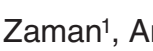
Comparative Genome Analysis of 2 *Mycobacterium Tuberculosis* Strains from Pakistan: Insights Globally Into Drug Resistance, Virulence, and Niche Adaptation

Evolutionary Bioinformatics
Volume 14: 1–9

DOI: 10.1177/1176934318790252



Asma Muhammad Yar[1], Ghanva Zaman[1], Annam Hussain[1], Yan Changhui[2], Azhar Rasul[3], Abrar Hussain[1], Zhu Bo[4], Habib Bokhari[5] and Muhammad Ibrahim[1]

[1]Genomics and Computational Biology Laboratory, COMSATS University Islamabad, Sahiwal Campus, Pakistan. [2]Department of Computer Science, North Dakota State University, Fargo, ND, USA. [3]Department of Zoology, Government College University, Faisalabad, Pakistan. [4]Key Laboratory of Urban Agriculture by Ministry of Agriculture of China, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China. [5]Laboratories of Microbiology and Public Health, COMSATS University Islamabad, Islamabad, Pakistan.

ABSTRACT: Multidrug-resistant *Mycobacterium tuberculosis* is a global threat particularly in developing countries like Pakistan. In this study, we identified 2 *M tuberculosis* strains, mnpk and swlpk, by 16S RNA genes, sequenced their draft genome, and compared the 2 genomes with reference strain H37Rv and gene expression analysis of selected virulent genes. Phylogenetic analysis of *M tuberculosis* strains, mnpk and swlpk, using 16S RNA genes revealed that the strains are closely related with reference strain H37Rv. The draft genome sequence of mnpk and swlpk contains 4305 and 4295 protein-coding genes, respectively, having 99.9% with high collinearity when compared with H37Rv. Although some important drug-resistant genes such as *fabG*, *faDE24*, and *iniA* were missing, genome mining also revealed key drug-resistant genes such as *katG*, *inhA*, *rpoA*, *rpoB*, and *rpoC* against first-line isoniazid and rifampicin drug. The strain mnpk and swlpk encodes 257 putative and 86 verified virulent genes including type 7 secretion system (T7SS) key genes. The variation in the expression profile of selected T7SS genes, particularly low expression level of *EspK*, raised concern that the mechanism of virulence of mnpk and swlpk might be different from H37Rv strains as *espK* is associated with ATPase *EccC1a* and *EccC1b* which showed high expression level. Briefly, this study shows that the strains mnpk and swlpk are linked with H37Rv having 99% similarity in genomes, but the absence of drug-resistant genes and variation in key genes' expression profile *espK*, *EccE1*, *PPE41*, and *espC* provide a rationale for the future investigation of *M tuberculosis* mnpk and swlpk pathogenesis via RNA sequencing, single-nucleotide polymorphisms, as well as gene manipulation.

KEYWORDS: *M tuberculosis*, genome sequence, MDR 4, ncRNA, outer membrane proteins, Pakistan

RECEIVED: January 9, 2018. **ACCEPTED:** June 18, 2018.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Higher Education Commission of Pakistan under the project of NRPU entitled "Sequencing and mapping of *Mycobacterium tuberculosis* from patients in Pakistan using next-generation sequencing (No-20-3658/R&D/HEC/14).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Muhammad Ibrahim, Department of Biosciences, COMSATS Institute of Information Technology, Sahiwal 57000, Pakistan.
Email: Ibrahim@ciitsahiwal.edu.pk

Introduction

Tuberculosis (TB) is a widespread infectious disease caused by a pathogen known as *Mycobacterium tuberculosis*. *Mycobacterium tuberculosis* is a gram-positive bacteria belonging to the Mycobacteriaceae family.[1] Tuberculosis, a chronic disease, is the primary source of lower respiratory tract illness and causes 10.4 million ill cases and 1.7 million deaths every year[2] (World Health Organization [WHO] report 2017). Unfortunately, in Pakistan, TB has remained ignored in the past and its prevalence has made a severe challenge for medical industry to deal with this disease.[3,4] More seriously, half of the TB cases in Pakistan are diagnosed in Punjab which is known as the most populous province of Pakistan.[4] Among many factors that contribute to TB prevalence and transmission in Pakistan, the predominant factors are incorrect treatment regimes, poverty, unawareness, poor sanitation, and default treatment.[5]

