OMICS A Journal of Integrative Biology
Volume 21, Number 7, 2017

DOI: 10.1089/omi.2017.0070

Whole Genome Sequencing of *Mycobacterium tuberculosis* Isolates From Extrapulmonary Sites

Kusum Sharma,[1,*] Renu Verma,[2,3,*] Jayshree Advani,[2,4,*] Oishi Chatterjee[2,5] Hitendra S. Solanki,[2,3] Aman Sharma,[6] Subhash Varma,[6] Manish Modi,[7] Pallab Ray,[1] Kanchan K. Mukherjee,[8] Megha Sharma,[1] Mandeed Singh Dhillion,[9] Mrutyunjay Suar,[3] Aditi Chatterjee,[2,10] Akhilesh Pandey,[2,11–14] Thottethodi Subrahmanya Keshava Prasad,[2,10,15] and Harsha Gowda[2,10]

Abstract

Tuberculosis (TB) remains one of the leading causes of morbidity and mortality worldwide. Extrapulmonary tuberculosis (EPTB) constitutes around 15–20% of TB cases in immunocompetent individuals. Extrapulmonary sites that are affected by TB include bones, lymph nodes, meningitis, pleura, and genitourinary tract. Whole genome sequencing has emerged as a powerful tool to map genetic diversity among *Mycobacterium tuberculosis* (MTB) isolates and identify the genomic signatures associated with drug resistance, pathogenesis, and disease transmission. Several pulmonary isolates of MTB have been sequenced over the years. However, availability of whole genome sequences of MTB isolates from extrapulmonary sites is limited. Some studies suggest that genetic variations in MTB might contribute to disease presentation in extrapulmonary sites. This can be addressed if whole genome sequence data from large number of extrapulmonary isolates becomes available. In this study, we have performed whole genome sequencing of five MTB clinical isolates derived from EPTB sites using next-generation sequencing platform. We identified 1434 nonsynonymous single nucleotide variations (SNVs), 143 insertions and 105 deletions. This includes 279 SNVs that were not reported before in publicly available datasets. We found several mutations that are known to confer resistance to drugs. All the five isolates belonged to East-African-Indian lineage (lineage 3). We identified 9 putative prophage DNA integrations and 14 predicted clustered regularly interspaced short palindromic repeats (CRISPR) in MTB genome. Our analysis indicates that more work is needed to map the genetic diversity of MTB. Whole genome sequencing in conjunction with comprehensive drug susceptibility testing can reveal clinically relevant mutations associated with drug resistance.

Keywords: coding DNA sequence, lineage, lymphadenitis, nonsynonymous, octal code

Introduction

Diagnosis and treatment of tuberculosis (TB) are regarded as major public health concern. According to World Health Organization (WHO) report, an estimated 10.4 million people developed TB and 1.4 million died from the disease in 2016. Emergence of drug resistance TB has further complicated management and control of the disease. Globally, 3.9% of new and 21% of previously treated TB cases were estimated to have multidrug-resistant TB (MDR-TB; WHO, 2016 TB report). Tuberculosis bacilli can infect any organ system in the body with pulmonary TB being its most

[1]Department of Medical Microbiology, PGIMER, Chandigarh, India.
[2]Institute of Bioinformatics, International Technology Park, Bangalore, India.
[3]School of Biotechnology, KIIT University, Bhubaneswar, India.
[4]Manipal University, Manipal, India.
[5]School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam, India.
Departments of [6]Internal Medicine, [7]Neurology, [8]Neurosurgery, and [9]Orthopedics, PGIMER, Chandigarh, India.
[10]YU-IOB Center for Systems Biology and Molecular Medicine, Mangalore, India.
[11]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland.
Departments of [12]Biological Chemistry, [13]Pathology, and [14]Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland.
[15]NIMHANS-IOB Proteomics and Bioinformatics Laboratory, Neurobiology Research Centre, National Institute of Mental Health and Neurosciences, Bangalore, India.
*These authors made an equal contribution as first authors.

common presentation. However, extrapulmonary tuberculosis (EPTB) including tuberculous meningitis, lymphadenitis, pleuritis, and endometritis are some of the other clinical manifestations (Fanning, 1999; Golden and Vikram, 2005; Meghdadi et al., 2015). In most cases, extrapulmonary infections are considered as continuation of primary pulmonary infection (Das et al., 2013). According to some estimates, EPTB accounts for 15–20% of the cases among immunocompetent and ~50% among immunocompromised individuals in India.

It has been proposed previously that pulmonary TB and EPTB have different underlying pathophysiology (Hasan et al., 2005; Oki et al., 2011) Genetic variations in host genes such as vitamin D receptor (Wilkinson et al., 2000), P2X7 (Fernando et al., 2007), interleukin-10, and interferon-γ (Ansari et al., 2009; Tso et al., 2005) have been proposed to play a role in genetic predisposition to EPTB. In addition, studies have shown that mutations in molecules associated with important functions relevant to pathogenesis are likely to affect key processes in bacteria such as secretion of proteins and glycolipids, which interact directly with the components of host cell (Glickman et al., 2000; Kolattukudy et al., 1997). Mutations in cell wall components, for example, affect the sliding motility of mycobacteria suggesting the role of cell wall in movement of pathogen (Martinez et al., 1999).

Several epidemiological studies and experiments using animal models have investigated whether specific lineages/genotypes are associated with disease presentation in extrapulmonary sites. At the moment, the findings are inconsistent and inconclusive to attribute contribution of specific genetic variations to disease presentation (Coscolla and Gagneux, 2014; Srilohasin et al., 2014). Whole genome sequencing provides superior resolution than currently available genotyping methods such as spoligotyping and mycobacterial interspersed repetitive-unit–variable-number tandem-repeats (MIRU-VNTR)-based methods (van Soolingen et al., 2016). Once whole genome sequencing data from large number of EPTB isolates becomes available, one could investigate whether there are genetic determinants associated with disease presentation in extrapulmonary sites.

Several large-scale whole genome sequencing efforts have been published for isolates from pulmonary infection (Casali et al., 2014; Zhang et al., 2013). However, similar efforts have not been undertaken to carry out whole genome sequencing of clinical isolates from extrapulmonary sites. Lack of data from extrapulmonary isolates has restricted our ability to map underlying genetic diversity among EPTB isolates and to determine if there are specific variations associated with EPTB isolates. Till date, whole genome data from 17 extrapulmonary isolates are available in the public domain (Chernyaeva et al., 2014; Das et al., 2013). In this study, we carried out whole genome sequencing of five clinical isolates from extrapulmonary sites and compared the genetic variations with data from pulmonary and extrapulmonary isolates that have been published before.

Materials and Methods

Mycobacterium tuberculosis *culture and DNA isolation for whole genome sequencing*

EPTB clinical isolates used in this study were cultured and maintained at the Department of Medical Microbiology, The Postgraduate Institute of Medical Education and Research (PGIMER), Chandigarh, India. The isolates were obtained from extrapulmonary sites of patients diagnosed with TB. The study was approved by PGIMER Institutional Ethics Committee. Five isolates, three from fine-needle aspiration cytology (FNAC from cervical Lymph nodes) and one each from cerebrospinal fluid (CSF) and joint aspirate pus were subjected to whole genome sequencing analysis. The isolates were processed as per the standard protocol and inoculated on Lowenstein–Jensen (LJ) slants (Difco Lowenstein Medium Base No. 244420).

The cultures were subsequently tested using commercially available SD BIO LINE TB Ag MPT64 Rapid kit to rule out Nontuberculous mycobacteria or any other infection. The isolates were subcultured using single colony growth on LJ for DNA isolation. DNA libraries were constructed with genomic DNA extracted using cetyltrimethylammonium bromide (CTAB) method (van Soolingen et al., 1991). Briefly, the bacterial cultures grown on LJ slant were harvested in 400 μL TE buffer and heat killed for 50 min at 95°C. The cell lysis was carried out by adding 30 μL of 20 mg/mL lysozyme (Merck Millipore No. 4403) for 2 h at 37°C. The lysate was further treated with 70 μL of 10% sodium dodecyl sulfate and 10 μL of 10 mg/mL Proteinase K (HiMedia Laboratories No. MB086) followed by 100 μL CTAB-NaCl (5 M) at 65°C for 30 min for protein denaturation and carbohydrate precipitation respectively.

