RESEARCH ARTICLE

TGS-TB: Total Genotyping Solution for *Mycobacterium tuberculosis* Using Short-Read Whole-Genome Sequencing

Tsuyoshi Sekizuka[1], Akifumi Yamashita[1], Yoshiro Murase[2], Tomotada Iwamoto[3], Satoshi Mitarai[4], Seiya Kato[5], Makoto Kuroda[1]*

1 Pathogen Genomics Center, National Institute of Infectious Diseases, Shinjyuku-ku, Tokyo, Japan, **2** Molecular Epidemiology Division, The Research Institute of Tuberculosis/Japan Anti-Tuberculosis Association, Kiyose, Tokyo, Japan, **3** Department of Infectious Diseases, Kobe Institute of Health, Chuo-ku, Kobe, Japan, **4** Bacteriology Division, Department of Mycobacterium Reference and Research, Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Kiyose, Tokyo, Japan, **5** Research Institute of Tuberculosis, Japan Anti-Tuberculosis Association, Kiyose, Tokyo, Japan

* makokuro@niid.go.jp

OPEN ACCESS

Citation: Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. (2015) TGS-TB: Total Genotyping Solution for *Mycobacterium tuberculosis* Using Short-Read Whole-Genome Sequencing. PLoS ONE 10(11): e0142951. doi:10.1371/journal.pone.0142951

Editor: Niyaz Ahmed, University of Hyderabad, INDIA

Received: May 21, 2015

Accepted: October 28, 2015

Published: November 13, 2015

Data Availability Statement: All data are fully available without restriction and TGS-TB web site: http://gph.niid.go.jp/tgs-tb/; the short-read sequences for seven Mtb isolates (JP01 to JP07) have been deposited in the DNA Data Bank of Japan (DDBJ; accession number: DRA003808).

Funding: This research was funded through a Grant-in-Aid for Research on Emerging and Re-emerging Infectious Diseases (H25-Shinko-Ippan-015) from the Ministry of Health Labour and Welfare Programs of Japan, and also through Research Program on Emerging and Re-emerging Infectious Diseases

Abstract

Whole-genome sequencing (WGS) with next-generation DNA sequencing (NGS) is an increasingly accessible and affordable method for genotyping hundreds of *Mycobacterium tuberculosis* (Mtb) isolates, leading to more effective epidemiological studies involving single nucleotide variations (SNVs) in core genomic sequences based on molecular evolution. We developed an all-in-one web-based tool for genotyping Mtb, referred to as the Total Genotyping Solution for TB (TGS-TB), to facilitate multiple genotyping platforms using NGS for spoligotyping and the detection of phylogenies with core genomic SNVs, IS*6110* insertion sites, and 43 customized loci for variable number tandem repeat (VNTR) through a user-friendly, simple click interface. This methodology is implemented with a KvarQ script to predict MTBC lineages/sublineages and potential antimicrobial resistance. Seven Mtb isolates (JP01 to JP07) in this study showing the same VNTR profile were accurately discriminated through median-joining network analysis using SNVs unique to those isolates. An additional IS*6110* insertion was detected in one of those isolates as supportive genetic information in addition to core genomic SNVs. The results of in silico analyses using TGS-TB are consistent with those obtained using conventional molecular genotyping methods, suggesting that NGS short reads could provide multiple genotypes to discriminate multiple strains of Mtb, although longer NGS reads ($\geq$300-mer) will be required for full genotyping on the TGS-TB web site. Most available short reads (~100-mer) can be utilized to discriminate the isolates based on the core genome phylogeny. TGS-TB provides a more accurate and discriminative strain typing for clinical and epidemiological investigations; NGS strain typing offers a total genotyping solution for Mtb outbreak and surveillance. TGS-TB web site: https://gph.niid.go.jp/tgs-tb/.

(15fk0108011h0003 and 15fk0108004h0001) from Japan Agency for Medical Research and development, AMED. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

An estimated 8.6 million people developed tuberculosis (TB) in 2012, and 1.3 million individuals died from this disease. WHO reported 450,000 new cases of multidrug resistant (MDR) *Mycobacterium tuberculosis* (Mtb) resistant to at least isoniazid and rifampicin worldwide [1]. Molecular genotyping of Mtb has been well developed [2]; three main typing methods, specifically IS*6110* restriction fragment length polymorphism (RFLP), spoligotyping (spacer oligonucleotide typing), and variable-number tandem repeat (VNTR) analysis, are currently used for fingerprinting Mtb strains to detect recent transmission.

IS*6110*-RFLP typing is a standard genotyping approach [3], but requires subculturing the isolates for several weeks to obtain sufficient DNA, and the typing procedure distinguishes a limited number of polymorphisms. Mtb contains 10 to 50 copies of a 36-bp direct repeat (DR) in clustered regularly interspaced palindromic repeats (CRISPRs), and the spacer sequences between DRs have different nucleotide sequences among strains. Thus, the pattern of spacers in a strain can be used for spoligotyping [4]. However, spoligotyping has less resolving power to discriminate among Mtb strains compared with IS*6110* genotyping [5]. The Mtb genome contains many mycobacterial interspersed repeat units (MIRUs) and MIRU-VNTR. MIRU-VNTR typing has progressed and is currently used to visualize the transmission of multiple Mtb strains, yielding intrinsically digital results that can be easily catalogued in a computer database [6]. Among more than 40 VNTR loci on the Mtb chromosome, MIRU-VNTR 15 and 24 loci have been proposed as the international standard [7]. However, the discriminatory power of this technique is not sufficient in countries such as East Asia and Russia with a high proportion of Beijing-type Mtb. Lineage- or sublineage-specific loci should be additionally investigated to increase the discriminative power of genotyping [8].

Genetic elements for molecular epidemiological genotyping techniques do provide adequate discriminatory power for distinguishing *M*. *tuberculosis* strains. However, the clustered strains defined using these methods might be distantly related, both genetically and historically, reflecting the low reliability of these tests to distinguish recent from past transmissions [9]. Thus, epidemiological investigations are typically needed to confirm recent transmission and remote infection.

