its DR mutations occurred after its separation from the common ancestor of the other five isolates.

The mutation G58R in the repair gene *mutT2* and the synonymous nucleotide substitution ggG/ggA in codon 12 of gene *ogt* were considered two genotyping markers of Modern Beijing isolates (Mestre et al. 2011). Figure 3 shows the distribution of these two mutations across the phylogeny (orange circles). Both mutations were present in all modern Beijing isolates except TCDC1, which is susceptible to antibiotics. However, the magenta triangles indicate that the mutation G58R in *mutT2* was found in two ancient isolates (Mtb.Thailand.1700793 and Mtb.United States.210) without the ggG/ggA mutation in *ogt* codon 12. Moreover, the mutation G58R in *mutT2* and ggG/ggA in ogt codon 12 were simultaneously identified in an Euro-American isolate (Mtb.Netherlands.N09900612). These findings suggest that G58R in mutT2 is not a modern Beijing lineage-specific mutation (Liu et al. 2016), although it is a common mutation in modern Beijing isolates. In addition, the substitution ggG/ggA in ogt codon 12 is not a unique mutation in modern Beijing strains.

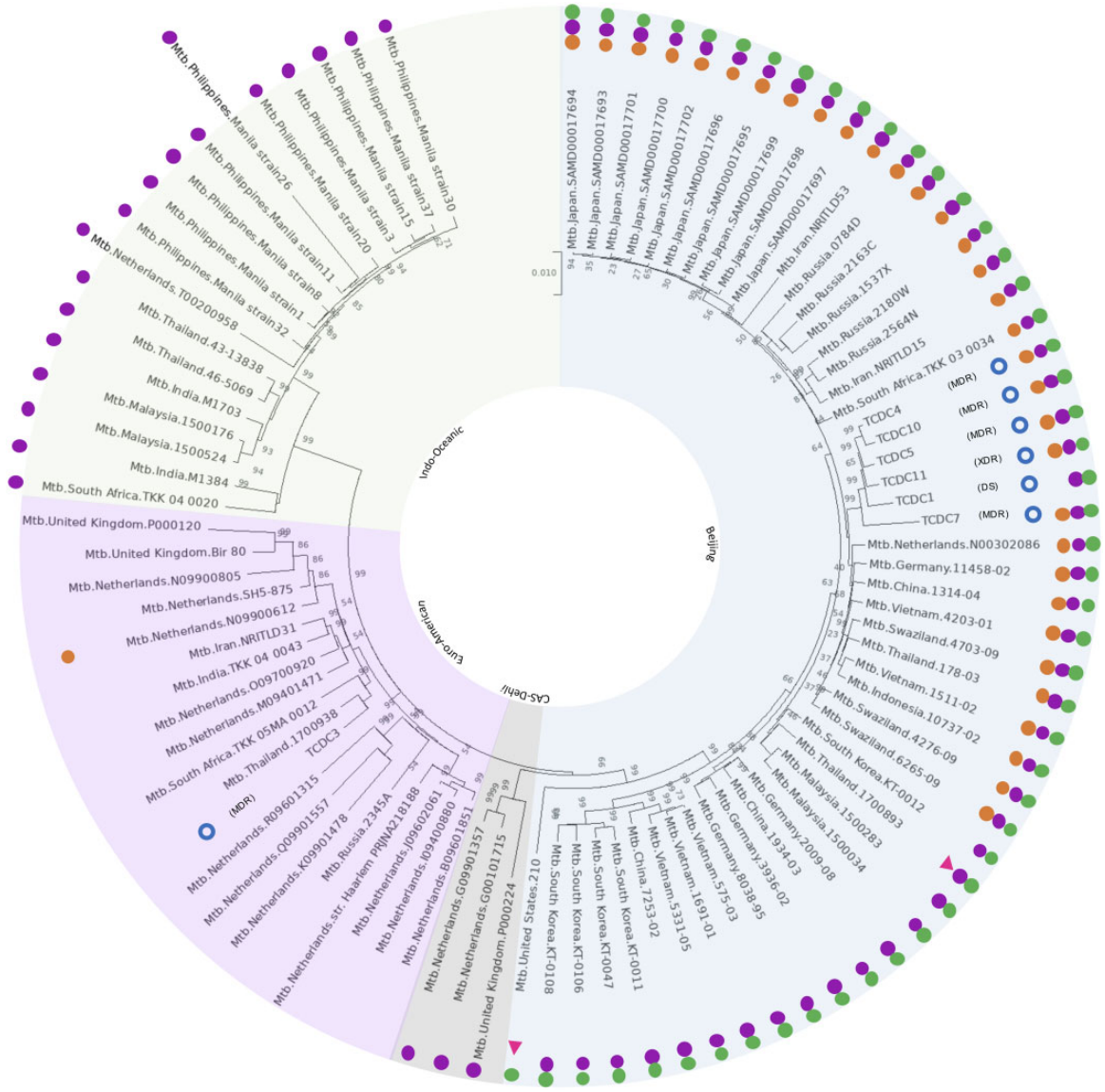

Fig. 3.—Phylogenetic tree of *M. tuberculosis* isolates. Each of the seven Taiwan isolates is indicated by a blue donut. Blue, gray, red, and green background indicate the Beijing, CAS-Dehli-, Euro-America, and Indo-Oceanic lineage *M. tuberculosis* isolates, respectively. The red circles denote Beijing lineage isolates with the potential genetic markers G58R in mutT2 (Rv1160) and ggG/ggA in codon 12 of gene ogt (Rv1316c). The magenta triangles indicate those isolates with mutT2 G58R only.

Genome Assembly and Annotation

The genomes of TCDC1, TCDC3, TCDC7, TCDC10, and TCDC11 were sequenced and assembled (supplementary table S2, Supplementary Material online; see Materials and Methods). Table 2 shows the assembly statistics and the NCBI accession number.

TCDC1, TCDC3, TCDC7, TCDC10, and TCDC11 were annotated to contain 4,116; 4,154; 4,370; 4,138, and 4,104 protein-coding genes, respectively (table 2). The 200 kb tandemly duplicated region in TCDC7 was annotated to contain 224 protein-coding genes and 1 tRNA gene, and the overlapping region of the two 200 kb tandem duplicates was found to contain 2 protein-coding genes.