The development of multidrug-resistant TB (MDR-TB) is the result of a number of mutational events, most of which are due to the phenomenon known as epistasis that leads to the formation of resistance to anti-TB drugs.[6] Like all bacteria, *M tuberculosis* accumulates genetics changes over time. For a long time, *M tuberculosis* has been known as highly conserved, having high sequence homology as well as low antigenic diversity.[7] The reason is that in *M tuberculosis*, unlike to other bacterial pathogens, horizontal gene transfer and resistance plasmids did not take part in the acquisition of drug resistance.[7] The advancement in next-generation sequencing technologies are proving promising methods to analysis and explore the entire genetic makeup of the bacteria.[8] These technologies have improved the diagnostics efficacy to scrutinize *M tuberculosis* strains as they move through space and time.[9] The genome sequencing, genetics, and physiological studies in TB, particularly the research on *M tuberculosis*, have largely been ignored in Pakistan, possibly due to insufficient funding. The aims of this study include genome sequencing, comparative genomics analysis, drugs

resistance, and virulence genes analysis following gene expression analyses of T7SS. Results from this study will help to understand and to measure the *M tuberculosis* disease severity in Pakistan by deciphering the sequence biology of *M tuberculosis*.

Materials and Methods

Culturing and identification of M tuberculosis by 16S rRNA gene sequencing

Mycobacterium tuberculosis isolates mnpk and swlpk were identified using 16S rRNA gene as described by El Amin et al[10] and Lawn et al.[11] Briefly, isolates of *M tuberculosis* cells were taken from Löwenstein-Jensen media and killed by adding ciprofloxacin (100 µg/mL) into the culture media before lysis and following incubation for an hour at 30°C. Lysozyme (20 µL) added into Eppendorf tube and stored for 2 hours at 37°C. After adding 30 µL sodium dodecyl sulfate (10%) and 3 µL proteinase K (0.2 mg/mL), tubes were gently vortexed and incubated for 20 minutes at 65°C. About 80 µL of 100 µL NaCl (0.1 M) and *N*-acetyl-*N, N, N*-trimethyl ammonium bromide (40 mM) was added. The Eppendorf tubes were vortexed till the whole solution turned milky and stored at 65°C for 15 minutes. About 750 µL of chloroform-isoamyl alcohol (24:1) was added following vortex and centrifugation till 5 minutes at room temperature at 13 000 rpm. It was followed by ethanol precipitation and collection of DNA in 110 µL TE buffer.

The 16S rRNA gene amplification was performed with universal primer set F-285 5′-AGAGTTTGATCCTGGC TCAG-3′ and Myc-264 3′-TGCACACAGGCCACAAG GGA-5′ as described by Gholoobi et al.[12] Total polymerase chain reaction (PCR) reaction mixture was set in a final volume of 25 µL using 12.5 µL of Taq Master Mix, 9.5 µL water, and 1 µL of DNA template, and forward and reverse primers were used. Negative control used in this experiment was the sample having no DNA template. The PCR conditions used for this reaction were as follows: 94°C for 1 minute, 60°C for 1 minute and 72°C for 1 minute for 35 cycles, and final elongation cycle at 72°C for 10 minutes. The PCR amplification was run at 1.5 % agarose gel electrophoresis and PCR product was eluted using DNA elution kit (Tiagin, TIANGEN Biotech, Beijing, China) and sent for sequencing (Macrogen Seoul, Korea) and obtained sequence were analyzed at NCBI (National Center for Biotechnology Information) using BLAST tool[13] with default parameters. To reveal evolutionary association of drug resistance proteins of *M tuberculosis* strains mnpk and swlpk, a phylogenetic analysis based on 16S rRNA genes was conducted through MEGA 7 using maximum likelihood methods with default parameters.[14]