Whole-genome sequencing (WGS) using next-generation DNA sequencing (NGS) has emerged as an increasingly accessible and affordable method for genotyping hundreds of Mtb isolates, leading to more effective epidemiological studies involving single nucleotide polymorphisms (SNPs) in the core genomic sequence based on the molecular evolutionary clock [9–14]. Genome-based clustering patterns are more consistent with contact tracing data and the geographical distribution of the cases compared with clustering patterns based on classical genotyping [15]. WGS facilitates the effective tracing of the Mtb complex (MTBC). Niemann et al. demonstrated that WGS revealed genomic heterogeneity among drug-susceptible and drug-resistant Mtb isolates with identical IS*6110* fingerprints and 23 out of 24 MIRU-VNTR loci [16]. Such heterogeneity is not detected using conventional MTBC genotyping, and some aspects of Mtb transmission dynamics could be missed or misinterpreted. When the overall genetic diversity of circulating clones is restricted, standard genotyping might not distinguish between relapse and exogenous re-infection. Bryant et al. demonstrated that WGS facilitates the differentiation of relapse and re-infection cases, with higher resolution through small (0 to 6 SNPs) and large (1,306 to 1,419 SNPs) distances [13]. It has been suggested that the mutation rate is constant at approximately 0.5 single nucleotide polymorphisms per genome per year in latent, active and re-activated diseases [9, 13, 17, 18]. Walker et al. established that most Mtb isolates were within five SNPs on the genome of another isolate obtained from the same individual or from a household contact [9].

Freely accessible web services facilitate the genotyping of isolated strains alone or in comparison with reference strains from major MTBC lineages. Currently, MIRU-VNTRplus web tools (http://www.miru-vntrplus.org) are available for analyzing MLVA data (MtbC15-9 type) in combination with other complementary typing data, including spoligotypes, regions of difference (RDs), SNPs in antimicrobial target genes and susceptibility information [19]. In addition to conventional genotyping tools, Kohl et al. recently provided core genomic multilocus sequence typing (cgMLST) tools to expand the approach for standardized WGS-based genotyping [20]. *In silico* spoligotyping is available using SpolPred to more accurately and rapidly determine a spoligotype [21]. The stand-alone, user friendly tool KvarQ has been reported to assign MTBC lineage/sublineages, SNP-barcodes and potential antimicrobial resistance (AMR) genes within 2 minutes based on SNP analysis directly from NGS short reads [22]. Similarly, PhyResSE also facilitates the lineage genotyping and AMR detection of Mtb using web-based tools [23]. The comprehensive genomic variation map for Mtb is available on the PolyTB web-based tool (http://pathogenseq.lshtm.ac.uk/polytb) to visualize the resulting MTBC genetic variations (74,039 SNPs, 4820 indels and 800 deletion sites) and important meta-data (e.g., *in silico* inferred strain-types and locations) at a genome and global scale [24]. Unexpectedly, newly obtained raw sequences cannot be analyzed using PolyTB, and conventional typing data are not available on the PolyTB resource. PhyTB illuminates *M. tuberculosis* genomic variation within epidemiological, geographical and phylogenic settings from 1,601 *M. tuberculosis* isolates, facilitating the assessment of genotype-phenotype associations [25].

WGS-based genotyping (cgMLST) [20], SpolPred [21], KvarQ [22] and PhyResSE [23] provide partial *in silico* genotyping analysis but do not include other conventional genotyping formats, such as IS*6110*-RFLP and MIRU-VNTR. This independent genotyping tool has not been integrated into one system. Thus, it can be laborious to determine each genotype through multiple genotyping programs.

Here, we present a Total Genotyping Solution for TB (TGS-TB) web-based tool to facilitate multiple genotyping formats using NGS for the analysis of phylogenies with core genomic single nucleotide variations (SNVs), linkage network analysis of outbreak strains, spoligotyping, the analysis of IS*6110* insertion sites and VNTRs (our customized short TR on 43 loci) through a user-friendly, simple click interface (Fig 1). The prediction of MTBC lineages/sublineages and potential AMRs based on the KvarQ script [22] is also included in TGS-TB web tools. Multiple NGS data obtained from outbreak strains can be accepted through TGS-TB, and the discrimination of strain-specific genotypes can be elucidated to investigate outbreaks, thereby contributing to TB surveillance.

Results and Discussion

Although WGS provides an adequate solution for molecular epidemiology, traditional genotyping methods are still used effectively. We have developed an all-in-one bioinformatics tool to implement both traditional and newly developed techniques (Fig 1). MiSeq is one of the best sequencers to accomplish this project, as MiSeq provides sufficient accuracy to assign SNVs and AMR-related genetic alterations and longer read lengths, up to 300-mer, to assign spoligotypes, IS*6110* insertion sites and sMIRU-VNTRs. Indeed, 350-mer (read1) x 250-mer (read2) paired-end sequencing resulted in 70% of 300-mer nucleotide sequences with Phred quality scores above 30 (S1 Fig), indicating sufficient accuracy for multiple *in silico* genotypings.

Basic function of TGS-TB

Paired-end fastq.gz files can be uploaded to the TGS-TB web page, and an e-mail announcement will be sent to users when all analyses are completed. The resulting basic information,