Equal volume of freshly prepared chloroform:isoamyl alcohol solution (24:1 ratio) were added for protein precipitation and centrifuged at 14,000 rpm for 15 min. The upper layer containing nucleic acid was collected and DNA was precipitated using 0.6 volumes of ice-cold isopropanol. The sample was centrifuged (14,000 rpm, for 15 min) and the pellet was washed twice with 70% ice-cold ethanol. DNA pellet was dried and resuspended in DNAase-free water. The quality of DNA was assessed on 1% agarose gel and the DNA was quantified on nanodrop spectrophotometer (ND-1000 UV-Vis). Spectrophotometer readings are provided in Supplementary Data. For library construction, DNA was fragmented and cleaned up using QIAquick columns (QIAGEN). The size distribution was checked by running aliquots of the samples on Agilent Bioanalyzer 7500 Nano chips. Illumina adapters were ligated to each fragment. Fragments of ~300 bases were separated using gel electrophoresis and subjected to paired end sequencing with read length of 100 bps and an average depth of 100X using Illumina MiSeq sequencer.

Mapping and variant detection using Mycobacterium tuberculosis *H37Rv reference genome*

The quality of the raw reads was checked using FastQC version 0.11.5 toolkit followed by trimming of adapters and low-quality bases with a Phred quality score of <20. Paired end reads were then mapped to *Mycobacterium tuberculosis* (MTB) H37Rv reference genome (NC000962.3) using Burrows-Wheeler Alignment Tool (BWA version 0.7.15) (Li and Durbin, 2009). The alignment files were subjected to local realignment and de-duplication using the Genome Analysis Toolkit (GATK version 3.6) and Picard.

Single nucleotide variations (SNVs) and insertion/deletions (INDELs) were called from each alignment file using GATK (McKenna et al., 2010) and Pindel version 0.2.5b8 (Ye et al., 2009). Variant filtering was carried out by removing variants

supported by <5 reads and INDELs supported by <7 reads. The SNVs and INDELs were annotated using in-house perl scripts. All variants identified in this study were manually inspected using Integrative Genomics Viewer (IGV version 2.3.86) (Robinson et al., 2011).

Principal component analysis

The whole genome sequencing data of pulmonary and extrapulmonary isolates were obtained from public domain (Chernyaeva et al., 2014). A total of 76,784 SNVs from 2003 isolates was used to generate the matrix of SNVs and isolates. This matrix was subjected to principal component analysis (PCA). PCA was performed using ggfortify package in R version 3.2.0.

Phylogenetic analysis and spoligotyping

Determination of MTB lineages/sublineages and *in silico* spoligotyping was carried out using KvarQ version 0.12.2 (Steiner et al., 2014). SNVs identified in each isolate and lineage-specific SNVs (Coll et al., 2014) were used to construct phylogenetic tree of the five isolates. The phylogenetic tree was constructed using maximum-likelihood method FastTree version 2.1.10 (Price et al., 2010) using a general time-reversible model with gamma correction for among-site rate variation. Calculation of 1000 bootstrap replicates provided support for nodes on the tree. The phylogenetic tree was visualized with Dendroscope version 3.5.8 (Huson and Scornavacca, 2012).

Prophage prediction in MTB

Prophages in EPTB isolates were identified using PHASTER (PHAge Search Tool Enhanced Release) http://phaster.ca (Arndt et al., 2016) web server. The genomes of five isolates were submitted to PHASTER. The two databases used by PHASTER for prophage prediction are the Bacteria database (updated December 2016) and the Prophage/Virus database (updated December 2016).

Data availability

Whole genome sequencing data were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database with accession PRJNA358480.

Results

Distribution of SNVs in EPTB clinical isolates

MTB genome comprises of 4.4 million base pairs and harbors ~4000 genes (Cole et al., 1998). The genome alignment of five EPTB clinical isolates was carried out using MTB H37Rv reference genome (NC_000962.3). Details of these five EPTB isolates are provided in Table 1. We identified 2341 genetic variations out of which 1434 were nonsynonymous SNVs in EPTB isolates (Supplementary Table S1). Of the total SNVs identified in these isolates, 279 nonsynonymous SNVs have not been previously reported in genome-based *Mycobacterium tuberculosis* variation (GMTV) database. This included SNVs resulting in premature stop codon in 23 genes. In addition, we also identified 143 insertions and 105 deletions (Supplementary Tables S2A and S3A). We observed 15 mutations corresponding to 9 genes that have been previously reported to confer resistance to first and second line antitubercular drugs (Table 3).

Lineage-specific SNVs and phylogenetic analysis

MTB strains are assigned to different lineages using methods such as IS6110-restriction fragment length polymorphism, spoligotyping, MIRU-VNTR, and large sequence polymorphisms. Molecular epidemiological studies have suggested that certain MTB strains are more prone to acquire drug resistance (Drobniewski et al., 2002; Shemyakin et al., 2004; Toungoussova et al., 2004). Related genotypes of MTB tend to be associated with specific geographic locations (Filliol et al., 2003; Friedman et al., 1997). These observations suggest the underlying difference among the clinical isolates that reflects their adaptation to specific host and geographical location.

Whole genome sequencing allows strain typing of MTB with a better resolution than the currently used methods. We performed phylogenetic analysis using SNVs data derived from whole genome sequencing. All five clinical isolates sequenced in this study belonged to lineage 3 (East-African-Indian), which is the most prevalent lineage in the Indian subcontinent. A phylogenetic tree was constructed using whole genome data from the current study and data from 17 EPTB isolates available in public domain (Fig. 1). The isolates were distributed among East-African-Indian, East-Asian (Beijing), Euro American (Ural), and New-1 Uganda, X-type lineage.

Whole genome sequencing data were also used to generate spoligotyping octal codes (Table 2). The octal codes also confirmed the isolates belong to East-African-Indian lineage. In the public domain, we found data for 45 pulmonary MTB isolates from lineage 3 (Ali et al., 2015; Chernyaeva et al., 2014). We compared the variants from these pulmonary isolates with those we observed in our EPTB isolates (lineage 3). Of the 1434 SNVs identified in EPTB isolates, 55% (793/1434) of the variations were found to be common between the two groups.

Table 1. Clinical Information for the Isolates Used in the Study

Sample ID	*Type*	*Age, years*	*Gender*	*Geographical location*	*Year of isolation*
PGI-14	CSF	58	Male	Chandigarh	2012
PGI-98	Joint aspirate pus	47	Male	Chandigarh	2013
PGI-100	FNAC (cervical lymph node)	38	Male	Punjab	2014
PGI-103	FNAC (cervical lymph node)	53	Male	Punjab	2013
PGI-155	FNAC (cervical lymph node)	28	Female	Punjab	2014

FNAC, fine-needle aspiration cytology; CSF, cerebrospinal fluid.