Genomic characteristics of TCDC 7 and TCDC11—including gene distribution, IS6110 repeat sequences, gc content, gc skew, and tandem duplications of each assembled genome—are shown in figure 5; those of TCDC1, TCDC3, and TCDC10 are shown in supplementary figure S5, Supplementary Material online. The GC skew across the genome is shown in the inner layer. In TCDC3, TCDC7, TCDC10, and TCDC11, the teal and olive areas, which illustrate the positive and the negative GC skew, respectively,

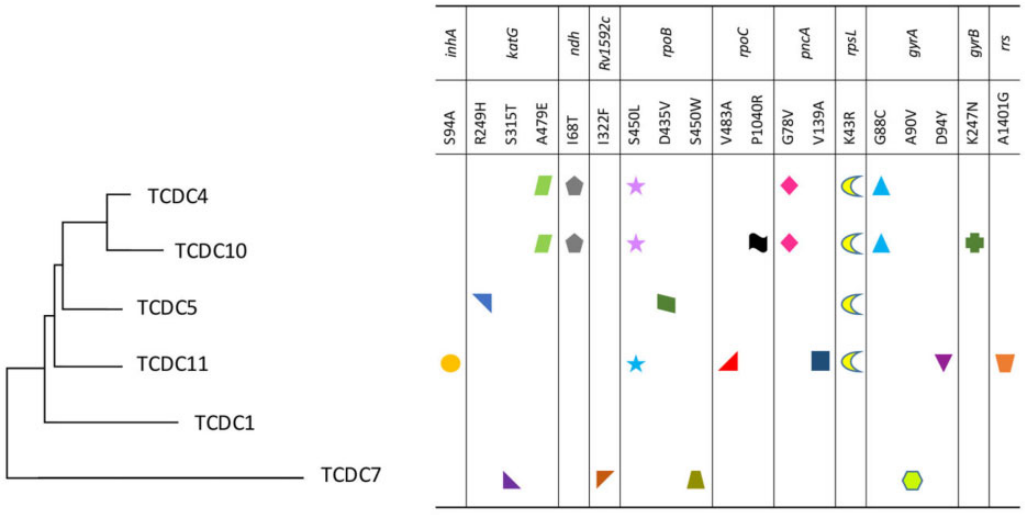

Fig. 4.—Independent occurrences of drug-resistant mutations. The drug-resistant mutations derived from the same origin are shown with the same symbol and color. The magenta and cyan stars show that rifampicin-resistant mutation S450L in *rpoB* occurred twice, whereas the other mutations occurred only once in the studied drug-resistant isolates. The phylogenetic relationships among the six isolates were taken from figure 3.

Table 2

Assembly Statistics and Numbers of Annotated Protein-Coding Genes, tRNAs, and rRNAs of Five *M. tuberculosis* Isolates

Isolate	TCDC1	TCDC3	TCDC7	TCDC10	TCDC11	H37Rv (Reference)
Drug resistance	DS	MDR	MDR	MDR	XDR	DS
Number of scaffolds	66	1	1	1	1	1
Total bases (bp)	4,366,071	4,409,825	4,641,184	4,419,577	4,418,417	4,411,532
N bases	161	12,980	2,401	4,226	0	0
Max scaffold size (bp)	333,253	4,409,825	4,641,184	4,419,577	4,418,417	4,411,532
Min scaffold size (bp)	399	4,409,825	4,641,184	4,419,577	4,418,417	4,411,532
N50	197,471	4,409,825	4,641,184	4,419,577	4,418,417	4,411,532
GC %	65.62	65.38	65.55	65.64	65.69	65.6
Spoligotype clade	Beijing	Euro-America	Beijing	Beijing	Beijing	Euro-America
Protein-coding genes	4,116	4,154	4,370	4,138	4,104	4,018
rRNAs	3	3	3	3	3	3
tRNAs	52	52	53	52	52	45
tmRNAs	1	1	1	1	1	0
NCBI ID	WUCS00000000	CP047258	CP047163	CP047164	CP046728	

indicate that the four genomes each with only one scaffold were indeed assembled into a single circular chromosome. Gray tiles at the fifth layer in figure 5*a* illustrate the 200 kb tandem duplicates in the TCDC7 genome.

Gene Gains and Losses among the TCDC Isolates

Among the 5 newly assembled genomes (table 2), TCDC11 has the smallest number of genes (4,104), although its assembly is likely the most complete. TCDC7 has the largest number of annotated genes (4,370), largely because the genome contains a 200 kb tandemly duplicated region that contains 224 genes. However, even after removing the 224 genes, it still has 42 genes more than TCDC11. Supplementary figure S8, Supplementary Material online shows the genes found in TCDC1, TCDC3, TCDC7, or/and TCDC10 but not in TCDC11 (hypothetical genes were excluded). For example, PE_PGRS14 is found in TCDC1, TCDC7, TCDC10, and partially in TCDC3, but is absent in TCDC11, whereas PE_PGRS56 and PE_PGRS57 are found in TCDC7 only.

The TCDC11 genes that are absent in TCDC1, TCDC3, TCDC7, or/and TCDC10 are shown in supplementary figure S9, Supplementary Material online. The 11 genes that were not found in TCDC3, which is Euro-American, might be Beijing lineage-specific genes. TCDC1 missed the largest number of TCDC11 genes, probably partly because its assembly is incomplete (i.e., only at the scaffold level). The discontinuity of the TCDC1 genome assembly makes the gene annotation of TCDC1 difficult, especially at the boundary of scaffolds. There are six TCDC11 genes that are not found in all other four assembled genomes. For example, TCDC11 contains two universal stress proteins (Mtb_TCDC11_2155 and