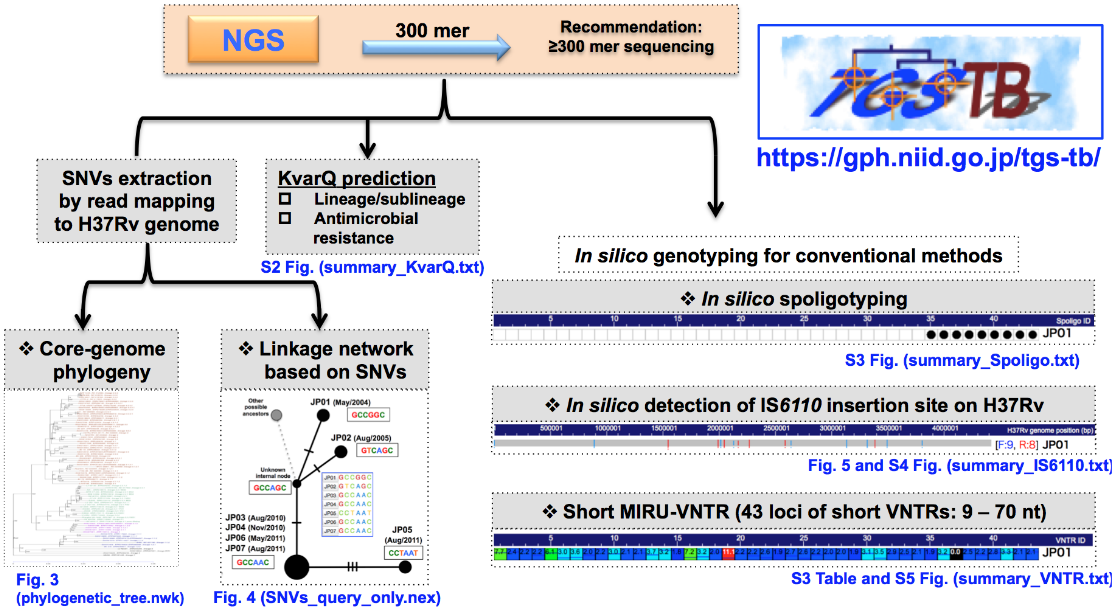

NGS
300 mer
Recommendation: ≥300 mer sequencing
TGS-TB
https://gph.niid.go.jp/tgs-tb/
SNVs extraction by read mapping to H37Rv genome
KvarQ prediction
□ Lineage/sublineage
□ Antimicrobial resistance
S2 Fig. (summary_KvarQ.txt)
In silico genotyping for conventional methods
❖ In silico spoligotyping
S3 Fig. (summary_Spoligo.txt)
❖ In silico detection of IS6110 insertion site on H37Rv
Fig. 5 and S4 Fig. (summary_IS6110.txt)
❖ Short MIRU-VNTR (43 loci of short VNTRs: 9 – 70 nt)
S3 Table and S5 Fig. (summary_VNTR.txt)
JP01
❖ Core-genome phylogeny
Fig. 3 (phylogenetic_tree.nwk)
❖ Linkage network based on SNVs
Other possible ancestors
JP01 (May/2004) GCCGGC
JP02 (Aug/2005) GTCAGC
Unknown internal node GCCAGC
JP03 (Aug/2010)
JP04 (Nov/2010)
JP06 (May/2011)
JP07 (Aug/2011)
GCCAAC
JP05 (Aug/2011) CCTAAT
JP01 GCCGGC
JP02 GTCAGC
JP03 GCCAAC
JP04 GCCAAC
JP05 CCTAAT
JP06 GCCAAC
JP07 GCCAAC
Fig. 4 (SNVs_query_only.nex)

Fig 1. Schematic representation of the TGS-TB system.

doi:10.1371/journal.pone.0142951.g001

such as number of trimmed map reads and the coverage region depth, is reported for all tested isolates (Fig 2). The respective results for lineage, AMR, core genome phylogeny (maximum-likelihood method), linkage network of outbreak strains, IS*6110* insertion, spoligotyping and sMIRU-VNTR typing can be viewed in a new window tab and can also be retrieved using the "download all" button.

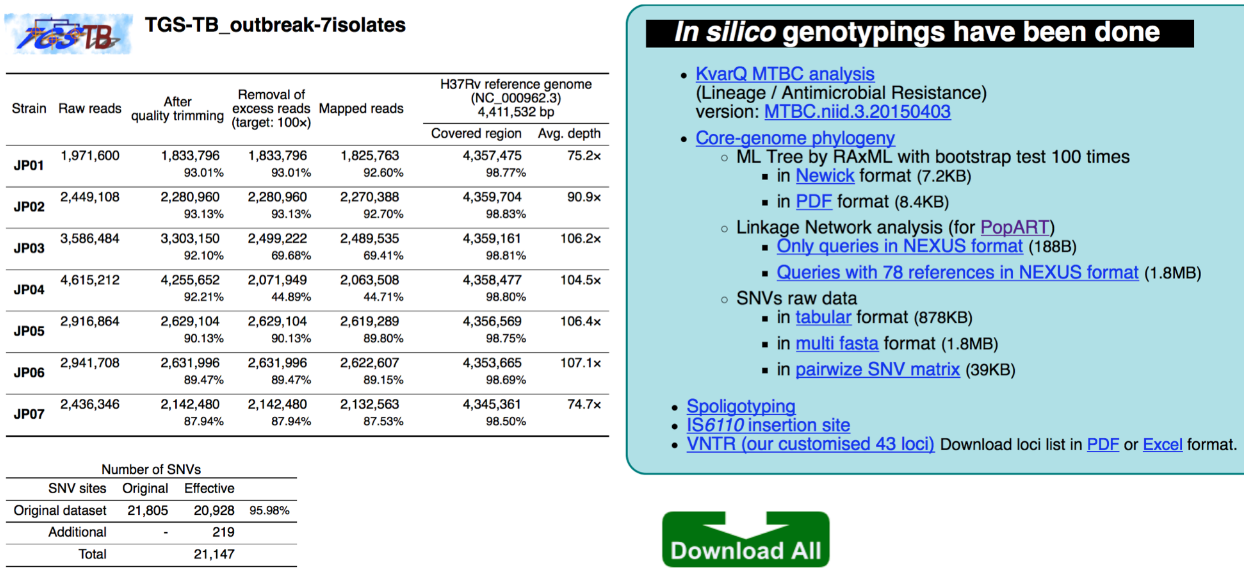
Core-genome phylogeny

To further characterize Mtb lineage analysis comprehensively based on core genome phylogeny, complete/draft genome sequences and short read archives for 78 isolates (S1 Table) were selected from ~2,400 public available short read archives (SRA) of Mtb strains.

In total, 21,805 core genome SNVs on the non-repetitive regions (S2 Table) are available in the TGS-TB, and 20,928 (95.98%) SNVs for the seven isolates (JP01 to JP07) were correctly extracted in the sample test (Fig 2). Additional query-specific novel SNVs can be identified in TGS-TB, and 219 additional strain-specific SNV sites can be implemented in the original dataset (21,805 core genome SNVs). A maximum-likelihood core genome phylogenetic tree is constructed based on the whole SNV dataset, including newly identified SNVs sites (Fig 3). The data for the original and query-specific SNVs can be downloaded as a tab-delimited file (summary_SNVs.txt) or fasta file (phylocoreGenome.fasta) for further phylogenetic analysis using more bootstrapping analyses or a Bayesian approach.

Linkage network analysis using SNVs

Although the core genome phylogeny database includes 78 references, it is too computationally intensive to determine the precise differences for only outbreak-related query isolates. Thus, query-specific SNV sites are simple datasets used to perform further epidemiological investigations. Among the tested seven isolates (four from the outbreak and three from related isolates with similar MIRU-VNTR loci [26]), 6 query-specific SNVs were extracted to investigate the molecular epidemiological markers to trace back the outbreak between patients. The



TGS-TB_outbreak-7isolates

Strain	Raw reads	After quality trimming	Removal of excess reads (target: 100×)	Mapped reads	H37Rv reference genome (NC_000962.3) 4,411,532 bp	
					Covered region	Avg. depth
JP01	1,971,600	1,833,796 93.01%	1,833,796 93.01%	1,825,763 92.60%	4,357,475 98.77%	75.2×
JP02	2,449,108	2,280,960 93.13%	2,280,960 93.13%	2,270,388 92.70%	4,359,704 98.83%	90.9×
JP03	3,586,484	3,303,150 92.10%	2,499,222 69.68%	2,489,535 69.41%	4,359,161 98.81%	106.2×
JP04	4,615,212	4,255,652 92.21%	2,071,949 44.89%	2,063,508 44.71%	4,358,477 98.80%	104.5×
JP05	2,916,864	2,629,104 90.13%	2,629,104 90.13%	2,619,289 89.80%	4,356,569 98.75%	106.4×
JP06	2,941,708	2,631,996 89.47%	2,631,996 89.47%	2,622,607 89.15%	4,353,665 98.69%	107.1×
JP07	2,436,346	2,142,480 87.94%	2,142,480 87.94%	2,132,563 87.53%	4,345,361 98.50%	74.7×