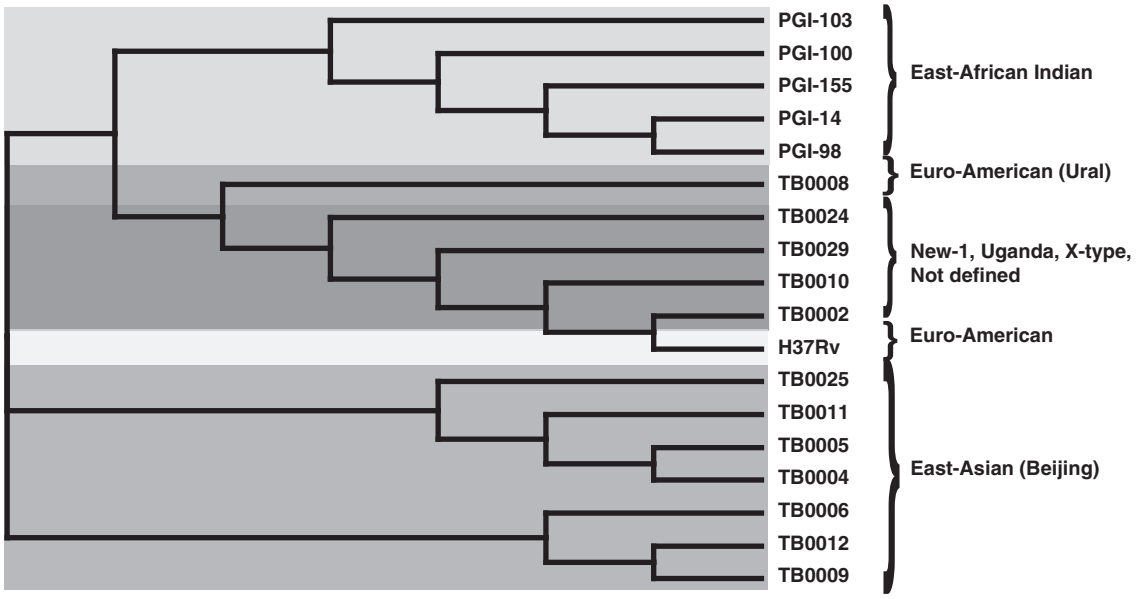

FIG. 1. Phylogenetic classification of extrapulmonary clinical isolates.

Nonsynonymous SNVs in extrapulmonary clinical isolates

To investigate genetic variation between pulmonary and extrapulmonary isolates, we combined our data ($n=5$) with whole genome data from 17 publically available datasets from EPTB isolates. It was compared with data from pulmonary TB isolates ($n=1981$) available in GMTV database (Chernyaeva et al., 2014). In total, 3487 SNVs were found to be common between pulmonary and extrapulmonary isolates whereas 2185 mutations were exclusive to EPTB isolates. Of these, 279 SNVs were exclusive to our dataset (Supplementary Table S4). SNVs associated with pulmonary and extrapulmonary isolates were subjected to PCA. The PCA revealed distinct clustering of SNVs associated with pulmonary TB and EPTB isolates (Fig. 2).

A significant number of variations were identified in the hyper-variable Pro-Glu (PE)-polymorphic GC-rich sequences (PGRS) regions in the genome. MTB harbors a large number of genes that code for motifs PE or Pro-Pro-Glu. A subgroup of this family contains PGRS and the other subgroup major polymorphic tandem repeats. These genes have been reported to play an important role in evasion of host immune response via antigenic variation (Sampson, 2011).

The cell wall of mycobacteria is highly complex and accounts for several intrinsic properties such as virulence, pathogenesis, formation of cords, and acid fastness (Akhtar et al., 2006; Goren et al., 1978; Kansal et al., 1998). It has been previously shown that there is an association between the pathogenic potential and cell wall structure of atypical mycobacteria. Cell wall is also associated with transport of various molecules across the bacterial cell. We observed mutation in transmembrane protein-carbonic anhydrase (Rv3273-R125P) in four out of five EPTB isolates sequenced in this study. Carbonic anhydrase is involved in transport of sulfate across the membrane (Gu et al., 2003).

Other membrane associated proteins with mutations included-Rv0361 (membrane protein), Rv0936 [phosphate-transport ATP-binding cassette (ABC) transporter integral membrane protein PstA2], and Rv1236 (sugar-transport ABC transporter integral membrane protein SugA). We observed mutations in Rv3303c gene (lpdA-C*472) leading to premature stop codon in all the five isolates. This mutation has been reported earlier in MTB clinical isolates. Comparison of INDELs identified in the current analysis also revealed differences in pulmonary TB and EPTB isolates. Of the 143 insertions identified, only 7 insertions (3 in coding DNA sequence [CDS] and 4 in intergenic regions) were reported in pulmonary TB clinical isolates (Supplementary Table S2B). Similarly, 7 of 105 deletions in CDS were reported in pulmonary TB isolates (Supplementary Table S3B).

Table 2. *In Silico* Spoligotyping Octal Codes

Sample ID	*Spoligotype octal code*	*Lineage*
PGI-14	703775740003771	East-African-Indian
PGI-98	703775740003771	East-African-Indian
PGI-100	703775740003771	East-African-Indian
PGI-103	703777740003171	East-African-Indian
PGI-155	773775777403771	East-African-Indian

Mutations in genes known to confer resistance to first and second-line antitubercular drugs

Mutations in certain genes are known to confer resistance to antitubercular drugs. These drug-resistant bacteria require extended TB treatment from 6 to 24 months. They are further classified as MDR-TB and extensively drug-resistant TB (XDR-TB). In MDR-TB, the mycobacteria are resistant to at least two major first line drugs—isoniazid (INH) and rifampicin. In addition to first-line drugs, resistance to any fluoroquinolone and at least one of the three injectable second-line drugs (i.e., amikacin, kanamycin, or capreomycin) (Iseman, 1993), is known as XDR-TB. Recent studies on MTB gene mutations have shown strong association of a few mutations with drug resistance (Cade et al., 2010). For example, mutation in katG leads to INH resistance, which is an effective first-line antitubercular drug.

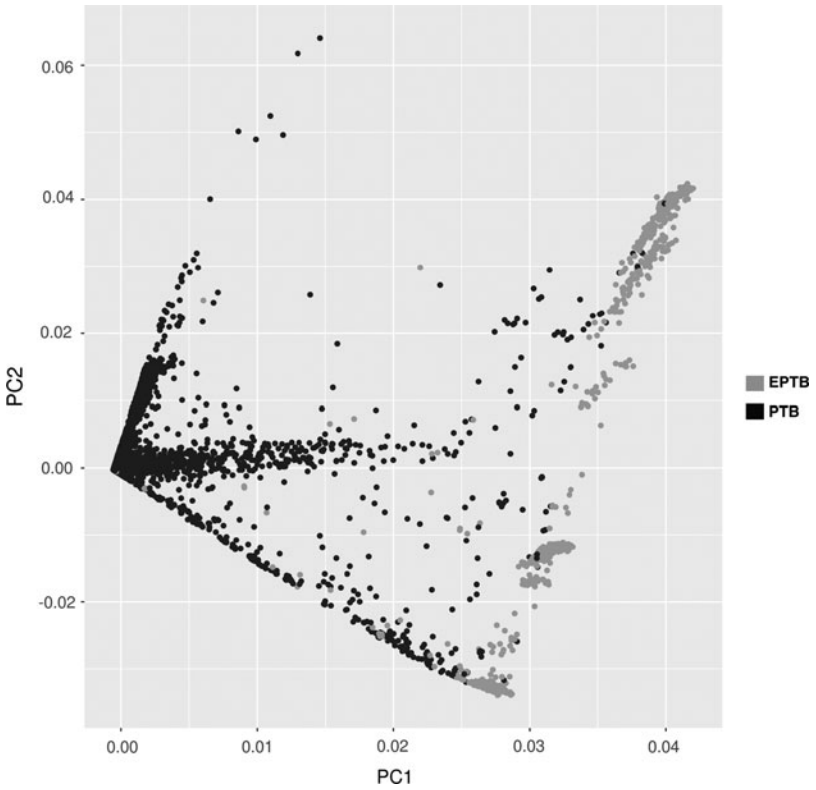

FIG. 2. Principal component analysis of nonsynonymous SNVs identified in EPTB isolates from current analysis and pulmonary TB isolates from genome data available in public domain. *Blue dots* represent SNVs in EPTB and *black dots* represent SNVs in pulmonary TB. EPTB, extrapulmonary tuberculosis; PTB, pulmonary TB; SNVs, single nucleotide variations.

KatG, a catalase-peroxidase, converts INH into an active form in MTB. Commercially available genotyping methods such as Xpert MTB/rifampicin (MTB/RIF) and Line Probe assays are used for detection of drug resistance-associated mutations (Massi et al., 2017; Nathavitharana et al., 2017; Saeed et al., 2017). In total, we identified 15 mutations that are associated with drug resistance (Fig. 3A). Three mutations (rpoB D435Y, rpoB L430P, and rpoB L452P) that are used in GeneXpert to diagnose MDR-TB were identified in 80% of the isolates. These isolates also harbored mutations in katG (R463L and S315T), conferring resistance to INH (Fig. 3B).