Genome sequencing, assembly, and annotation

The whole genome sequence of *M tuberculosis* strains mnpk and swlpk from Pakistani patients was done using Illumina MiSeq 300PE throughput 2M which produced, on average, more than 20× coverage with a total of 140 127 reads for strain mnpk and more than 20× coverage with a total of 139 989 reads for strain swlpk. The de novo genomic assembly was conducted using Geneious pro V10.[15] Assembled genome sequences of strains were annotated using Rapid Annotation Subsystem Technology (RAST) server,[16] tRNAscan-SE 1.21,[17] and RNAamer V1.2[18] which provide high-quality functional annotation.

Genome alignment, similarity, and visualization

The genome alignment of *M tuberculosis* swlpk and mnpk was conducted using Mauve software package,[19] which is employed for conducting an alignment of multiple genomes to look at the highly similar subsequences, evolutionary events such as inversions, rearrangement, and likely to reveal the correct global alignment. The multiple genome alignment in the form of graphical map of genome sequence provides a means to quickly view the characteristics of specific genomic regions and study of genomic level evolutionary dynamics. The 2 genomes were also further compared using CGView tool.[20] A circular genomic map of *M tuberculosis* strains was generated using CGView that represent circular genomes into graphical map resulting base composition plots, sequence features, and analysis such as GC skew, GC content, and number of RNAs.

Drug-resistant gene prediction and characterizations

In a previous study, Ze-Jia et al[21] reported various drug resistance genes in *M tuberculosis* strain H37Rv. The sequences of drug-resistant genes were retrieved from *M tuberculosis* strain H37Rv via UniProt and NCBI databases and were used as reference. Using these genes as query, a BLAST search was conducted against *M tuberculosis* strains mnpk and swlpk at RAST server with following parameters, cut-off value 10, ie, 1e−30, and filter size 0.

The drug resistance proteins identified in mnpk and swlpk using reference strain H37Rv were further in silico characterized by performing protein domain, protein family search analysis using NCBI's Conserved Domain Database[22] and InterPro.[23] Protein family search was conducted in drug resistance genes of both strains using Pfam database.[24]

Virulence factors distribution and analysis

The Virulence Factors Database (VFDB)[25] was used for retrieving virulence genes and all the virulent genes used in this study were from *M tuberculosis* H37Rv. The FASTA sequences of encoded virulent genes of H37Rv were used as query sequences to search each matching gene in the 2 *M tuberculosis* strains, mnpk and swlpk, genomes using BLAST too at RAST server (50% coverage and 90% identity thresholds).

RNA extraction and quantitative real-time PCR

ESX-1 to ESX-5 are the 5 T7SS present in *M tuberculosis*. Three out of these 5 systems are necessary for mycobacterial virulence and/or viability.[26] To understand the molecular

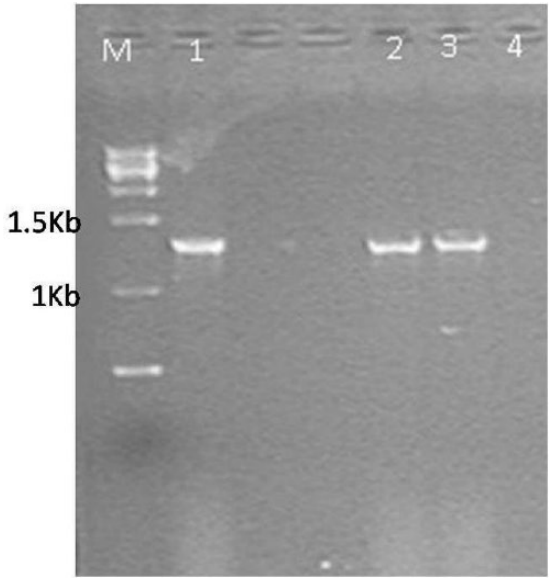


Figure 1. DNA band of amplified 16S rRNA gene M: 10 kb DNA ladder, 1 and 2: 16S rRNA bands of *Mycobacterium tuberculosis* patients from Multan regions. 3: 16S rRNA bands of *M tuberculosis* patients from Sahiwal regions. 4: Negative control.