Number of SNVs

SNV sites	Original	Effective	
Original dataset	21,805	20,928	95.98%
Additional	-	219	
Total		21,147	

***In silico* genotypings have been done**

- KvarQ MTBC analysis
  (Lineage / Antimicrobial Resistance)
  version: MTBC.niid.3.20150403
- Core-genome phylogeny
  - ML Tree by RAxML with bootstrap test 100 times
    - in Newick format (7.2KB)
    - in PDF format (8.4KB)
  - Linkage Network analysis (for PopART)
    - Only queries in NEXUS format (188B)
    - Queries with 78 references in NEXUS format (1.8MB)
  - SNVs raw data
    - in tabular format (878KB)
    - in multi fasta format (1.8MB)
    - in pairwize SNV matrix (39KB)
- Spoligotyping
- IS*6110* insertion site
- VNTR (our customised 43 loci) Download loci list in PDF or Excel format.

Fig 2. Sample results obtained from TGS-TB. The NGS reads of seven Mtb isolates were investigated, and the resulting basic information, such as number of trimmed map reads and the coverage region-depth, is shown. In total, 21,805 core-genome SNVs are available in the TGS-TB; 20,928 (95.98%) SNVs are characterized, and 219 additional strain-specific SNVs sites can be implemented in the original dataset. The respective results for lineage, AMR, core genome phylogenetic tree (maximum-likelihood method with x100 bootstrapping), spoligotyping, IS*6110* insertion and sMIRU-VNTR typing can be viewed in a new window tab and retrieved using the "download all" button. The KvarQ script predicts the lineages/sublineages and AMRs, and the sample queries are assigned as a lineage 2/Beijing sublineage without AMRs (S2 Fig). The AMR target list in the original KvarQ (v2.0) program has been improved with the addition of more reliable genetic alterations for the *embA*, *gyrA*, *katG*, *pncA*, *rpoB*, *rpsL*, *rrs* and *inhA* genes (S1 Text).

doi:10.1371/journal.pone.0142951.g002

differences based on these 6 SNV sites can be visualized through a median-joining network using PopART software (Fig 4). The MIRU-VNTR loci of the four outbreak isolates (JP03, JP04, JP06 and JP07) were investigated prior to NGS, indicating that three additional isolates (JP01, JP02 and JP05) were closely related to the four outbreak isolates. The MIRU-VNTR results did not show sufficient discrimination among the seven isolates ranging from 2004 to 2011. NGS and TGS-TB analyses were performed to identify the SNVs among those isolates. No unique SNVs were detected for the four outbreak isolates (JP03, JP04, JP06 and JP07), and the two past isolates (JP01 and JP02) with two SNVs difference could be one outbreak source or potential ancestors related to this outbreak. It has been reported that the estimated rate of change in DNA sequences was 0.3–0.5 single nucleotide polymorphisms per genome per year [9, 13, 17, 18]; thus, the two detected SNVs represent a reasonable nucleotide substitution rate between the outbreak (2010/2011) and past isolates (2004/2005). In contrast, JP05 was isolated at a close time point, with four outbreak isolates in December 2010, and three SNV sites were detected when compared with four outbreak isolates. One additional IS*6110* insertion was detected in JP05 (Figs 4B and 5), strongly suggesting that JP05 did not appear to be involved in the outbreak. Such SNV networks facilitate the detection of epidemiological factors, and the obtained network is consistent with the field study (data not shown).

In silico genotyping instead of conventional methods

In addition to Mtb genotyping using conventional methods, detection of the IS*6110* insertion (S3 Fig) and *in silico* spoligotyping (S4 Fig) are also available using the all-in-one TGS-TB analysis. *In silico* detection of the IS*6110* insertion provides precise information compared with IS*6110*-RFLP genotyping. The determination of the insertion sites and the orientation on the Mtb genome suggested the distinct type of genotypic evolution involving core genome SNVs,