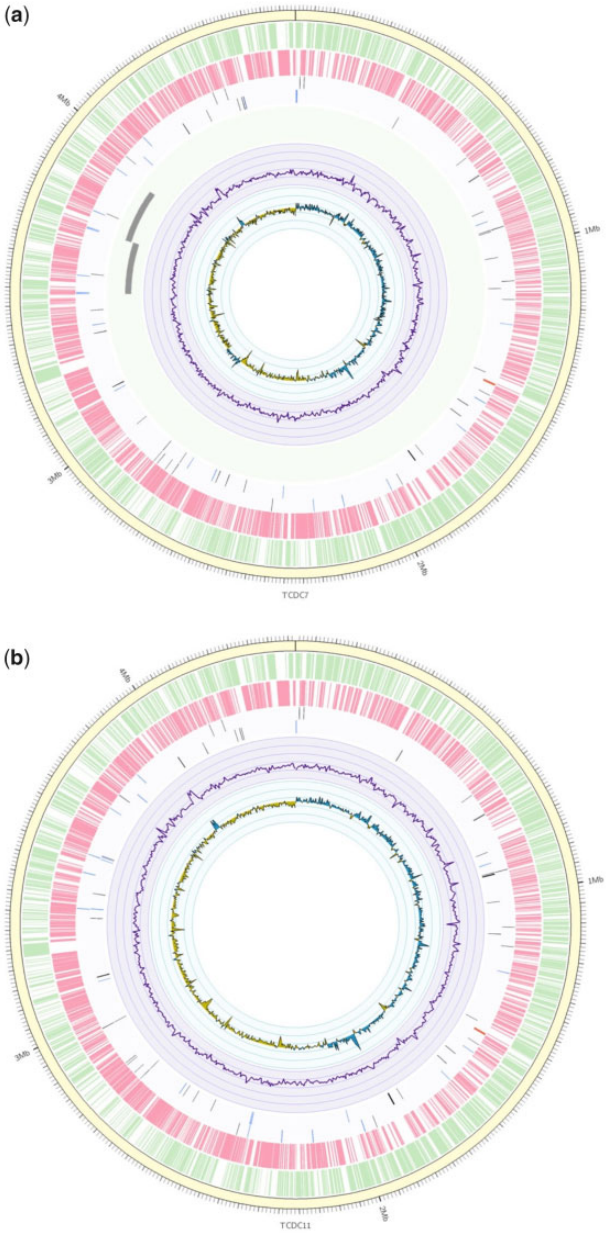

Fig. 5.—Assemblies of TCDC7 and TCDC11 genomes. (*a*) The yellow outer layer shows the circular scaffold (chromosome) of TCDC7. The green and pink bars are protein-coding genes in the forward and reverse strands, respectively. Red, black, and blue bars are rRNA, tRNA, and IS6110 insertion fragments in the forward (layer 4) and reverse (layer 5) strands, respectively; the purple curve in the second inner layer illustrates the GC content across the genome. The GC skew across the genome is shown in the inner layer by teal (positive GC skew) and olive (negative GC skew) areas. The gray tiles in the sixth layer of (*a*) show the 200 kb duplication in TCDC7. (*b*) Genome assembly of TCDC11.

Mtb_TCDC11_2471); TCDC3, TCDC7, and TCDC10 each contains only one; and TCDC1 contains none. The two universal stress proteins in TCDC11 might have helped it survive under the stress of various anti-TB drugs.

TCDC3, which belongs to the Euro-American lineage, has some genes with functions not found in the four Beijing lineage genomes, including PE_PGRS13, 50S ribosomal protein L28, acylamidase, dihydroorotate dehydrogenase, fumarate reductase, LysR family transcriptional regulator, NADPH dehydrogenase, potassium transporter Trk prevent-host-death protein, cyclic pyranopterin monophosphate synthase, and a few transcriptional factors, membrane proteins, transposases, etc. Conversely, TCDC3 lacks some genes that are found in the four Beijing lineage genomes, including EsxI, GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol dimannoside mannosyltransferase, PE_PGRS54, peptidase M22, pterin-4-alpha-carbinolamine dehydratase, sulfite oxidase, and a transporter.

Genes Present in TCDC Genomes but Absent in H37Rv

Supplementary figure S6*a–e*, Supplementary Material online shows the numbers of protein-coding genes in the 5 TCDC genomes assembled in this study that are not found in the H37Rv gene models: 133 for TCDC1, 127 for TCDC3, 136 for TCDC7, 119 for TCDC10, and 115 for TCDC11. We used blastp to search these H37Rv-absent protein-coding genes in the non-redundant (nr) database and identified eight (in TCDC1), one (in TCDC3), one (in TCDC7) and two (in TCDC10) H37Rv-absent protein-coding genes that are not found in the nr database. In contrast, the 115 H37Rv-absent genes identified in TCDC11 are all found in the nr database. The nr hits of H37Rv-absent protein-coding genes have been annotated in the following four species of the *M. tuberculosis* complex: *M. tuberculosis*, *M. orygis*, *M. canettii*, and *M. bovis* (left panels of supplementary fig. S6*f–j*, Supplementary Material online). The right panels of supplementary figure S6*f–j*, Supplementary Material online show the functional classifications of the top nr hits of H37Rv-absent protein-coding genes in each assembly. The functional comparison of H37Rv-absent protein-coding genes among five annotated genomes is shown in supplementary figure S10, Supplementary Material online. Half of the H37Rv-absent protein-coding genes are found in all five TCDC assemblies. In addition, we also looked for H37Rv-absent non-protein-coding genes. Blastn results showed that all H37Rv-absent non-protein-coding genes are in the nucleotide (nt) database.

The H37Rv protein-coding genes that were not found in one or more of the five TCDC isolates are shown in supplementary figure S6*k–o*, Supplementary Material online and their functions are shown in supplementary figure S11, Supplementary Material online. For example, PE_PGRS14 was not found in TCDC11, while the two membrane-associated serine proteases PE19 and PPE26 were not found only in TCDC3. Ten H37Rv genes were not found only in TCDC1, including PE_PGRS53, two-component sensor histidine kinase DosT, the universal stress protein family proteins, etc. The losses of both universal stress proteins (one is completely lost and the other one is partially lost) may be