In addition, we also identified mutations known to confer resistance to ethambutol, fluoroquinolones, and streptomycin (Table 3). It is suggested that presence of drug resistance mutations costs fitness of bacteria in absence of drug (de Vos et al., 2013). However, compensatory mutations in rpoA and rpoC gene have shown to restore the fitness of rifampicin-resistant bacteria (Brandis and Hughes, 2013; Brandis et al., 2012). We observed compensatory mutation-rpoC H525 in one of the EPTB isolates in our data along with rpoB S450L mutation. Mutation in gid gene, which is known to confer resistance to streptomycin, resulted in premature stop codon (gid Q125*). Interestingly, this mutation was identified in four out of five isolates analyzed in this study.

Functional classification of EPTB specific mutations

Genes carrying nonsynonymous mutations that were exclusive to EPTB isolates were subjected to gene ontology (GO) analysis. GO enrichment analysis for biological process, molecular function, and subcellular localization was carried out using protein analysis through evolutionary relationship (PANTHER) classification system. GO terms with $p \leq 0.05$ were considered as significant (Mi et al., 2017). Categorization of proteins based on biological process showed enrichment of metabolic processes. We observed enrichment of catalytic activity, ion binding, transferase activity, and adenosine triphosphate binding.

Subcellular localization of proteins identified 10 major categories including cell wall, external encapsulating structure, cell membrane, cytoplasm, cell periphery, and plasma membrane (Fig. 4). MTB cell wall represents a wide range of complex lipids and lipoglycans on its cell surface (Converse et al., 2003). These cell wall lipids are known to play an important role in pathogenesis of MTB (Bhuwan et al., 2016). Genes associated with their biosynthesis and transport are important in determining the interaction of pathogen with host.

Prophage distribution in EPTB isolates

Phage-mediated horizontal gene transfer has earlier been reported to influence bacterial virulence and its response to antibiotics in *Escherichia coli* (Fortier and Sekulovic, 2013). Choo et al., have earlier reported rapid phage-mediated evolution in *Mycobacterium abscessus* suggesting the role of phage integration in controlling important physiological functions in bacteria.

In this study, we used PHASTER-a phage identification tool to find phage DNA incorporated in MTB genome (Arndt et al., 2016). Whole genome sequencing data from EPTB

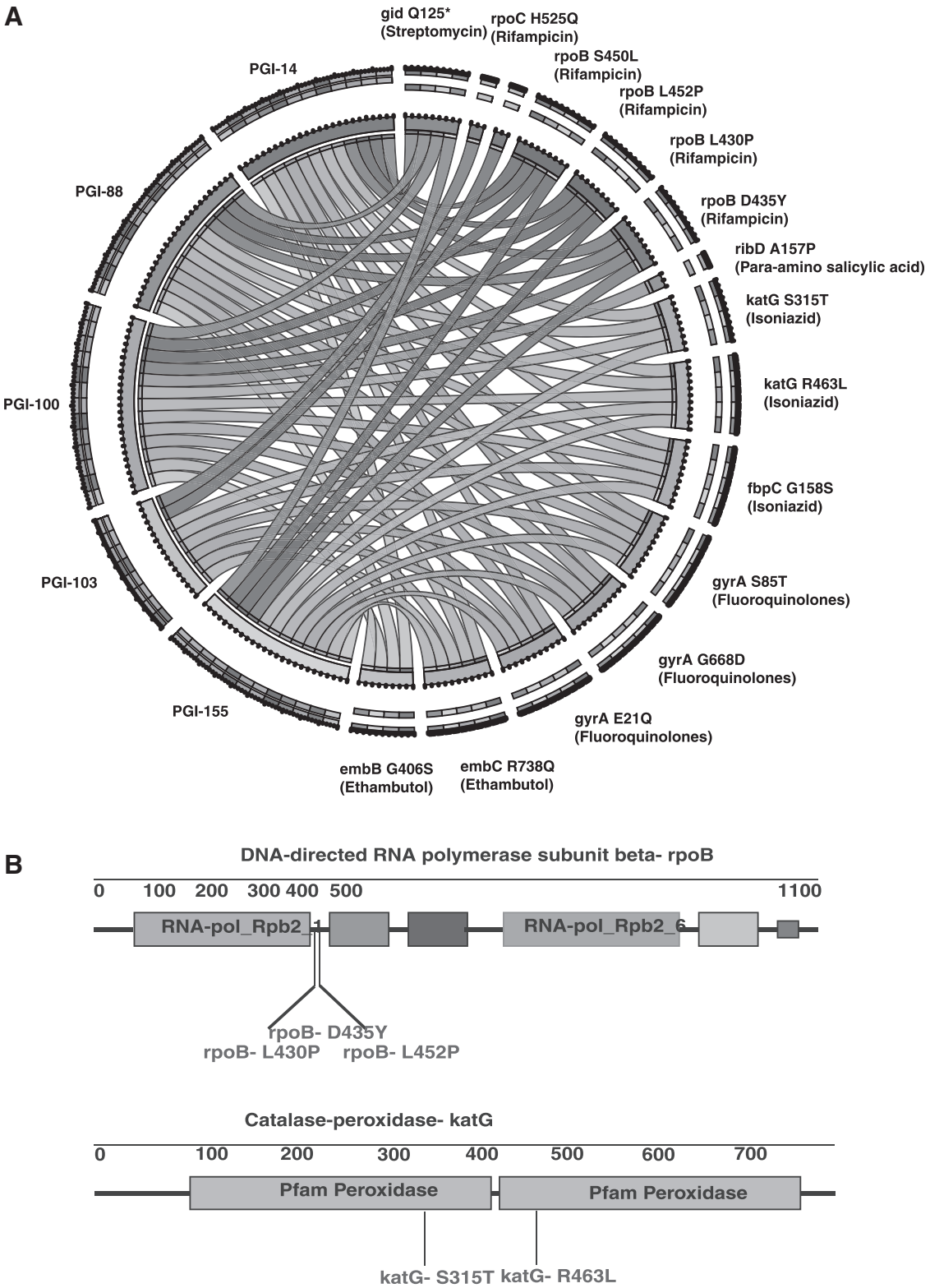

FIG. 3. **(A)** Circos plot depicting mutation frequency in genes known to confer drug resistance. **(B)** Mutations identified in catalytic domains of rpoB and katG conferring resistance to rifampicin and isoniazid respectively.

clinical isolates were used for prophage identification. In total, we identified nine different putative prophages in EPTB isolates (Supplementary Fig. S1). In addition, we also identified attachment sites, *att*L and *att*R along with integrase enzyme in all the five isolates. Of the nine prophages identified in our analysis, Mycobacterium phage Kratio DNA was present in all the MTB genomes whereas three prophages—Mycobacterium phage Larva, Mycobacterium phage Peg-Leg, and Mycobacterium phage GUmbie were present in 80% of the isolates (Fig. 5).