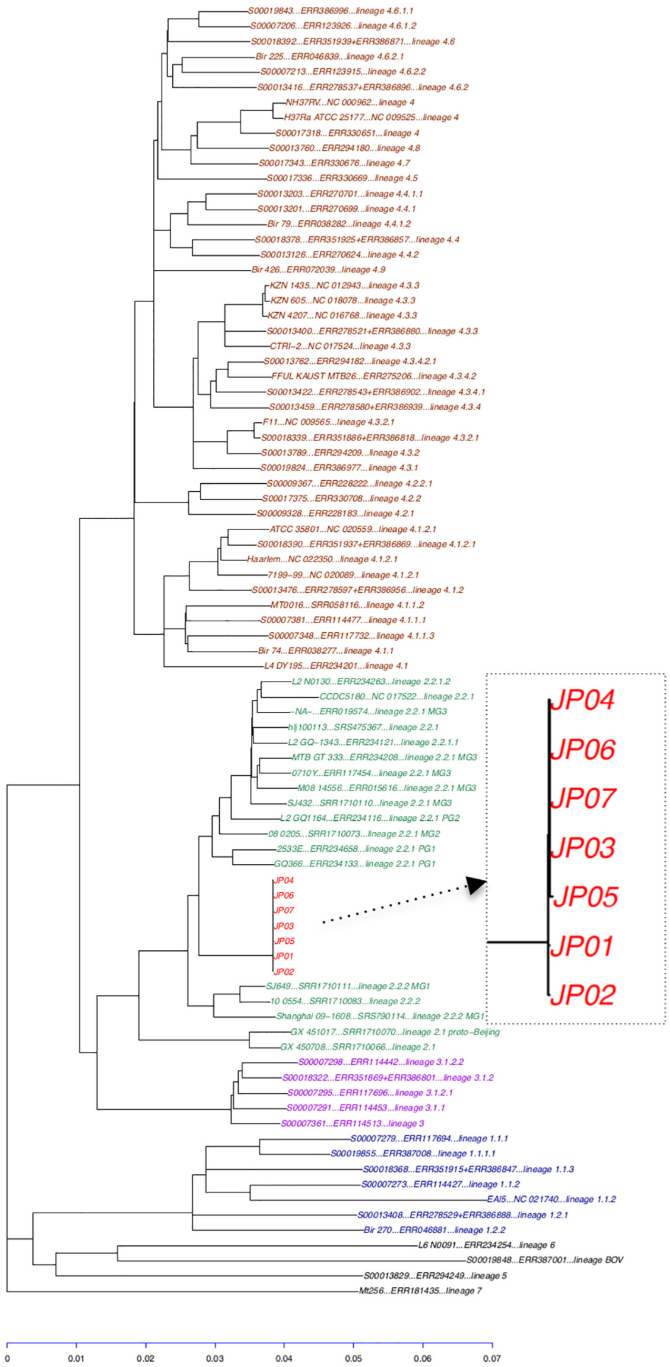

S00019843...ERR386996...lineage 4.6.1.1
S00007206...ERR123926...lineage 4.6.1.2
S00018392...ERR351939+ERR386871...lineage 4.6
Bir 225...ERR046839...lineage 4.6.2.1
S00007213...ERR123915...lineage 4.6.2.2
S00013416...ERR278537+ERR386896...lineage 4.6.2
NH37RV...NC 000962...lineage 4
H37Ra ATCC 25177...NC 009525...lineage 4
S00017318...ERR330651...lineage 4
S00013760...ERR294180...lineage 4.8
S00017343...ERR330676...lineage 4.7
S00017336...ERR330669...lineage 4.5
S00013203...ERR270701...lineage 4.4.1.1
S00013201...ERR270699...lineage 4.4.1
Bir 79...ERR038282...lineage 4.4.1.2
S00018378...ERR351925+ERR386857...lineage 4.4
S00013126...ERR270624...lineage 4.4.2
Bir 426...ERR072039...lineage 4.9
KZN 1435...NC 012943...lineage 4.3.3
KZN 605...NC 018078...lineage 4.3.3
KZN 4207...NC 016768...lineage 4.3.3
S00013400...ERR278521+ERR386880...lineage 4.3.3
CTRI–2...NC 017524...lineage 4.3.3
S00013762...ERR294182...lineage 4.3.4.2.1
FFUL KAUST MTB26...ERR275206...lineage 4.3.4.2
S00013422...ERR278543+ERR386902...lineage 4.3.4.1
S00013459...ERR278580+ERR386939...lineage 4.3.4
F11...NC 009565...lineage 4.3.2.1
S00018339...ERR351886+ERR386818...lineage 4.3.2.1
S00013789...ERR294209...lineage 4.3.2
S00019824...ERR386977...lineage 4.3.1
S00009367...ERR228222...lineage 4.2.2.1
S00017375...ERR330708...lineage 4.2.2
S00009328...ERR228183...lineage 4.2.1
ATCC 35801...NC 020559...lineage 4.1.2.1
S00018390...ERR351937+ERR386869...lineage 4.1.2.1
Haarlem...NC 022350...lineage 4.1.2.1
7199–99...NC 020089...lineage 4.1.2.1
S00013476...ERR278597+ERR386956...lineage 4.1.2
MT0016...SRR058116...lineage 4.1.1.2
S00007381...ERR114477...lineage 4.1.1.1
S00007348...ERR117732...lineage 4.1.1.3
Bir 74...ERR038277...lineage 4.1.1
L4 DY195...ERR234201...lineage 4.1
L2 N0130...ERR234263...lineage 2.2.1.2
CCDC5180...NC 017522...lineage 2.2.1
–NA–...ERR019574...lineage 2.2.1 MG3
hlj100113...SRS475367...lineage 2.2.1
L2 GQ–1343...ERR234121...lineage 2.2.1.1
MTB GT 333...ERR234208...lineage 2.2.1 MG3
0710Y...ERR117454...lineage 2.2.1 MG3
M08 14556...ERR015616...lineage 2.2.1 MG3
SJ432...SRR171011 0...lineage 2.2.1 MG3
L2 GQ1164...ERR234116...lineage 2.2.1 PG2
08 0205...SRR1710073...lineage 2.2.1 MG2
2533E...ERR234658...lineage 2.2.1 PG1
GQ366...ERR234133...lineage 2.2.1 PG1
JP04
JP06
JP07
JP03
JP05
JP01
JP02
SJ649...SRR1710111...lineage 2.2.2 MG1
10 0554...SRR1710083...lineage 2.2.2
Shanghai 09–1608...SRS790114...lineage 2.2.2 MG1
GX 451017...SRR1710070...lineage 2.1 proto–Beijing
GX 450708...SRR1710066...lineage 2.1
S00007298...ERR114442...lineage 3.1.2.2
S00018322...ERR351869+ERR386801...lineage 3.1.2
S00007295...ERR117696...lineage 3.1.2.1
S00007291...ERR114453...lineage 3.1.1
S00007361...ERR114513...lineage 3
S00007279...ERR117694...lineage 1.1.1
S00019855...ERR387008...lineage 1.1.1.1
S00018368...ERR351915+ERR386847...lineage 1.1.3
S00007273...ERR114427...lineage 1.1.2
EAI5...NC 021740...lineage 1.1.2
S00013408...ERR278529+ERR386888...lineage 1.2.1
Bir 270...ERR046881...lineage 1.2.2
L6 N0091...ERR234254...lineage 6
S00019848...ERR387001...lineage BOV
S00013829...ERR294249...lineage 5
Mt256...ERR181435...lineage 7
JP04
JP06
JP07
JP03
JP05
JP01
JP02
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07

Fig 3. The core genome phylogeny obtained by the maximum-likelihood method with x100 bootstrapping.

doi:10.1371/journal.pone.0142951.g003

implying that novel Mtb clones can be easily detected as variations of the IS*6110* insertion. One of the Mtb isolates, JP05, showed a heterozygous population of the IS*6110* insertion sequencing, showing additional insertions at minor sites. Approximately 10% of the population might be mixed or contaminated in the original isolates, implying that the detection of the hetero-population could be valuable in distinguishing Mtb mixed-infections.