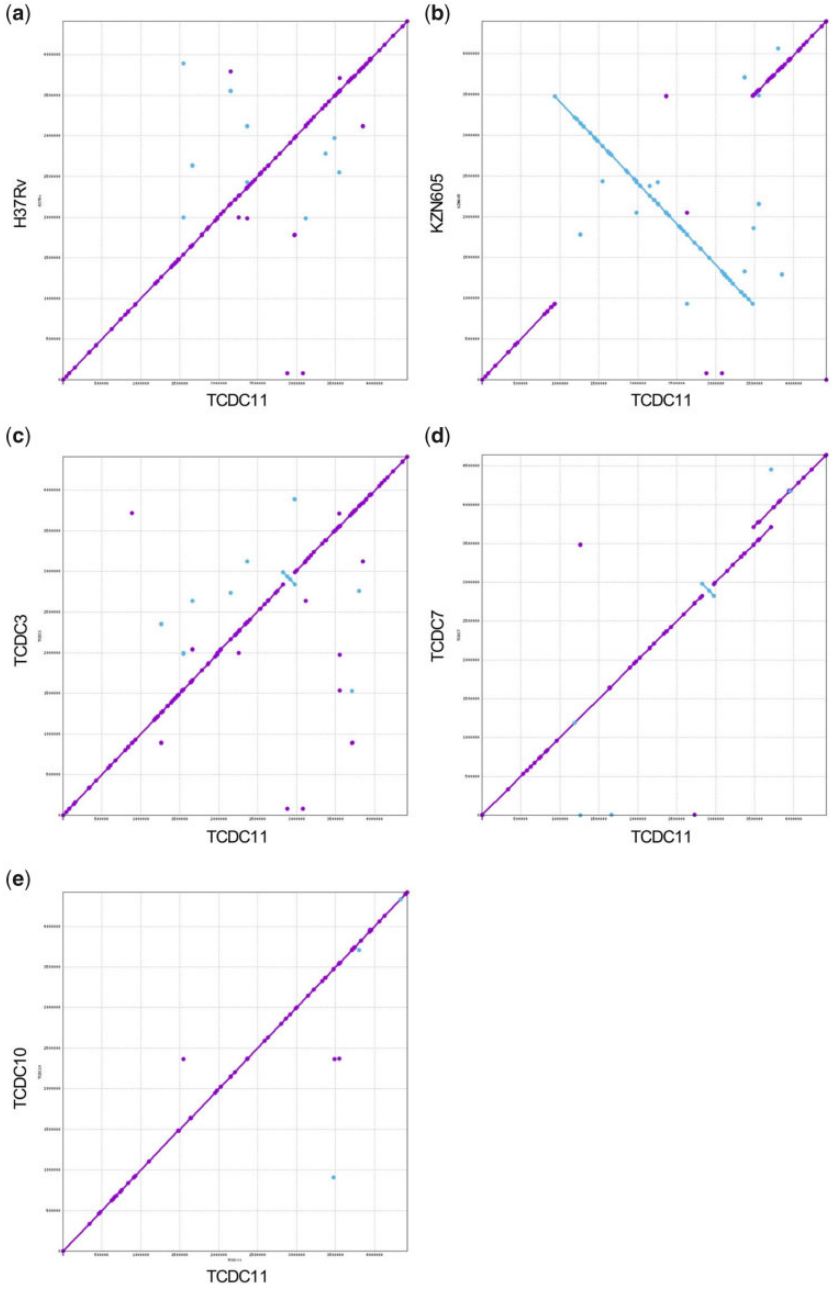

Fig. 6.—Dotplots of TCDC11 against other *M. tuberculosis* genomes. (*a*) Reference H37Rv, (*b*) the published KZN605, (*c*) TCDC3, (*d*) TCDC7, and (*e*) TCDC10. The purple lines show the regions of a genome that could be aligned to the TCDC11 genome by Mummer, whereas the blue lines show the inversion fragments. Repeat sequences IS6110 are not shown.

the cause of TCDC1 susceptible to antibiotics. The incomplete assembly of TCDC1 may have missed the prediction of some protein-coding genes.

Structural Changes

The dotplots of the TCDC11 genome against the H37Rv and KZN605 genomes show that the TCDC11 genome aligns well with the H37Rv genome, but has a large inversion with respect to the KZN genome (fig. 6*a* and *b*). The dotplot of TCDC11 against TCDC3 show that TCDC11 has more rearrangement fragments with respect to TCDC3 than to TCDC7 and TCDC10. In other words, TCDC11 is more distantly related to TCDC3 than to TCDC7 and TCDC10. This observation is consistent with the phylogeny in figure 3 and to the lineage typing in table 2. However, TCDC3 and TDCD7 share the 100 kb inversion (fig. 6*c* and *d*). This inversion was not found in KZN605, F11, and CDC1551 (three Euro-American

lineage isolates) or in CCDC5180 and CCDC5179 (two Beijing lineage isolates) (data not shown). A simple explanation for this observation is that this inversion arose once in the ancestor of TCDC3 and once in the ancestor of TCDC7 after its separation from the ancestor of TCDC1 (see fig. 3). In addition, a tandem duplication is also shown in the dotplot in figure 6*d*. This duplication may have arisen in TCDC7. The dotplots in figure 6*c*–*e* suggest that large-scale genomic rearrangements did not occur frequently during *M. tuberculosis*' evolution in Taiwan, although they were found in a cluster of Russian Beijing B0/W148 isolates (Mokrousov et al. 2012; Bespyatykh et al. 2019).

Constructing a Reference Genome for Beijing Lineage Isolates

To construct a Beijing lineage reference genome, we started with the assembled genome of TCDC11 because it is of very high quality and circularized into one chromosome without any unknown base. TCDC11, however, is resistant to eight anti-TB drugs. To facilitate the identification of drug-resistant mutations, we generated a TCDC11-based DS reference genome, named BLrg, for Beijing lineage isolates (Supplementary file S12, Supplementary Material online). BLrg was derived from TCDC11 by base corrections using Pilon (Walker et al. 2014) and the HiSeq reads of TCDC1, which is DS. The sequences of the unique genes in TCDC1, TCDC7, and TCDC10 were then sequentially added to the end of BLrg by putting NN between every two genes added. Thus, BLrg is DS and contains all genes found in the assembled TCDC11, TCDC1, TCDC 7, and TCDC10 genomes.

BLrg is 4,457,385 bp long, including 175 N's. BLrg contains a total of 4,163 protein-coding genes, including 212 genes not found in H37Rv.

To show that BLrg is a better reference than H37Rv for Beijing isolates, we arbitrarily selected two Beijing isolates from figure 3, that is, Mtb.Japan.SAMD00017702 and Mtb.Vietnam.1691-01. Then we computed the number of sequence differences between Mtb.Japan.SAMD00017702 (and Mtb.Vietnam.1691-01) and BLrg, and between each of them and H37Rv. There are, respectively, 1,545 and 1,421 single nucleotide mutations (SNMs) and 137 and 143 indels for the 2 comparisons using H37Rv as the reference, but only 431 and 574 SNMs and 41 and 71 indels for the 2 comparisons using BLrg as the reference. Clearly, using BLrg as the reference simplifies the analysis of Beijing isolate sequences.

Finally, using BLrg as the reference and the NGS data of the 6 Beijing lineage isolates, we identified the same 19 DR mutations as above, that is, when using H37Rv as the reference.