Table 3. Mutations Known to Confer Resistance to First and Second-Line Antitubercular Drugs

Sample_ID			*PGI-14*	*PGI-98*	*PGI-100*	*PGI-103*	*PGI-155*
Category			*CSF*	*FNAC*	*FNAC*	*FNAC*	*Joint aspirate pus*
Gene	*Mutation*	*Drug*					
embB	embB G406S	Ethambutol	√	√	√	X	√
embC	embC R738Q	Ethambutol	√	√	√	√	√
gyrA	gyrA E21Q	Fluoroquinolones	√	√	√	√	√
gyrA	gyrA G668D	Fluoroquinolones	√	√	√	√	√
gyrA	gyrA S95T	Fluoroquinolones	√	√	√	√	√
fbpC	fbpC G158S	Isoniazid	√	√	√	√	√
katG	katG R463L	Isoniazid	√	√	√	√	√
katG	katG S315T	Isoniazid	√	√	√	X	√
ribD	ribD A157P	Para-aminosalisylic acid	X	X	√	X	X
rpoB	rpoB D435Y	Rifampicin	√	√	√	X	√
rpoB	rpoB L430P	Rifampicin	√	√	√	X	√
rpoB	rpoB L452P	Rifampicin	√	√	√	X	√
rpoB	rpoB S450L	Rifampicin	X	X	X	√	X
rpoC	rpoC H525Q	Rifampicin	X	X	X	√	X
gid	gid Q125*	Streptomycin	√	√	√	X	√

CRISPR identification in extrapulmonary clinical isolates

Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) have recently been identified as an inheritable part of bacterial genome (Peng et al., 2014). CRISPR is a component of bacterial immunity. An estimated 40% of bacteria and 90% of archaea carry CRISPR in their genome. A CRISPR in general is composed of arrays that are repetitive elements interspersed with spacers. Spacers originate from short fragments of foreign DNA (Peng et al., 2014). We used CRISPRFinder tool (Grissa et al., 2007) to identify putative CRISPR loci in MTB genomes of clinical isolates. We identified 8 known CRISPR loci in EPTB isolates (Supplementary Table S5). Two CRISPRs were shared between PGI-14 and PGI-100 isolates and remaining CRISPRs were unique to individual isolates. In addition, 14 possible CRISPRs not reported earlier in MTB were also predicted of which CRISPR-1 was shared between all the 5 five isolates whereas CRISPR 2 and 3 were shared among 3 and 2 MTB isolates respectively (Fig. 6).

Mutations in noncoding RNA

Previous efforts on sequencing-based analysis of mycobacterial genomes have revealed genes for noncoding RNAs that include intergenic small RNAs (sRNAs) (Arnvig et al., 2011; Miotto et al., 2012). These sRNAs can regulate gene expression in response to environmental changes (Song and Wai, 2009). We identified two mutations in noncoding RNAs (mcr11 G26D, mcr3 S14L) in all the EPTB isolates and one mutation in (F6-T9P) noncoding RNA in isolate PGI-98. We also found two novel mutations in 23S ribosomal RNA-rrl (V123A and K429R) that have not been previously reported (Table 4). These mutations in *rrl* gene are associated with Linezolid (second line drug) resistance.

Discussion

There are two clinical manifestations of TB: pulmonary TB and EPTB. Several clinical isolates from pulmonary TB have been sequenced over the years and have provided significant insights into genetic diversity and drug resistance pattern. However, there is lack of data related to EPTB isolates. The genetic diversity and drug resistance pattern among extrapulmonary isolates remains largely unexplored. Several reports have earlier provided the evidence that *Mycobacterium tuberculosis* complex strains vary in terms of virulence and pathogenicity in model organisms (Coscolla and Gagneux, 2010). It is proposed, MTB strains causing EPTB have a better ability to invade macrophages, have a higher replication rate in macrophages and animal models, can disseminate, and are associated with decreased survival among animals, compared with strains that cause non-EPTB (Garcia de Viedma et al., 2005; Wong et al., 2007).

Genetic studies on humans and MTB indicate coevolution and adaptation of mycobacteria to changing population (Banuls et al., 2015; David et al., 2015). Several studies have shown association of genetic variation with disease presentation, treatment outcome, and frequency of transmission (Caws et al., 2008; Faksri et al., 2011). Currently, the evidence to attribute specific genetic variations to disease presentation in pulmonary/extrapulmonary sites is inconclusive. Given the complexity of genomic diversity within the MTB strains and clinical phenotypes, systematic analysis of isolates from various infection sites are needed. In this study, whole genome sequencing analysis of MTB isolated from CSF, FNAC from cervical lymph nodes and joint aspirate pus was performed.

In addition to known mutations, several novel SNVs were identified. Comparison of variants from extrapulmonary and pulmonary isolates revealed several genetic variations that were only observed in EPTB. The significance of these genomic variants in the pathogenesis of EPTB is unclear. We also identified several mutations that are known to confer drug resistance. Whole genome sequencing in conjunction with comprehensive drug susceptibility testing can reveal novel mutations that confer drug resistance. The data would be valuable to develop next generation diagnostics to identify and monitor drug resistance.

Based on limited dataset, it is not possible to determine whether specific genetic variations are associated with EPTB

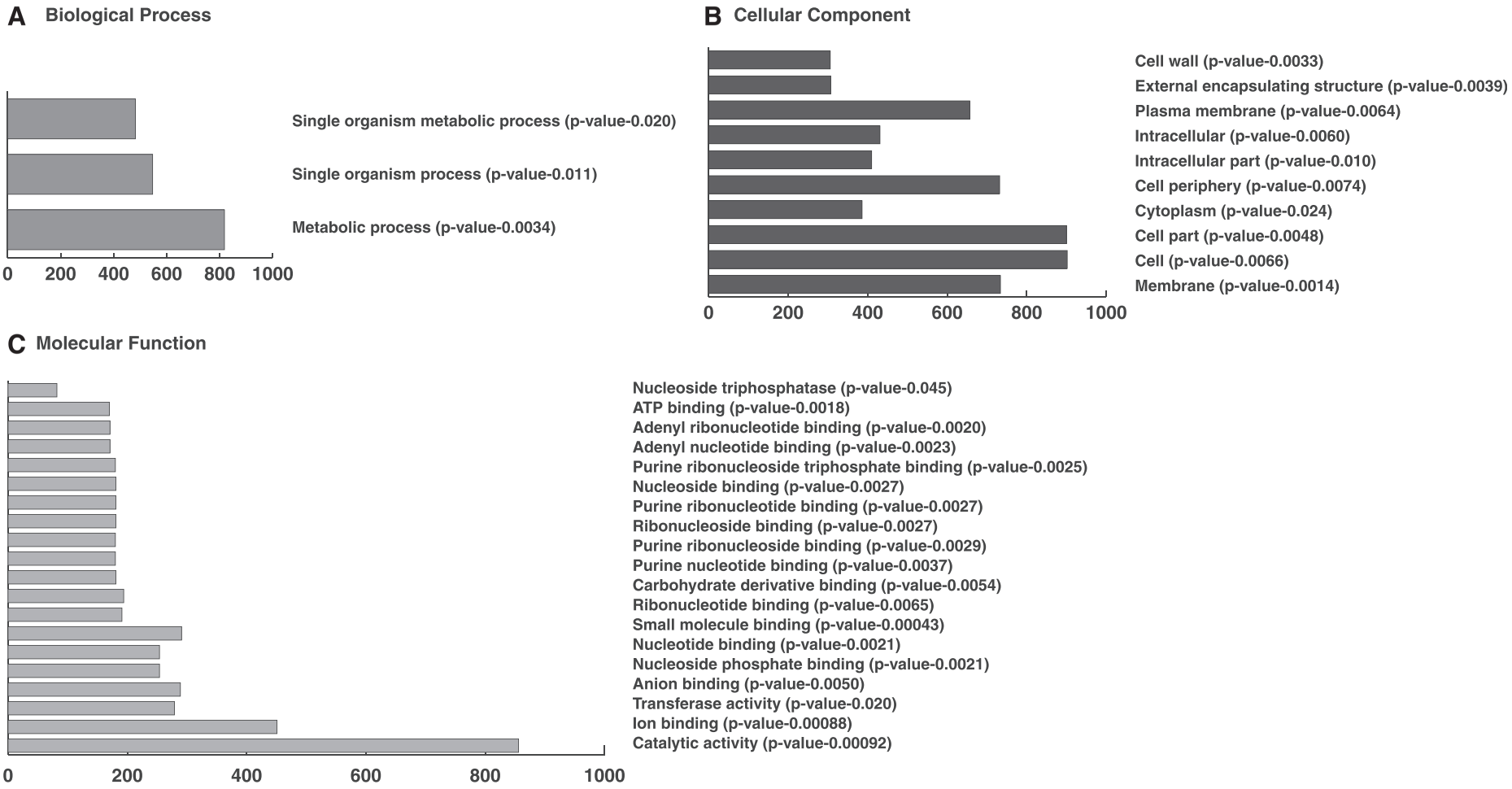

FIG. 4. Gene ontology analyses of genes with SNVs exclusive to extrapulmonary isolates compared to pulmonary isolates.