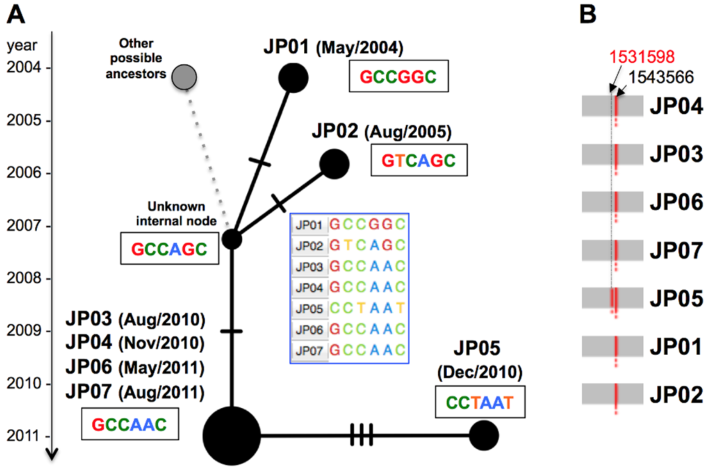


Fig 4. Median-joining network of the seven outbreak isolates based on the detected core genomic variations. A) The variations are summarized as nexus format files (.nex), and PopART visualizes the epidemiological linkages among the isolates through a user specified network method. The bars on the edge indicate the number of SNVs between the nodes (isolates). B) In addition to three SNV differences between JP05 and the outbreak isolates (JP03, JP04, JP06, and JP07), an additional IS*6110* insertion was detected at the 1,531,598 nt genome position in JP05, suggesting that JP05 could be unrelated to the outbreak, although the VNTR profile is consistent.

doi:10.1371/journal.pone.0142951.g004

Short MIRU-VNTR (sMIRU-VNTR)

MIRU-VNTR is a valuable discriminative genotyping tool, but NGS short reads are not a sufficient length to assign all loci of MIRU-VNTR (15 or 24 loci set). We originally assigned the 43 loci of short tandem repeats (TR), including 10 loci (MIRU02, ETRC, MIRU10, MIRU16, MIRU20, Mtub30, ETRB, MIRU27, Mtub34 and MIRU39) of the 24-loci MIRU-VNTR (S3 Table and S5 Fig). The selected loci are short TRs, designated as short MIRU-VNTRs (sMIRU-VNTRs). The detected repeats units are not always integers due to partial sequence repeat units.

Conclusions

The results of *in silico* analysis through TGS-TB are consistent with conventional molecular genotyping methods (S6 Fig), suggesting that MiSeq NGS short reads could provide sufficient multiple genotypes at once to discriminate multiple strains of Mtb during an outbreak. The WGS approach is highly affordable for the characterization of the Mtb strains described above, but comprehensive genotyping tools, including traditional genotyping, have been expected to integrate all genetic information thus far. The all-in-one tool, TGS-TB, provides a more accurate and discriminative strain typing for clinical and epidemiological investigations. Thus, NGS strain typing could offer a total genotyping solution for Mtb outbreak and surveillance. The genotype information obtained for all Mtb isolates can be deposited into an integrated database for the future surveillance of outbreak and global infections.

Materials and Methods

Mycobacterium tuberculosis(Mtb) strains

The complete and draft genomic sequences for Mtb strains were retrieved from the GenBank database and NCBI SRA projects (ERP000111, ERP000192, ERP000276, ERP000436, ERP000520, ERP001567, ERP001731, ERP001885, ERP002611, SRP002589), and the following *in silico* analyses were performed using H37Rv (NC_000962.3) as a reference genome

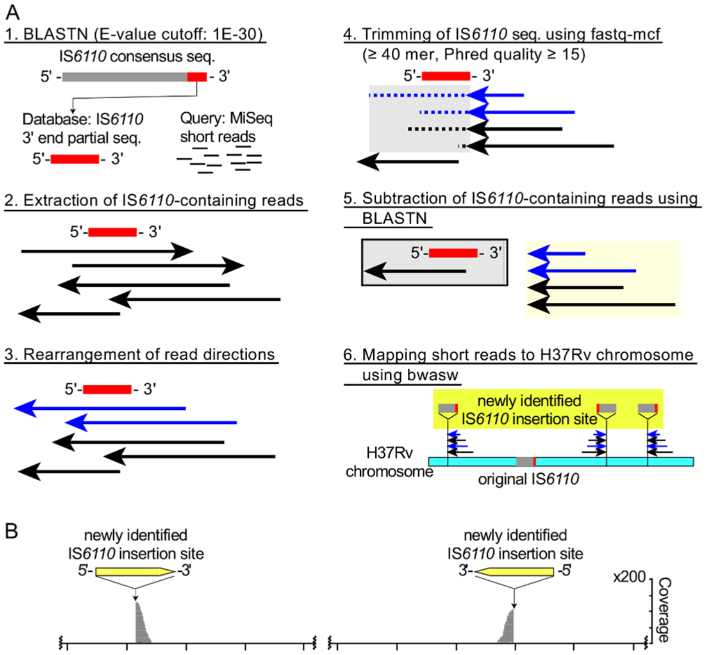

A
1. BLASTN (E-value cutoff: 1E-30)
IS6110 consensus seq.
5' - - 3'
Database: IS6110 3' end partial seq.
5'- - 3'
Query: MiSeq short reads
2. Extraction of IS6110-containing reads
5'- - 3'
3. Rearrangement of read directions
5'- - 3'
4. Trimming of IS6110 seq. using fastq-mcf
(≥ 40 mer, Phred quality ≥ 15)
5'- - 3'
5. Subtraction of IS6110-containing reads using BLASTN
5'- - 3'
6. Mapping short reads to H37Rv chromosome using bwasw
newly identified IS6110 insertion site
H37Rv chromosome
original IS6110
B
newly identified IS6110 insertion site
5'- -3'
newly identified IS6110 insertion site
3'- -5'
x200
Coverage

Fig 5. Schematic representation of the IS*6110* insertion detection strategy. A) The IS*6110* sequence (Acc.# X94955 and X94956)-positive short reads are collected (A1-2), rearranged (A3), trimmed (A4), subtracted (A5) and mapped to the Mtb H37Rv chromosome (NC_000962.3) [30] using BWA-SW mapping [29]. B) Typical read mapping profile for the detection of the IS*6110* insertion site in both directions.

doi:10.1371/journal.pone.0142951.g005

sequence. The performance and accuracy of TGS-TB were examined using seven Japanese Mtb isolates associated with local outbreaks (JP01, JP02, JP03, JP04, JP05, JP06 and JP07) (S1 Table).

NGS

Genomic DNA from Mtb isolates was purified through the benzyl chloride method using the ISOPLANT kit (WAKO, Osaka, Japan). A genomic DNA library for NGS was prepared using the Nextera XT DNA Sample Prep Kit (Illumina, San Diego, CA, USA), followed by insert size selection using 1% TAE agarose electrophoresis to obtain an insert of approximately 600 bp. Sequencing was performed on MiSeq (Illumina, San Diego, CA, USA) using the MiSeq Reagent Kit v3 (600 cycle) with 350-mer x 250-mer paired-end short reads and 96-sample multiplexing (S1 Fig).