Discussion

Detection of DR Mutations

The DR is a major challenge to TB control. In this study, NGS was applied to identify DR mutations in six Beijing lineage *M. tuberculosis* isolates with different DR profiles. We focused on analyzing nonsynonymous mutations in the coding regions and mutations in the upstream regions of the known drug-resistant genes. Using 2 new rules, we identified 19 mutations associated with DR, including 13 known and 6 novel mutations. How these mutations were identified and their biological backgrounds and implications are discussed below.

Isoniazid-Resistant Mutations

Isoniazid, a major first-line anti-TB drug, is a prodrug that kills *M. tuberculosis* cells by stopping cell wall synthesis. The prodrug is transformed into the activated form, isonicotinic acyl radical, by KatG (catalase-peroxidase). Isonicotinic acyl radical and oxidized nicotinamide adenine dinucleotide (NAD^+) form the isonicotinic acyl–reduced nicotinamide adenine dinucleotide (NADH) complex, which binds to protein InhA and inhibits the synthesis of mycolic acids in the bacterial cell (Rozwarski et al. 1998). The concentration ratio of NADH to NAD^+ is regulated by *ndh* (Vilchèze et al. 2005). Therefore, mutations in *inhA*, *katG*, and *ndh* may cause isoniazid resistance. The transcription of *Rv1592c*, a gene with unknown function, is induced by isoniazid and mutations in this gene were found in isoniazid-resistant isolates (Ramaswamy et al. 2003; Aragón et al. 2006). Therefore, we also analyzed the mutations in *Rv1592c*. The mutations we found in these four genes are shown in figure 7*a*.

As the mutation R463L in *katG* and the mutation I322V in *Rv1592c* are found in all six isolates (fig. 7*a*), including TCDC1 (a DS isolate), they are considered unrelated to isoniazid resistance. In addition, the mutation R463L in *katG* is a known lineage marker found in both isoniazid-resistant and -susceptible isolates and is thus considered not associated with isoniazid resistance (Torres et al. 2015).

In TCDC11, an XDR-TB isolate, the resistance to isoniazid can probably be explained by the known isoniazid-resistant mutation S94A in *inhA* (fig. 7*a*), although it is a low-level INH-R mutation (Vilchèze et al. 2006). Therefore, the C-1T mutation in the promoter of *inhA* is not a candidate mutation for isoniazid resistance.

In TCDC4 and TCDC10, two MDR isolates, we found two novel mutations *katG* A479E and *ndh* I68T, but no other potential isoniazid-resistant mutations (fig. 7*a*). We therefore consider these two novel mutation candidate mutations for isoniazid resistance. In TCDC7, the mutation *katG* S315T is known to be associated with isoniazid resistance (Yu et al. 2003), so the new mutation *Rv1592c* I322F is probably not an isoniazid-resistant mutation (fig. 7*a*).

In TCDC5, the mutation *katG* R249H is a known mutation for isoniazid-resistance (Brossier et al. 2016) and we found no other candidate mutation for isoniazid resistance (fig. 7*a*).

Fig. 7.—Identified mutations in known DR genes associated with first- and second-line anti-TB drugs. Green and red rectangles denote the known and novel candidate drug-resistant mutations, respectively. The seven drugs studied are presented separately in figure (a)–(f). A mutation that is only identified in DR isolates in which there are no known DR mutations found is considered a DR mutation. The gray rectangles denote mutations deemed not associated with drug resistance because they are found in both DR and DS isolates.

Rifampicin-Resistant Mutations

Rifampicin, a major first-line anti-TB drug, binds to the β-subunit of bacterial DNA-dependent RNA polymerase and inhibits the RNA synthesis of *M. tuberculosis*. Some mutations in the *rpoB* gene, which encodes the β-subunit of bacterial DNA-dependent RNA polymerase, are known to be associated with resistance to rifampicin (Brandis et al. 2012). In addition, fitness-compensatory mutations of rifampicin-resistant *M. tuberculosis* have been found in three bacterial DNA-dependent RNA polymerase genes—*rpoA*, *rpoB*, and *rpoC*—which encode the α, β, and β' subunits of RNA polymerase, respectively (Hughes and Brandis 2013). We therefore

examined the mutations in the *rpoA*, *rpoB*, and *rpoC* genes, and figure 7*b* shows the mutations found in *rpoB* and *rpoC* in different *M. tuberculosis* isolates.

The two amino acids D435 and S450 in *rpoB* form hydrogen bonds with the critical rifampicin hydroxyl groups at O1 and O2 (Campbell et al. 2001). Therefore, mutations at these two positions are considered candidate mutations for rifampicin resistance (fig. 7*b*).

In TCDC4, the known rifampicin-resistant mutation *rpoB* S450L is found.

In TCDC5, the known rifampicin-resistant mutation *rpoB* D435V is found.

In TCDC7, the known rifampicin-resistant mutation *rpoB* S450W is found (Casali et al. 2016).

In TCDC10, as in TCDC4, the known rifampicin-resistant mutation *rpoB* S450L is found. On the other hand, *rpoC* P1040R is a candidate novel mutation for compensatory rifampicin resistance mutation.

In TCDC11, *rpoB* S450L is a known rifampicin-resistant mutation and *rpoC* V483A is considered a compensatory mutation associated with rifampicin resistance.

Pyrazinamide-Resistant Mutations

Pyrazinamide is activated by PZase, the product of *pncA*, and transformed into POA, which disrupts the assembly of the *M. tuberculosis* cell membrane by inhibiting fatty acid synthase I (Palomino and Martin 2014). Hence, mutations in *pncA* may cause pyrazinamide resistance. Figure 7*c* shows the mutations we found in *pncA*.

In TCDC4 and TCDC10, the mutation *pncA* G78V is considered a candidate novel mutation for pyrazinamide resistance.

In TCDC11, *pncA* V139A, a mechanism-unknown mutation found in pyrazinamide-resistant isolates is identified. A previous study showed that the POA efflux rate was significantly lower in pyrazinamide-resistant isolates than in pyrazinamide-susceptible isolates (Zimic et al. 2012). The POA efflux rate of the pyrazinamide-resistant isolate with *pncA* V139A was higher than the average rate of pyrazinamide-susceptible isolates (Zimic et al. 2012). However, we still consider *pncA* V139A a candidate DR mutation of pyrazinamide because it is the only mutation in *pncA* found in TCDC11 and was found in other PZA-resistant isolates (Sheen et al. 2017). As will be seen later, this conclusion was supported by our functional assay.