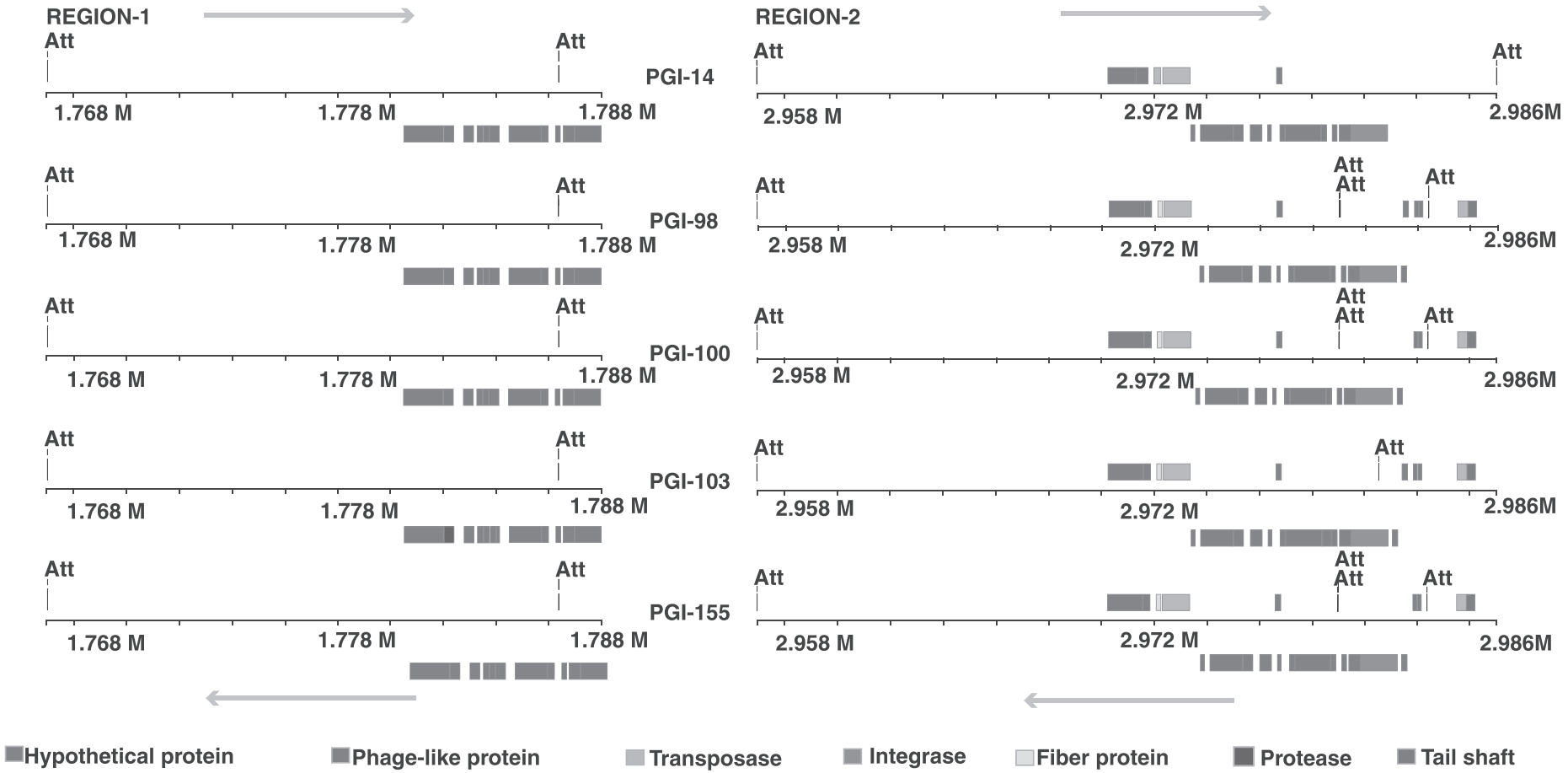

FIG. 5. The structure of the putative prophages found in the EPTB isolates. These prophages consist of phage-associated phage-like proteins and enzymes-like integrase, which assist in the integration of phage into bacteria.

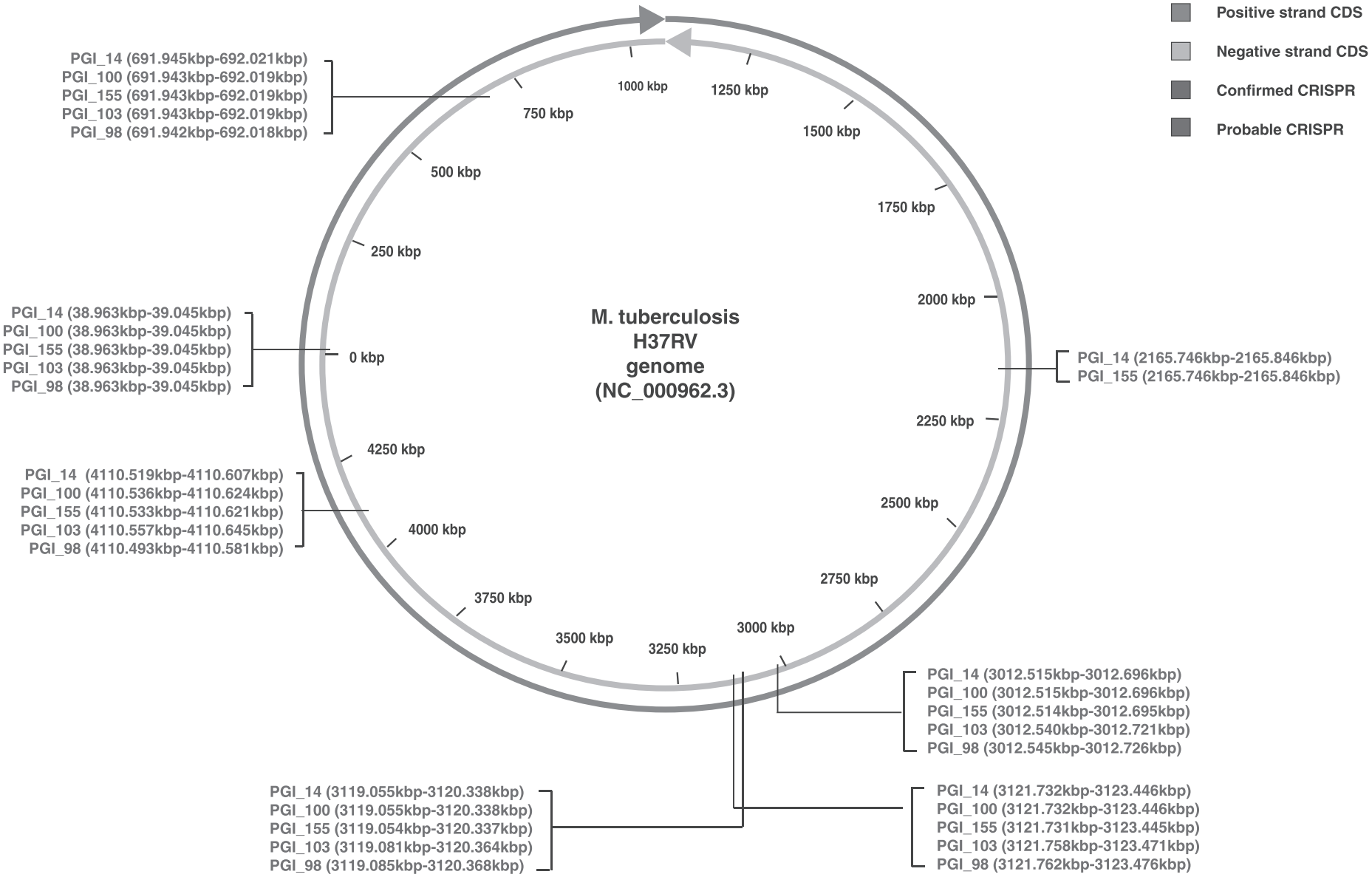

FIG. 6. Known and predicted clustered regularly interspaced short palindromic repeats (CRISPR) identified in EPTB isolates.