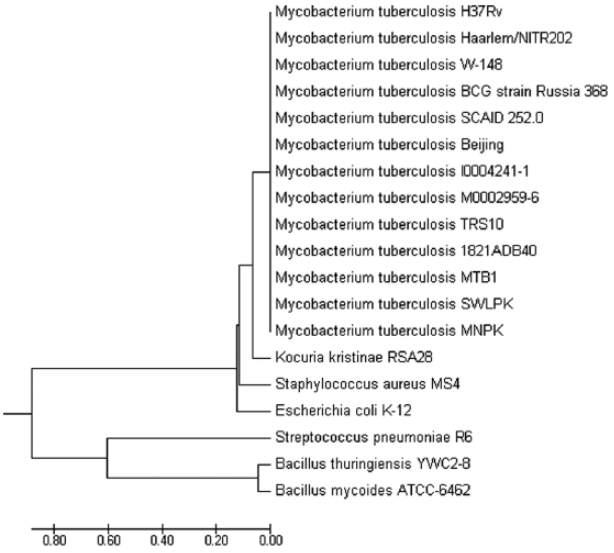


Figure 2. Phylogenetic analysis of *Mycobacterium tuberculosis* strains mnpk and swlpk.

mechanisms of resistance pathogenicity based on the available genomics data, in this study, we further evaluated *M tuberculosis* mnpk and swlpk strains and gene expression profiles of selected and previously verified T7SS.[27] Culture was prepared for RNA extraction as described by Wang et al,[28] using RNA extraction kit according to the manufacturer's instructions (Qiagen, Germantown, MD, USA). About 1.5 µg of RNA was reverse transcribed according to the manufacturer's instructions (Qiagen, USA) with following thermal cycling conditions: 25°C for 10 minutes, 42°C for 60 minutes, and 85°C for 5 minutes. The complementary DNA (cDNA) was kept at −20°C. Two cDNA preparations were made. The assay was conducted using cDNA directly as a template for quantitative reverse transcription PCR (qRT-PCR) on PikoReal Real-Time PCR System (Thermo Scientific). Comparative gene analysis for normalized expression of target transcripts profile was determined using the ΔΔCT method, where CT is the threshold cycle.

Results

Genomic DNA extraction and 16S rRNA gene identification

The PCR reaction was performed using 16S rRNA gene primers and results showed the amplification of 16S rRNA genes at size approximately 1400 bp (base pairs) when compared with 100 bp markers (Figure 1) which lead to the confirmation of isolate as *M tuberculosis* at molecular level. Amplified products were sequenced which give the size of 1.4 kb gene sequence. Global alignment and comparative analysis were conducted at NCBI BLAST analyses using NCBI BLASTn for evolutionary analysis. The results showed that *M tuberculosis* strain has 100% identity with reference strain H37Rv, whereas other close neighbors are *M tuberculosis* strain MTB1 another drug-resistant strain (Figure 2). The partial sequences were submitted to the NCBI and the accession numbers obtained based on 16S rRNA gene of *M tuberculosis* strains mnpk and swlpk are KY271751 and KY287640, respectively. Evolutionary relationships among various biological species or other entities are shown in the form of evolutionary tree (Figure 2) and those strains are out of the cluster of *M tuberculosis* endorsing that the partial sequence of *M tuberculosis* strains is conserved within *M tuberculosis* species.

Genome data acquisition and analysis

The sequenced and assembled genomes of 2 *M tuberculosis* strains mnpk and swlpk consisting of 20 and 12 contigs, respectively, were used in this study using H37Rv as reference. The estimated assembled genome size of strain mnpk is 4.4 Mb and swlpk 4.39 Mb with GC contents of 65.6% (Table 1). The size of strain mnpk as well as swlpk is smaller than *M tuberculosis* strain H37Rv whose total estimated size is 4.41 Mb. The strains mnpk and swlpk encode total 4295 protein-coding sequence (CDS) and 4305 CDS, respectively. Genome comparative analysis at RAST reveals that the closest neighbor of mnpk and swlpk is *M tuberculosis* strain H37Rv which belongs to the Beijing genotypes.