Streptomycin-Resistant Mutations

Streptomycin binds to helices 1, 18, 27, and 44 of 16S ribosomal RNA and S12 ribosomal protein (Demirci et al. 2013). Therefore, mutations in 16S ribosomal RNA and *rpsL*, which encodes the S12 ribosomal protein, may cause streptomycin resistance. In addition, mutations in *gidB* were also found in streptomycin-resistant isolates (Wong et al. 2011). We therefore investigated the mutations in *rrs*, *rpsL*, and *gidB*. We found the known mutation *rpsL* K43R (Spies et al. 2011) in all four streptomycin-resistant isolates, TCDC4, TCDC5, TCDC10, and TCDC11 (fig. 7*d*). The mutation *gidB* E92D was detected in all *M. tuberculosis* isolates including the streptomycin-susceptible one and is thus not a candidate mutation for streptomycin resistance (Spies et al. 2011).

In TCDC11, the nucleotide substitution *rrs* A1401G was found (fig. 7*f*). However, it was known to be associated with high-level resistance to KM and AMK and only low-level to CPM (Hobbie et al. 2006); it is therefore not the cause of streptomycin resistance in TCDC11.

Fluoroquinolone-Resistant Mutations

The second-line anti-TB drugs include fluoroquinolone and aminoglycoside/polypeptide drugs. Ofloxacin is one of the fluoroquinolone drugs (Ramaswamy and Musser 1998) that binds *M. tuberculosis* DNA gyrase and inhibits DNA gyrase to relax positive supercoils of DNA. Mutations in the quinolone resistance determining region (QRDR) in *gyrA* and *gyrB* (QRDR-A and QRDR-B) are known to be associated with fluoroquinolone resistance (Piton et al. 2010). Therefore, we analyzed the mutations in the *gyrA* and *gyrB* genes (fig. 7*e*), and six nonsynonymous mutations in the *gyrA* gene and one nonsynonymous mutation in the *gyrB* gene were found in the six *M. tuberculosis* isolates. The mutations *gyrA* E21Q, S95T, and G668D were found in both ofloxacin-susceptible and ofloxacin-resistant *M. tuberculosis*, so they are not candidates for ofloxacin resistance (Farhat et al. 2016). The other three mutations (fig. 7*e*) are discussed below.

In TCDC4, the G88C mutation in *gyrA* is located at one of the fluoroquinolone binding sites, QRDR-A (Matrat et al. 2006). It is a known fluoroquinolone-resistance mutation.

In TCDC7, the A90V mutation in *gyrA* is detected, and like G88C, A90V is at a fluoroquinolone binding site of gyrase A. It is a known mutation for fluoroquinolone resistance (Matrat et al. 2006).

In TCDC11, the D94Y mutation in *gyrA* is a known fluoroquinolone-resistant mutation (Matrat et al. 2006).

In TCDC10, as in TCDC4, the mutation G88C in *gyrA* was found. In addition, a novel mutation in *gyrB*, K247N is found. It is not located in QRDR and has not been previously detected in fluoroquinolone-resistant isolates. Our functional assay verified the association between this mutation and ofloxacin resistance (see next section).

Second-Line Injectable Drug-Resistant Mutations

The second-line injectable drugs for tuberculosis treatment include the polypeptide antibiotic capreomycin and the aminoglycoside antibiotics amikacin and kanamycin, which bind to the A site of the 30S subunit of *M. tuberculosis* ribosome,

causing incorrect translation (Poehlsgaard and Douthwaite 2005). In the six *M. tuberculosis* isolates under study, only TCDC11 resists the second-line injectable drugs, and we found a known mutation associated with second-line injectable drug (fig. 7*f*). The adenine to guanine substitution at position 1401 in *rrs* represses the binding of aminoglycosides to the A site of *M. tuberculosis* ribosome RNA (Hobbie et al. 2006). The association between *rrs* A1401G and DR was considered almost 100% specific to both kanamycin and amikacin, whereas the specificity to capreomycin is lower (Jugheli et al. 2009).

Experimental Support of Candidate Drug-Resistant Mutations

The functional assays of two known and two novel drug-resistant mutations showed that in each of these four mutations, the affinity of antibiotics to their targets was reduced, raising the capacity of *M. tuberculosis* to resist antibiotics. One of the experimentally supported novel DR mutations associated with ofloxacin is K247N in *gyrB*, which is outside the QRDR-B region that is defined between residue positions 500 and 540 (Pantel et al. 2012). The mechanism of K247N causing a decrease in the efficacy of ofloxacin is unclear as it is not located in the enzyme catalytic core (Piton et al. 2010). Further structural analysis may be conducted to figure out the DR mechanism of K247N.

Kanamycin and amikacin are aminoglycosides and are important second-line injectable drugs. They bind to the 16S-ribosome RNA and kill bacteria by producing incorrect translation. Structural evidence has shown that the 6′ amino group in ring I of aminoglycoside cannot interact with guanine 1401 of ribosome RNA via a hydrogen bond. In addition, repulsion between the 6′ amino group of ring I and the N1/N2 amino groups of guanine was also observed. Therefore, the substitution of adenine to guanine at 1401 of the ribosomal RNA prevents aminoglycoside from binding to the RNA and decreases the capacity of aminoglycoside to kill *M. tuberculosis* (Hobbie et al. 2006). We conducted a functional assay to confirm that nucleotide substitution A1401G in *rrs* leads to a weaker kanamycin binding to *rrs* and causes DR.

Genome Sequencing, Assembly, and Annotation

This study applied three NGS technologies to conduct genome sequencing and assembly of five Taiwan *M. tuberculosis* isolates, including one Euro-American lineage and four Beijing lineage isolates. The obtained circular genomes of one XDR and three MDR isolates are the first Taiwan *M. tuberculosis* assemblies in circular chromosome form and can be used as the reference for genomic study of Taiwan isolates.