Table 4. List of Mutations Present in Noncoding RNA

Sample ID		*PGI-14*	*PGI-100*	*PGI-103*	*PGI-98*	*PGI-155*
Category		*CSF*	*FNAC*	*FNAC*	*FNAC*	*Joint aspirate pus*
Noncoding RNA	*Mutation*					
mcr11	mcr11 G26D	√	√	√	√	√
mcr3	mcr3 S14L	√	√	√	√	√
rrl	rrl V123A	X	X	√	X	X
Rrl	rrl K429R	X	X	√	X	X
F6	F6 T9P	X	X	X	√	X

isolates as compared to pulmonary isolates. This warrants large-scale whole genome sequencing efforts similar to those carried out on pulmonary isolates. In addition, identification of several novel SNVs in our isolates also highlights underrepresentation of genetic data from MTB strains from Indian subcontinent, in public databases. Future efforts to sequence large number of isolates from the Indian subcontinent will expand our understanding of genetic diversity in MTB and also reveal novel genetic determinants of drug resistance.

Conclusions

Several studies have been carried out to map genetic diversity of MTB and investigate its implications on pathogenesis, transmissibility, drug resistance, and susceptibility. Advent of next generation sequencing technologies has enabled whole genome sequencing of hundreds of MTB. Most of these studies have been carried out on clinical isolates from pulmonary TB. There are limited studies on whole genome sequencing of clinical isolates from extrapulmonary sites. We carried out whole genome sequencing of five isolates from extrapulmonary sites. These isolates carried 279 SNVs that are not reported before. In addition, large-scale efforts are needed to carry out whole genome sequencing of EPTB isolates to systematically map genetic diversity and understand genotype–phenotype correlations.

Acknowledgments

T.S.K.P. is the recipient of the Department of Science and Technology (DST)-IDP research grant ''Development of epitope based diagnostic gadget for detection of MTB in the Indian population'' from the DST, Government of India. K.S. is thankful to HRTP-NIH (HIV Research Training Program) for financial support in the form of 3 months fellowship for learning the analysis of whole genome sequencing (WGS) at Johns Hopkins University. K.S. acknowledges the financial support to New York University's School of Medicine, New York University under HRTP-NIH for 3 months fellowship.

We thank the Infosys Foundation for research support to the Institute of Bioinformatics. R.V. is a recipient of Senior Research Fellowship from University Grants Commission (UGC), Government of India. J.A. is a recipient of Senior Research Fellowship from Council of Scientific and Industrial Research (CSIR), Government of India. O.C. is a recipient of INSPIRE Fellowship from the DST, Government of India.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Akhtar P, Srivastava S, Srivastava A, Srivastava M, Srivastava BS, and Srivastava R. (2006). Rv3303c of *Mycobacterium tuberculosis* protects tubercle bacilli against oxidative stress *in vivo* and contributes to virulence in mice. Microbes Infect 8, 2855–2862.

Ali A, Hasan Z, McNerney R, et al. (2015). Whole genome sequencing based characterization of extensively drug-resistant *Mycobacterium tuberculosis* isolates from Pakistan. PLoS One 10, e0117771.

Ansari A, Talat N, Jamil B, et al. (2009). Cytokine gene polymorphisms across tuberculosis clinical spectrum in Pakistani patients. PLoS One 4, e4778.

Arndt D, Grant JR, Marcu A, et al. (2016). PHASTER: A better, faster version of the PHAST phage search tool. Nucleic Acids Res 44, W16–W21.

Arnvig KB, Comas I, Thomson NR, et al. (2011). Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. PLoS Pathog 7, e1002342.

Banuls AL, Sanou A, Anh NT, and Godreuil S. (2015). *Mycobacterium tuberculosis*: Ecology and evolution of a human bacterium. J Med Microbiol 64, 1261–1269.

Bhuwan M, Arora N, Sharma A, et al. (2016). Interaction of *Mycobacterium tuberculosis* virulence factor RipA with chaperone MoxR1 is required for transport through the TAT secretion system. MBio 7, e02259.

Brandis G, and Hughes D. (2013). Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. J Antimicrob Chemother 68, 2493–2497.

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

Cade CE, Dlouhy AC, Medzihradszky KF, Salas-Castillo SP, and Ghiladi RA. (2010). Isoniazid-resistance conferring mutations in *Mycobacterium tuberculosis* KatG: Catalase, peroxidase, and INH-NADH adduct formation activities. Protein Sci 19, 458–474.

Casali N, Nikolayevskyy V, Balabanova Y, et al. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet 46, 279–286.

Caws M, Thwaites G, Dunstan S, et al. (2008). The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog 4, e1000034.

Chernyaeva EN, Shulgina MV, Rotkevich MS, et al. (2014). Genome-wide *Mycobacterium tuberculosis* variation (GMTV) database: A new tool for integrating sequence variations and epidemiology. BMC Genomics 15, 308.

Choo SW, Wee WY, Ngeow YF, et al. (2014). Genomic reconnaissance of clinical isolates of emerging human pathogen

Mycobacterium abscessus reveals high evolutionary potential. Sci Rep 4, 4061.

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

Coll F, McNerney R, Guerra-Assuncao JA, et al. (2014). A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun 5, 4812.

Converse SE, Mougous JD, Leavell MD, Leary JA, Bertozzi CR, and Cox JS. (2003). MmpL8 is required for sulfolipid-1 biosynthesis and *Mycobacterium tuberculosis* virulence. Proc Natl Acad Sci U S A 100, 6121–6126.

Coscolla M, and Gagneux S. (2010). Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov Today Dis Mech 7, e43–e59.

Coscolla M, and Gagneux S. (2014). Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin Immunol 26, 431–444.

Das S, Roychowdhury T, Kumar P, et al. (2013). Genetic heterogeneity revealed by sequence analysis of *Mycobacterium tuberculosis* isolates from extra-pulmonary tuberculosis patients. BMC Genomics 14, 404.

David S, Mateus AR, Duarte EL, et al. (2015). Determinants of the sympatric host-pathogen relationship in tuberculosis. PLoS One 10, e0140625.

de Vos M, Muller B, Borrell S, et al. (2013). Putative compensatory mutations in the rpoC gene of rifampin-resistant *Mycobacterium tuberculosis* are associated with ongoing transmission. Antimicrob Agents Chemother 57, 827–832.

Drobniewski F, Balabanova Y, Ruddy M, et al. (2002). Rifampin- and multidrug-resistant tuberculosis in Russian civilians and prison inmates: Dominance of the Beijing strain family. Emerg Infect Dis 8, 1320–1326.

Faksri K, Drobniewski F, Nikolayevskyy V, et al. (2011). Epidemiological trends and clinical comparisons of *Mycobacterium tuberculosis* lineages in Thai TB meningitis. Tuberculosis (Edinb) 91, 594–600.

Fanning A. (1999). Tuberculosis: 6. Extrapulmonary disease. CMAJ 160, 1597–1603.

Fernando SL, Saunders BM, Sluyter R, et al. (2007). A polymorphism in the P2X7 gene increases susceptibility to extrapulmonary tuberculosis. Am J Respir Crit Care Med 175, 360–366.

Filliol I, Driscoll JR, van Soolingen D, et al. (2003). Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J Clin Microbiol 41, 1963–1970.

Fortier LC, and Sekulovic O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. Virulence 4, 354–365.

Friedman CR, Quinn GC, Kreiswirth BN, et al. (1997). Widespread dissemination of a drug-susceptible strain of *Mycobacterium tuberculosis*. J Infect Dis 176, 478–484.

Garcia de Viedma D, Lorenzo G, Cardona PJ, et al. (2005). Association between the infectivity of *Mycobacterium tuberculosis* strains and their efficiency for extrarespiratory infection. J Infect Dis 192, 2059–2065.

Glickman MS, Cox JS, and Jacobs WR Jr. (2000). A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. Mol Cell 5, 717–727.