Comparative genomics features analysis

We investigated the overall genome similarities as well as differences between *M tuberculosis* strains mnpk and swlpk with reference strain H37Rv by alignment using Mauve 2.3.1, CGView, and RAST tools. The *M tuberculosis* strains mnpk and swlpk were highly syntenic when compared with H37Rv

Table 1. Genomic feature of *Mycobacterium tuberculosis* mnpk and swlpk.

NAME OF STRAINS	*M TUBERCULOSIS* MNPK	*M TUBERCULOSIS* SWLPK	*M TUBERCULOSIS* H37RV
Genomic size, bp	4 409 295	4 391 906	4 411 532
GC content, %	65.6	65.6	65.6
No. of coding sequences	4305	4295	4367
No. of RNAs	48	48	50
Contigs	20	12	1

Figure 3. Mauve alignment of the genome of *Mycobacterium tuberculosis* strains mnpk and swlpk and H37Rv reference strain.

genomes validating their close relationship (Figure 3). The Mauve results showed that colored blocks represent the individual locally collinear blocks (LCBs), and the homologous LCBs among the 3 strains are connected. Overall, the syntenic regions are shown as colored rounded boxes and unique regions in the genomes are shown as white/gray areas in Figure 3. The regions mutual with a subset of 3 genomes or segments seem to be conserved among all the genomes. Because *M tuberculosis* genomes of strains mnpk and swlpk are unfinished, these results only include the genomic rearrangement that occurred within contigs and the actual number of rearrangements is probably higher.

A circular genome map of strains mnpk and swlpk was generated using CGView using assembled contigs. The circular genome of strains mnpk and swlpk displays the CDS, open reading frame, GC content, number of RNAs, and GC skew (Figure 4) where the first 2 rings represent the CDS and the number of RNAs (tRNA, rRNA, and sRNA), on the forward and reverse strands. The GC contents which *M tuberculosis* encode and are above the genome average (65.6%) are shown by black plot with the peaks extending toward the outside of the circle, whereas those GC contents which are lower than the genome average are extending toward the center mark segments. The innermost plot represents GC skew. Green color shows the positive G+C mean excess of guanine over cytosine, whereas purple color shows negative G–C mean excess of cytosine over guanine (Figure 4).

The predicted protein sequences were annotated to various *clusters of orthologous* groups (COG) categories. Some differences in protein numbers among COG categories of *M tuberculosis* strains mnpk and swlpk and H37Rv genomes were identified (including those listed as protein numbers for *M tuberculosis* strains mnpk and swlpk and H37Rv genomes (Figure 5). This list contains all genes associated with various key function of *M tuberculosis* such as virulence, disease and defense genes, membrane transport genes stress responses genes, drug resistance genes and regulation, and cell signaling genes. The details of these genes are shown and described in Figure 5. It showed several new insights such as the better understanding of the *M tuberculosis* strains', swlpk and mnpk, origin as an obligate pathogen and population genetic characteristics and its molecular evolution both within and between hosts following many features related to antibiotic resistance.