Gene annotation was conducted by Prokka (Seemann 2014) with the addition of third-party tools. However, some genes were incorrectly annotated and some were missed. In order to correct the prediction errors, we developed a reference-guided gene model reannotation pipeline to adjust the gene models. With the developed reannotation pipeline, the gene sequences of 850 TCDC11 genes were revised, 232 short TCDC11 annotated sequences were removed, and 51 missed annotated genes were added to the gene models. We compared the sequence alignment percentage before and after reannotation to evaluate the performance of the reference-guided gene model reannotation pipeline. The sequences of removed genes and revised genes before correction were blasted against the H37Rv genes to identify the alignment percentage of Prokka annotation. The alignment percentage of each gene is defined by the product of alignment identity and the percentage of the sequence aligned, which is the ratio of alignment length to aligned gene length of H37Rv. Adjusted sequences of revised genes and new genes were used to compute the alignment percentage after gene reannotation. Histograms of sequence alignment percentage before and after gene model revision are illustrated by blue and dark magenta bars in supplementary figure S7, Supplementary Material online. The sequence alignment percentage of adjusted gene models is significantly higher than the Prokka one (P-value $= 5\times10^{-63}$), indicating that our reference-guided gene model reannotation pipeline effectively corrected annotation errors.

Toxin–Antitoxin System Gene Mutations

The toxin–antitoxin (TA) system is essential for bacteria to adapt to external stress. Toxin MazF3, MazF6, and MazF9 of the ribonuclease MazEF TA system are considered to respond to antibiotics and may induce drug tolerance (Tiwari et al. 2015). Our genome assembly and annotation led to the identification of the MazEF TA system in all *M. tuberculosis* isolates we studied. However, point mutations were found in *mazF3*, *mazF6*, and *mazF8* in the isolates (table 3). The mutations T65I in MazF3 and G41V in MazF8 are found in all of the six Taiwan Beijing lineage isolates studied, including DS and resistant ones. These two mutations are thus not associated with DR.

The mutation G41V in MazF8 was found in all of the Beijing lineage isolates in figure 3 (green circles), consistent with the view that G41V is a Beijing lineage marker mutation (Mikheecheva et al. 2017). On the other hand, the mutation T61I in MazF3 was found in all isolates in figure 3 (purple circles) except all Euro-American ones and Mtb.United.States.210, which is the farthest Beijing isolate in figure 3. In other words, the Euro-American lineage is the only lineage that lacks both mutations, T61I in mazF3 and G41V in MazF8, so the lack of both mutations can be considered a genetic marker for Euro-American lineage isolates.

In addition to the point mutation T65I in MazF3, the mutation tGg/tAg that caused a premature stop codon at the tenth amino acid in MazF6 was also identified in TCDC11,

Table 3
Mutations of MazEF Toxin–Antitoxin Systems in TCDC Isolates

Toxin MazF	a.a. (codon) change	Isolates
MazF3	T65I (aCc/aTc)	TCDC1, TCDC4, TCDC5, TCDC7, TCDC10, TCDC11
MazF6	W10* (tGg/tAg)	TCDC11
MazF6	G59S (Ggc/Agc)	TCDC3
MazF8	G41V (gGt/gTt)	TCDC1, TCDC4, TCDC5, TCDC7, TCDC10, TCDC11

which resists eight antibiotics. Overexpression of *mazF6* results in the inhibition of *M. tuberculosis* cell growth; simultaneous deletion of *mazF3*, *mazF6*, and *mazF9* reduces TB persistence (Tiwari et al. 2015). We presume that the truncated MazF6 product and the mutated MazF3 product are cofunctional and increase the persistence of TCDC11.

In TCDC3, the Euro-American lineage MDR isolate, the mutation G59S in MazF6 was found. It has been known that a deletion of *mazF3*, *mazF6*, or *mazF9* does not suppress cell growth (Tiwari et al. 2015). Hence, the mutation G59S in MazF6 is not considered to be associated with DR in TCDC3.

In addition, another Beijing lineage marker mutation T16A (Aca/Gca) in toxin gene *vapC37* was found in all Taiwan DR Beijing lineage isolates, but not the DS isolate (Sala et al. 2014; Mikheecheva et al. 2017; Zaychikova et al. 2018). As *vapC37* has been suggested to be responsible for a toxin in latent infection, T16A in *vapC37* might be a virulence mutation.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

All the library construction and next-generation sequencing experiments were carried out by the High Throughput Genomics Core Facility of the Biodiversity Research Center in Academia Sinica, Taiwan. This study was supported by Academia Sinica (AS-Summit-108) and the Ministry of Science and Technology, Taiwan (MOST 106-2321-B-001-013 and MOST 107-2311-B-001-016-MY3).

Literature Cited

Aragón LM, et al. 2006. Rapid detection of specific gene mutations associated with isoniazid or rifampicin resistance in *Mycobacterium tuberculosis* clinical isolates using non-fluorescent low-density DNA microarrays. J Antimicrob Chemother. 57(5):825–831.

Bespyatykh J, et al. 2019. Proteogenomic analysis of *Mycobacterium tuberculosis* Beijing B0/W148 cluster strains. J Proteomics 192:18–26.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27(4):578–579.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Brandis G, Wrande M, Liljas L, Hughes D. 2012. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. Mol Microbiol. 85(1):142–151.

Brossier F, Boudinet M, Jarlier V, Petrella S, Sougakoff W. 2016. Comparative study of enzymatic activities of new KatG mutants from low- and high-level isoniazid-resistant clinical isolates of *Mycobacterium tuberculosis*. Tuberculosis (Edinb) 100:15–24.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Campbell EA, et al. 2001. Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. Cell 104(6):901–912.

Casali N, et al. 2016. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. PLoS Med. 13(10):e1002137.

Chin C-S, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10(6):563–569.

Chuang P-H, Wu M-H, Fan S-Y, Lin K-Y, Jou R. 2016. Population-based drug resistance surveillance of multidrug-resistant tuberculosis in Taiwan, 2007-2014. PLoS One 11(11):e0165222.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80–92.

Cole ST, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393(6685):537–544.

Coll F, et al. 2015. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 7(1):51.

Coll F, et al. 2012. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. Bioinformatics 28(22):2991–2993.

Demay C, et al. 2012. SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. Infect Genet Evol. 12(4):755–766.

Demirci H, et al. 2013. A structural basis for streptomycin-induced misreading of the genetic code. Nat Commun. 4(1):1355.

Eddy S. 1998. Profile hidden Markov models. Bioinformatics 14(9):755–763.

Farhat MR, et al. 2016. Gyrase mutations are associated with variable levels of fluoroquinolone resistance in *Mycobacterium tuberculosis*. J Clin Microbiol. 54(3):727–733.

Hobbie SN, et al. 2006. A genetic model to investigate drug-target interactions at the ribosomal decoding site. Biochimie 88(8):1033–1043.