Golden MP, and Vikram HR. (2005). Extrapulmonary tuberculosis: An overview. Am Fam Physician 72, 1761–1768.

Goren MB, Cernich M, and Brokl O. (1978). Some observations of mycobacterial acid-fastness. Am Rev Respir Dis 118, 151–154.

Grissa I, Vergnaud G, and Pourcel C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35, W52–W57.

Gu S, Chen J, Dobos KM, Bradbury EM, Belisle JT, and Chen X. (2003). Comprehensive proteomic profiling of the membrane constituents of a *Mycobacterium tuberculosis* strain. Mol Cell Proteomics 2, 1284–1296.

Hasan Z, Zaidi I, Jamil B, Khan MA, Kanji A, and Hussain R. (2005). Elevated *ex vivo* monocyte chemotactic protein-1 (CCL2) in pulmonary as compared with extra-pulmonary tuberculosis. BMC Immunol 6, 14.

Huson DH, and Scornavacca C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol 61, 1061–1067.

Iseman MD. (1993). Treatment of multidrug-resistant tuberculosis. N Engl J Med 329, 784–791.

Kansal RG, Gomez-Flores R, and Mehta RT. (1998). Change in colony morphology influences the virulence as well as the biochemical properties of the *Mycobacterium avium* complex. Microb Pathog 25, 203–214.

Kolattukudy PE, Fernandes ND, Azad AK, Fitzmaurice AM, and Sirakova TD. (1997). Biochemistry and molecular genetics of cell-wall lipid biosynthesis in mycobacteria. Mol Microbiol 24, 263–270.

Li H, and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Martinez A, Torello S, and Kolter R. (1999). Sliding motility in mycobacteria. J Bacteriol 181, 7331–7338.

Massi MN, Biatko KT, Handayani I, et al. (2017). Evaluation of rapid GeneXpert MTB/RIF method using DNA tissue specimens of vertebral bones in patients with suspected spondylitis TB. J Orthop 14, 189–191.

McKenna A, Hanna M, Banks E, et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303.

Meghdadi H, Khosravi AD, Ghadiri AA, Sina AH, and Alami A. (2015). Detection of *Mycobacterium tuberculosis* in extrapulmonary biopsy samples using PCR targeting IS6110, rpoB, and nested-rpoB PCR cloning. Front Microbiol 6, 675.

Mi H, Huang X, Muruganujan A, et al. (2017). PANTHER version 11: Expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45, D183–D189.

Miotto P, Forti F, Ambrosi A, et al. (2012). Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. PLoS One 7, e51950.

Nathavitharana RR, Cudahy PG, Schumacher SG, Steingart KR, Pai M, and Denkinger CM. (2017). Accuracy of line probe assays for the diagnosis of pulmonary and multidrug-resistant tuberculosis: A systematic review and meta-analysis. Eur Respir J 49.

Oki NO, Motsinger-Reif AA, Antas PR, Levy S, Holland SM, and Sterling TR. (2011). Novel human genetic variants associated with extrapulmonary tuberculosis: A pilot genome wide association study. BMC Res Notes 4, 28.

Peng L, Pei J, Pang H, Guo Y, Lin L, and Huang R. (2014). Whole genome sequencing reveals a novel CRISPR system in industrial *Clostridium acetobutylicum*. J Ind Microbiol Biotechnol 41, 1677–1685.

Price MN, Dehal PS, and Arkin AP. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.

Robinson JT, Thorvaldsdottir H, Winckler W, et al. (2011). Integrative genomics viewer. Nat Biotechnol 29, 24–26.

Saeed M, Iram S, Hussain S, Ahmed A, Akbar M, and Aslam M. (2017). GeneXpert: A new tool for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis*. J Pak Med Assoc 67, 270–274.

Sampson SL. (2011). Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol 2011, 497203.

Shemyakin IG, Stepanshina VN, Ivanov IY, et al. (2004). Characterization of drug-resistant isolates of *Mycobacterium tuberculosis* derived from Russian inmates. Int J Tuberc Lung Dis 8, 1194–1203.

Song T, and Wai SN. (2009). A novel sRNA that modulates virulence and environmental fitness of Vibrio cholerae. RNA Biol 6, 254–258.

Srilohasin P, Chaiprasert A, Tokunaga K, et al. (2014). Genetic diversity and dynamic distribution of *Mycobacterium tuberculosis* isolates causing pulmonary and extrapulmonary tuberculosis in Thailand. J Clin Microbiol 52, 4267–4274.

Steiner A, Stucki D, Coscolla M, Borrell S, and Gagneux S. (2014). KvarQ: Targeted and direct variant calling from FastQ reads of bacterial genomes. BMC Genomics 15, 881.

Toungoussova OS, Caugant DA, Sandven P, Mariandyshev AO, and Bjune G. (2004). Impact of drug resistance on fitness of *Mycobacterium tuberculosis* strains of the W-Beijing genotype. FEMS Immunol Med Microbiol 42, 281–290.

Tso HW, Ip WK, Chong WP, Tam CM, Chiang AK, and Lau YL. (2005). Association of interferon gamma and interleukin 10 genes with tuberculosis in Hong Kong Chinese. Genes Immun 6, 358–363.

van Soolingen D, Hermans PW, de Haas PE, Soll DR, and van Embden JD. (1991). Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: Evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J Clin Microbiol 29, 2578–2586.

van Soolingen D, Jajou R, Mulder A, and de Neeling H. (2016). Whole genome sequencing as the ultimate tool to diagnose tuberculosis. Int J Mycobacteriol 5(Suppl 1), S60–S61.

Wilkinson RJ, Llewelyn M, Toossi Z, et al. (2000). Influence of vitamin D deficiency and vitamin D receptor polymorphisms on tuberculosis among Gujarati Asians in west London: A case-control study. Lancet 355, 618–621.

Wong KC, Leong WM, Law HK, et al. (2007). Molecular characterization of clinical isolates of *Mycobacterium tuberculosis* and their association with phenotypic virulence in human macrophages. Clin Vaccine Immunol 14, 1279–1284.

Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871.

Zhang H, Li D, Zhao L, et al. (2013). Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet 45, 1255–1260.

Address correspondence to:
Harsha Gowda, PhD
Faculty Scientist
Institute of Bioinformatics
7th Floor, Discoverer Building
International Tech Park
Bangalore 560066
India

E-mail: harsha@ibioinformatics.org

Thottethodi Subrahmanya Keshava Prasad, PhD
Faculty Scientist
Institute of Bioinformatics
7th Floor, Discoverer Building
International Technology Park
Bangalore 560066
India

E-mail: keshav@ibioinformatics.org

Kusum Sharma, MD
Department of Medical Microbiology
PGIMER
Sector 12
Chandigarh 160012
India

E-mail: sharmakusum9@yahoo.co.in

Abbreviations Used

ABC = ATP-binding cassette
CDS = coding DNA sequence
CRISPR = clustered regularly interspaced short palindromic repeats
CSF = cerebrospinal fluid
CTAB = cetyltrimethylammonium bromide
EPTB = extrapulmonary tuberculosis
FNAC = fine-needle aspiration cytology
GATK = Genome Analysis Toolkit
GMTV = *Mycobacterium tuberculosis* variation
GO = gene ontology
INDELs = insertions and deletions
INH = isoniazid
LJ = Lowenstein–Jensen
MDR-TB = multidrug-resistant TB
MIRU-VNTR = mycobacterial interspersed repetitive-unit–variable-number tandem-repeat
MTB = *Mycobacterium tuberculosis*
NCBI = National Center for Biotechnology Information
PANTHER = protein analysis through evolutionary relationships
PCA = principal component analysis
PE = Pro-Glu
PGIMER = Postgraduate Institute of Medical Education and Research
PGRS = polymorphic GC-rich sequences
PHASTER = PHAge Search Tool Enhanced Release
RIF = rifampicin
SRA = Sequence Read Archive
SNVs = single nucleotide variations
sRNAs = small RNAs
TB = tuberculosis
WHO = World Health Organization
XDR-TB = extensively drug-resistant TB