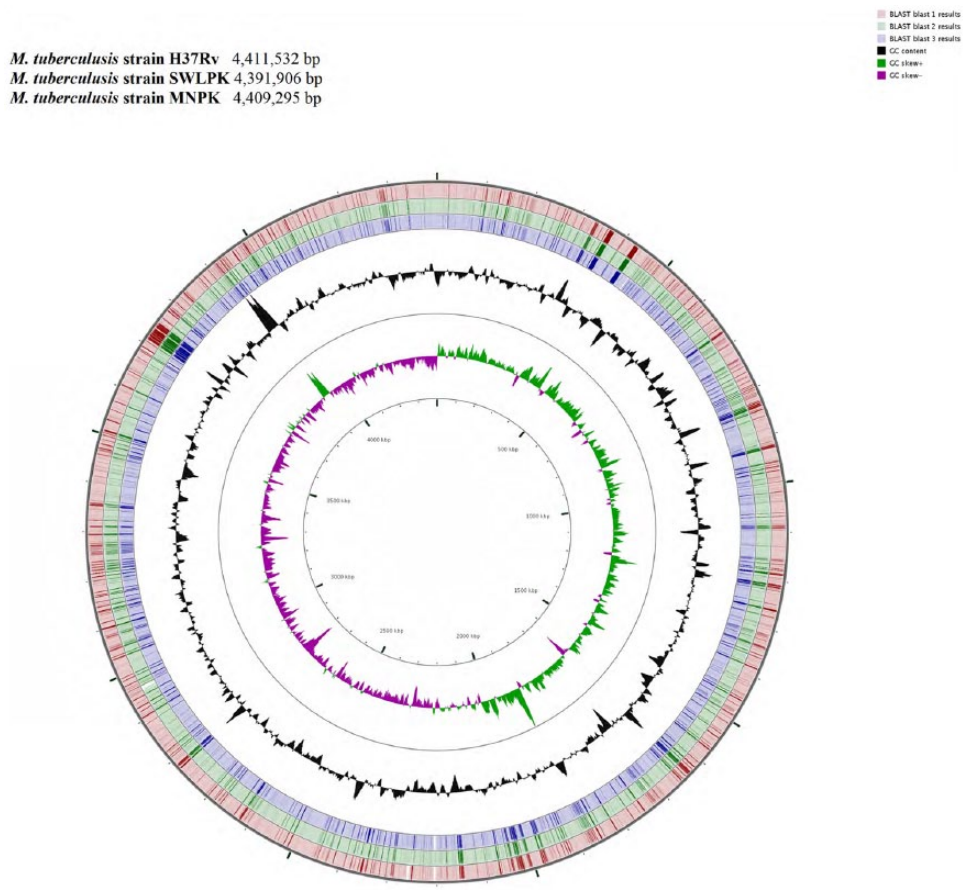


Figure 4. Circular representation of *Mycobacterium tuberculosis* strains mnpk and swlpk.