Hughes D, Brandis G. 2013. Rifampicin resistance: fitness costs and the significance of compensatory evolution. Antibiotics 2(2):206–216.

Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11(1):119.

Jou R, Chiang C-Y, Huang W-L. 2005. Distribution of the Beijing family genotypes of *Mycobacterium tuberculosis* in Taiwan. J Clin Microbiol. 43(1):95–100.

Jugheli L, et al. 2009. High level of cross-resistance between kanamycin, amikacin, and capreomycin among *Mycobacterium tuberculosis* isolates from Georgia and a close relation with mutations in the rrs gene. Antimicrob Agents Chemother. 53(12):5064–5068.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35(6):1547–1549.
Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5(2):R12.
Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35(9):3100–3108.
Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32(1):11–16.
Li H, et al.; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.
Liu Q, et al. 2016. Genetic features of *Mycobacterium tuberculosis* modern Beijing sublineage. Emerg Microbes Infect. 5(1):e14–e18.
Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1(1):18.
Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380.
Matrat S, et al. 2006. Functional analysis of DNA gyrase mutant enzymes carrying mutations at position 88 in the A subunit found in clinical strains of *Mycobacterium tuberculosis* resistant to fluoroquinolones. Antimicrob Agents Chemother. 50(12):4170–4173.
Mestre O, et al. 2011. Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS One 6(1):e16020.
Mikheecheva NE, Zaychikova MV, Melerzanov AV, Danilenko VN. 2017. A Nonsynonymous SNP catalog of *Mycobacterium tuberculosis* virulence genes and its use for detecting new potentially virulent sublineages. Genome Biol Evol. 9(4):887–899.
Mokrousov I, et al. 2012. Russian 'successful' clone B0/W148 of *Mycobacterium tuberculosis* Beijing genotype: a multiplex PCR assay for rapid detection and global screening. J Clin Microbiol. 50(11):3757–3759.
Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics 13(Suppl 14):S8.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.
Palomino JC, Martin A. 2014. Drug resistance mechanisms in *Mycobacterium tuberculosis*. Antibiotics 3(3):317–340.
Pantel A, et al. 2012. Extending the definition of the GyrB quinolone resistance-determining region in *Mycobacterium tuberculosis* DNA gyrase for assessing fluoroquinolone resistance in *M. tuberculosis*. Antimicrob Agents Chemother. 56(4):1990–1996.
Piton J, et al. 2010. Structural insights into the quinolone resistance mechanism of *Mycobacterium tuberculosis* DNA gyrase. PLoS One 5(8):e12245.
Poehlsgaard J, Douthwaite S. 2005. The bacterial ribosome as a target for antibiotics. Nat Rev Microbiol. 3(11):870–881.
Ramaswamy S, Musser JM. 1998. Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. Tuber Lung Dis. 79(1):3–29.
Ramaswamy SV, et al. 2003. Single nucleotide polymorphisms in genes associated with isoniazid resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 47(4):1241–1250.
Reed MB, Gagneux S, Deriemer K, Small PM, Barry CE. 2007. The W-Beijing lineage of *Mycobacterium tuberculosis* overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated. J Bacteriol. 189(7):2583–2589.
Rodríguez-Castillo JG, et al. 2017. Comparative genomic analysis of *Mycobacterium tuberculosis* Beijing-like strains revealed specific genetic variations associated with virulence and drug resistance. Infect Genet Evol. 54:314–323.
Rozwarski DA, Grant GA, Barton DH, Jacobs WR, Sacchettini JC. 1998. Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. Science 279(5347):98–102.
Sala A, Bordes P, Genevaux P. 2014. Multiple toxin-antitoxin systems in *Mycobacterium tuberculosis*. Toxins (Basel) 6(3):1002–1020.
Salian S, et al. 2012. Structure-activity relationships among the kanamycin aminoglycosides: role of ring I hydroxyl and amino groups. Antimicrob Agents Chemother. 56(12):6104–6108.
Sandgren A, et al. 2009. Tuberculosis drug resistance mutation database. PLoS Med. 6(2):e2.
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.
Sheen P, et al. 2017. A multiple genome analysis of *Mycobacterium tuberculosis* reveals specific novel genes and mutations associated with pyrazinamide resistance. BMC Genomics 18(1):769.
Spies FS, et al. 2011. Streptomycin resistance and lineage-specific polymorphisms in *Mycobacterium tuberculosis* gidB gene. J Clin Microbiol. 49(7):2625–2630.
Tiwari P, et al. 2015. MazF ribonucleases promote *Mycobacterium tuberculosis* drug tolerance and virulence in guinea pigs. Nat Commun. 6(1):6059.
Torres JN, et al. 2015. Novel katG mutations causing isoniazid resistance in clinical *M. tuberculosis* isolates. Emerg Microbes Infect. 4(1):e42–e49.
Vilchèze C, et al. 2005. Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria. Antimicrob Agents Chemother. 49(2):708–720.
Vilchèze C, et al. 2006. Transfer of a point mutation in *Mycobacterium tuberculosis* inhA resolves the target of isoniazid. Nat Med. 12(9):1027–1029.
Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9(11):e112963.
Wattam AR, et al. 2017. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 45(D1):D535–D542.
Wong SY, et al. 2011. Mutations in gidB confer low-level streptomycin resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 55(6):2515–2522.
Woodcock J, Moazed D, Cannon M, Davies J, Noller HF. 1991. Interaction of antibiotics with A- and P-site-specific bases in 16S ribosomal RNA. EMBO J. 10(10):3099–3103.
World Health Organization. 2018. Global tuberculosis report 2018. World Health Organization.
Yu S, Girotto S, Lee C, Magliozzo RS. 2003. Reduced affinity for isoniazid in the S315T mutant of *Mycobacterium tuberculosis* KatG is a key factor in antibiotic resistance. J Biol Chem. 278(17):14769–14775.
Zaychikova MV, et al. 2018. Single nucleotide polymorphisms of Beijing lineage *Mycobacterium tuberculosis* toxin-antitoxin system genes: their role in the changes of protein activity and evolution. Tuberculosis (Edinb) 112:11–19.
Zhang H, et al. 2013. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 45(10):1255–1260.
Zimic M, et al. 2012. Pyrazinoic acid efflux rate in *Mycobacterium tuberculosis* is a better proxy of pyrazinamide resistance. Tuberculosis (Edinb) 92(1):84–91.

Associate editor: Rachel Whitaker
Highlights editor: George Zhang