KvarQ prediction

MTBC lineages/sublineages and potential AMRs were determined using the KvarQ script according to the manufacturer's instructions [22]. The AMR target list has been improved to detect more reliable genetic alterations for the *embA*, *gyrA*, *katG*, *pncA*, *rpoB*, *rpsL*, *rrs* and *inhA* genes from TBDReaMDB [27] (S1 Text).

Core-genome phylogenetic and linkage network analysis

Prior to *in silico* genotyping, the adapter sequences were trimmed from the short reads, and low quality bases with a Phred score less than 15 were eliminated using the Skewer program, to obtain sequences at least 50-mer in length [28]. The remaining reads are mapped using the

BWA-mem program [29] with the Mtb H37Rv chromosome (NC_000962.3) reference genome sequence [30]. Reliable SNV sites with at least a 5x coverage depth and a Phred score of at least 20 were selected from the mapping. The SNV sites on the repeat regions of the H37RV genome (S2 Table), which was assigned based on the GenBank annotation file (NC_000962.3) and newly identified by the NUCmer program [31], was excluded for further core-genome phylogeny analysis because those SNVs sites are considered unreliable. A total of 21,805 SNV sites were extracted as an original dataset from Mtb with 12 complete genomes and 66 SRAs (S1 Table and Fig 2). Maximum likelihood phylogenetic analysis of all concatenated SNV alleles was performed using RAxML v8.2.0 [32] with 1,000 bootstrap iterations.

To investigate the epidemiological linkage between patients, the information from queries for isolate-specific genes or the above mentioned reference genomes can be downloaded as a nexus format file to visualize linkage networks, such as the median-joining network method using PopART (http://popart.otago.ac.nz).

In silico detection of IS*6110* insertion site

The IS*6110* insertion site was detected in the following manner (Fig 5A): IS*6110* sequence positive short reads are collected from all short reads (Fig 5A1–2); the direction of collected reads was rearranged to the direction of the IS*6110* sequence (Fig 5A3); trimming and subtraction of the IS*6110* sequence (Acc.# X94955 and X94956) was performed according to an adapter trimming procedure (Fig 5A4–5); and the resulting trimmed short reads were mapped to the Mtb H37Rv chromosome (NC_000962.3) [30] through BWA-SW mapping [29] (Fig 5A6). The insertion site was detected at high coverage peaks on the chromosome DNA (Fig 5B).

In silico spoligotyping

In silico spoligotyping was performed through a blastn search using 43 spacer sequences as a query [4] against the obtained short reads, and double mismatches with homology were considered a positive threshold.

Short MIRU-VNTR

Instead of MIRU-VNTR [7, 19], we demonstrated that short VNTR loci can be used for further genotyping through short read sequencing. Possible candidates for additional VNTR loci with core sequences of ≥9 bp were extracted from the H37Rv genome according to the microorganisms tandem repeats database (http://minisatellites-rec.igmors.u-psud.fr/GPMS/) [33] using the default settings. Although NGS short reads are not sufficient for conventional MIRU-VNTR (MtbC15-9 type) because of the use of sequences up to 300 bp in length, 43 loci can be reasonably assigned as discriminative VNTR loci. Variable repeat units were assigned as follows: the corresponding short reads for respective VNTR loci were collected, followed by trimming of unique core genomic sequences at both ends of the TR. Using the remaining TR sequences, variable repeat units were counted based on the repeat core sequence. The selected loci were short repeat units, referred to as short MIRU-VNTR (sMIRU-VNTR) (S3 Table).

Conventional molecular genotyping methods

Three molecular genotyping methods based on spoligotyping, VNTR and IS*6110*-RFLP were performed on the seven Japanese clinical strains (JP01 to JP07). Spoligotyping was performed according to the standardized protocol using an in-house membrane [4]. An optimized 24-loci MIRU-VNTR analysis [7] was conducted as previously described [26], and IS*6110*-RFLP was performed according to a standardized protocol [3].

Supporting Information

S1 Fig. Web image on the BaseSpace analysis for MiSeq sequencing. NGS was performed on a MiSeq NGS sequencer (Illumina, San Diego, CA, USA) using the MiSeq Reagent Kit v3 (600 cycle) with 350-mer x 250-mer paired-end short reads.
(PDF)

S2 Fig. KvarQ prediction for lineages/sublineages and antimicrobial resistance.
(PDF)

S3 Fig. The results of the *in silico* detection of IS*6110* insertion sites. The red and sky-blue vertical bars on the H37Rv reference genome indicate the forward and reverse IS*6110* insertions, respectively.
(PDF)

S4 Fig. The results of *in silico* spoligotyping using 43 spacer oligos. The filled circles indicate positive homology to each oligonucleotide sequence.
(PDF)

S5 Fig. The results of *in silico* sMIRU-VNTR typing using 43 customized loci with tandem repeats (TRs). The detected number of TRs is shown on each locus and visualized using a color variation scale. Black and gray boxes indicate no detection of TRs and lower depths, respectively.
(PDF)

S6 Fig. Conventional molecular genotyping results for the seven outbreak isolates examined in this study.
(PDF)

S1 Table. List of Mtb strains used in this study with information of the MTBC lineage/sublineage.
(PDF)

S2 Table. Repeat regions in the *M. tuberculosis* H37Rv genome (NC_000962.3).
(PDF)

S3 Table. Locus information for short MIRU-VNTR.
(PDF)

S1 Text. # Version: MTBC.niid.3.20150403 #. The AMR target list for KvarQ prediction has been improved to detect more reliable genetic alterations for the *embA*, *gyrA*, *katG*, *pncA*, *rpoB*, *rpsL*, *rrs* and *inhA* genes from TBDReaMDB [27].
(TXT)

Acknowledgments

The authors are grateful to a Public Health Center in Shinjuku City for providing the clinical isolates used in this study. The study protocol was approved by the institutional medical ethics committee of the Research Institute of Tuberculosis/Japan Anti-Tuberculosis Association in Japan (Approval No. RIT/IRB 27–10).

Author Contributions

Conceived and designed the experiments: TS YM SM SK MK. Performed the experiments: TS YM. Analyzed the data: YM TI TS. Contributed reagents/materials/analysis tools: TS YM AY. Wrote the paper: TS YM MK.

References

1. Global tuberculosis report. [Internet]. http://www.who.int/tb/publications/global_report/gtbr13_main_text.pdf (accessed Jan 10, 2014). 2013

2. Barnes PF, Cave MD. Molecular epidemiology of tuberculosis. The New England journal of medicine. 2003; 349(12):1149–56. doi: 10.1056/NEJMra021964 PMID: 13679530.

3. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. Journal of clinical microbiology. 1993; 31(2):406–9. PMID: 8381814; PubMed Central PMCID: PMC262774.

4. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. Journal of clinical microbiology. 1997; 35(4):907–14. PMID: 9157152; PubMed Central PMCID: PMC229700.

5. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PW, Martin C, et al. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. Journal of clinical microbiology. 1999; 37(8):2607–18. PMID: 10405410; PubMed Central PMCID: PMC85295.

6. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. Journal of clinical microbiology. 2001; 39(10):3563–71. doi: 10.1128/JCM.39.10.3563-3571.2001 PMID: 11574573; PubMed Central PMCID: PMC88389.

7. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. Journal of clinical microbiology. 2006; 44(12):4498–510. doi: 10.1128/JCM.01392-06 PMID: 17005759; PubMed Central PMCID: PMC1698431.

8. Allix-Beguec C, Wahl C, Hanekom M, Nikolayevskyy V, Drobniewski F, Maeda S, et al. Proposal of a consensus set of hypervariable mycobacterial interspersed repetitive-unit-variable-number tandem-repeat loci for subtyping of *Mycobacterium tuberculosis* Beijing isolates. Journal of clinical microbiology. 2014; 52(1):164–72. doi: 10.1128/JCM.02519-13 PMID: 24172154; PubMed Central PMCID: PMC3911419.

9. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. The Lancet Infectious diseases. 2013; 13(2):137–46. doi: 10.1016/S1473-3099(12)70277-3 PMID: 23158499; PubMed Central PMCID: PMC3556524.

10. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nature reviews Genetics. 2012; 13(9):601–12. doi: 10.1038/nrg3226 PMID: 22868263.

11. Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. The Lancet Respiratory medicine. 2014; 2(4):285–92. doi: 10.1016/S2213-2600(14)70027-X PMID: 24717625.

12. Ford C, Yusim K, Ioerger T, Feng S, Chase M, Greene M, et al. *Mycobacterium tuberculosis*—heterogeneity revealed through whole genome sequencing. Tuberculosis. 2012; 92(3):194–201. doi: 10.1016/j.tube.2011.11.003 PMID: 22218163; PubMed Central PMCID: PMC3323677.

13. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. The Lancet Respiratory medicine. 2013; 1(10):786–92. doi: 10.1016/S2213-2600(13)70231-5 PMID: 24461758; PubMed Central PMCID: PMC3861685.

14. Luo T, Yang C, Peng Y, Lu L, Sun G, Wu J, et al. Whole-genome sequencing to detect recent transmission of *Mycobacterium tuberculosis* in settings with a high burden of tuberculosis. Tuberculosis. 2014; 94(4):434–40. doi: 10.1016/j.tube.2014.04.005 PMID: 24888866.

15. Roetzer A, Diel R, Kohl TA, Ruckert C, Nubel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS medicine. 2013; 10(2):e1001387. doi: 10.1371/journal.pmed.1001387 PMID: 23424287; PubMed Central PMCID: PMC3570532.

16. Niemann S, Koser CU, Gagneux S, Plinke C, Homolka S, Bignell H, et al. Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PloS one. 2009; 4(10):e7407. doi: 10.1371/journal.pone.0007407 PMID: 19823582; PubMed Central PMCID: PMC2756628.

17. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, et al. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. Nature genetics. 2011; 43(5):482–6. doi: 10.1038/ng.811 PMID: 21516081; PubMed Central PMCID: PMC3101871.

18. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis. 2013; 13:110. doi: 10.1186/1471-2334-13-110 PMID: 23446317; PubMed Central PMCID: PMC3599118.

19. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic acids research. 2010; 38(Web Server issue):W326–31. doi: 10.1093/nar/gkq351 PMID: 20457747; PubMed Central PMCID: PMC2896200.

20. Kohl TA, Diel R, Harmsen D, Rothganger J, Walter KM, Merker M, et al. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. Journal of clinical microbiology. 2014; 52(7):2479–86. doi: 10.1128/JCM.00567-14 PMID: 24789177; PubMed Central PMCID: PMC4097744.

21. Coll F, Mallard K, Preston MD, Bentley S, Parkhill J, McNerney R, et al. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. Bioinformatics. 2012; 28(22):2991–3. doi: 10.1093/bioinformatics/bts544 PMID: 23014632; PubMed Central PMCID: PMC3496340.

22. Steiner A, Stucki D, Coscolla M, Borrell S, Gagneux S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC genomics. 2014; 15:881. doi: 10.1186/1471-2164-15-881 PMID: 25297886; PubMed Central PMCID: PMC4197298.

23. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. Journal of clinical microbiology. 2015. doi: 10.1128/JCM.00025-15 PMID: 25854485.

24. Coll F, Preston M, Guerra-Assuncao JA, Hill-Cawthorn G, Harris D, Perdigao J, et al. PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis. 2014; 94(3):346–54. doi: 10.1016/j.tube.2014.02.005 PMID: 24637013; PubMed Central PMCID: PMC4066953.

25. Benavente ED, Coll F, Furnham N, McNerney R, Glynn JR, Campino S, et al. PhyTB: Phylogenetic tree visualisation and sample positioning for *M*. *tuberculosis*. BMC bioinformatics. 2015; 16:155. doi: 10.1186/s12859-015-0603-3 PMID: 25968323; PubMed Central PMCID: PMCPMC4429496.

26. Murase Y, Mitarai S, Sugawara I, Kato S, Maeda S. Promising loci of variable numbers of tandem repeats for typing Beijing family *Mycobacterium tuberculosis*. Journal of medical microbiology. 2008; 57(Pt 7):873–80. doi: 10.1099/jmm.0.47564-0 PMID: 18566146.

27. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS medicine. 2009; 6(2):e2. doi: 10.1371/journal.pmed.1000002 PMID: 19209951; PubMed Central PMCID: PMC2637921.

28. Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC bioinformatics. 2014; 15:182. doi: 10.1186/1471-2105-15-182 PMID: 24925680; PubMed Central PMCID: PMC4074385.

29. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26(5):589–95. doi: 10.1093/bioinformatics/btp698 PMID: 20080505; PubMed Central PMCID: PMC2828108.

30. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998; 393(6685):537–44. doi: 10.1038/31159 PMID: 9634230.

31. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5(2):R12. doi: 10.1186/gb-2004-5-2-r12 PMID: 14759262; PubMed Central PMCID: PMC395750.

32. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. doi: 10.1093/bioinformatics/btu033 PMID: 24451623; PubMed Central PMCID: PMC3998144.

33. Denoeud F, Vergnaud G. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. BMC bioinformatics. 2004; 5:4. doi: 10.1186/1471-2105-5-4 PMID: 14715089; PubMed Central PMCID: PMC331396.