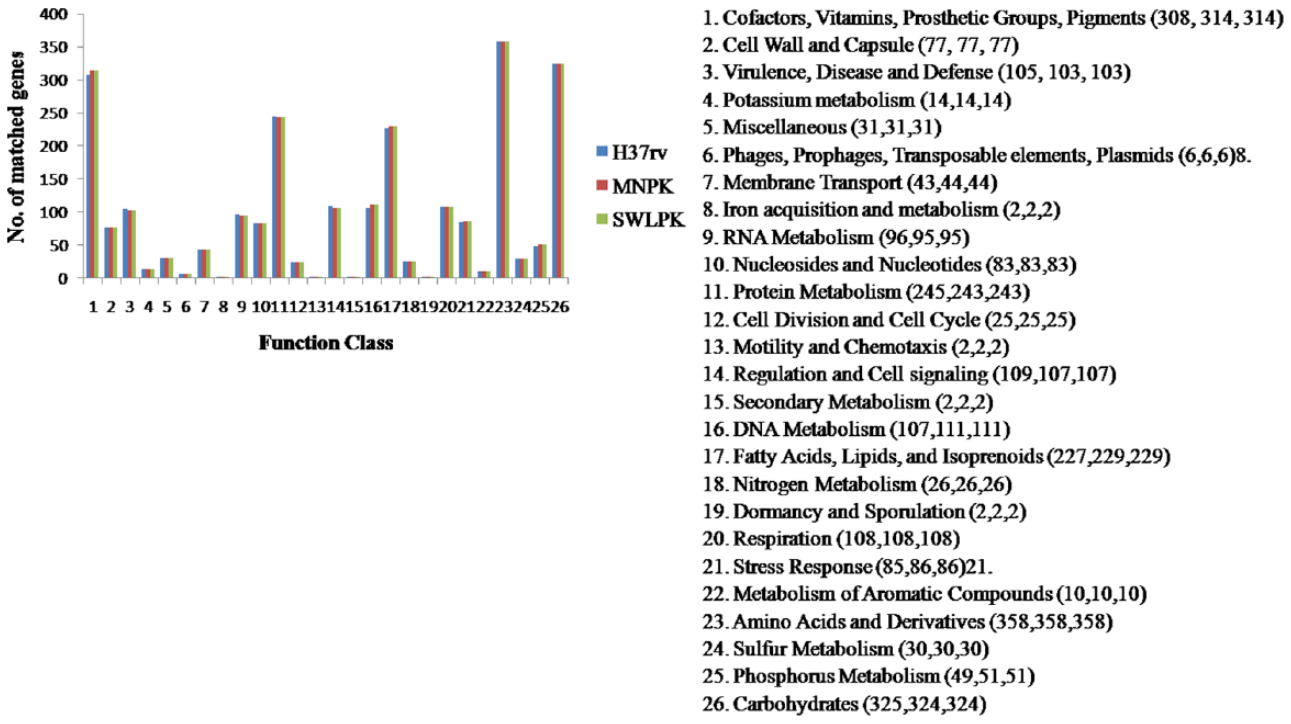


Figure 5. Classification of genes based on cluster of orthologous groups (COG).

Table 2. The drug resistance–associated genes present in mnpk and swlpk.

DRUGS	GENES INVOLVED IN RESISTANCE		
	MNPK	SWLPK	H37RV
Isoniazid (INH)	katG, inhA, ahpC, fabG	katG, inhA, ahpC, fabG	katG, inhA, ahpC, fabG, faDE24, iniA, oxyR
Rifampicin (RMP)	rpoA, rpoB, rpoB	rpoA, rpoB, rpoC	rpoA, rpoB, rpoC
Capreomycin (CPM)	tlyA	tlyA	tlyA, Rrs
Ethambutol (EMB)	embA, embB, embC	embA, embB, embC	embA, embB, embC, embR, iniA, rmlD
Ethionamide (ETH)	ethA	inhA, katG	EthA
Ofloxacin (OFX)	gyrA, gyrB	gyrA, gyrB	gyrA, gyrB
Streptomycin (SM)	gidB	gidB	gidB, rpsL, rrs

Drug resistance gene prediction and analysis

Mycobacterium tuberculosis strains mnpk and swlpk encode several drug-resistant genes (Table 2) against key drugs such as isoniazid, rifampicin, capreomycin, ethambutol (EMB), ethionamide, and ofloxacin. The uptake of drug resistance genes listed in Table 2 not only reveals the association of these strains with Beijing family genotype of *M tuberculosis* but is also vital for development of drug resistance genes–based markers for Pakistan-resourced strains of *M tuberculosis*. Apart from the drug-resistant genes present in strains mnpk and swlpk (Table 2), several drug-resistant genes of *M tuberculosis* already reported in reference strain H37Rv were not found such as faDE24 and iniA which are associated with isoniazid genes; rpoB which is associated with rifampicin gene; embR, iniA, and rmlD, which are ethambutol-associated genes; and rpsL and rrs which are associated with streptomycin (SM) (Table 2). Moreover, the missing genes may therefore have impact on the efficacy of antimicrobial agents, particularly when their presence has not been diagnosed.

Table 3 enlists the protein domains and families associated with drug-resistant genes present in mnpk and swlpk. Our results showed that all the drug resistance genes of *M tuberculosis* strains mnpk and swlpk against first-line or second-line drugs encode similar domain as well as family. No differences were noted among their family-based functional analysis. The rpoA and rpoC genes of strains mnpk and swlpk that participate in RIF resistance belong to DNA-directed RNA polymerase, and the subunit beta prime family encodes the same family. Moreover, gene paralogs were seen such as all the genes, embA, embB, and embC, retrieved from strains mnpk as well as swlpk genomes associated with ethambutol drug resistance belong to the same family Arabinose_trans_C.

Table 3. Protein families of drug resistance genes present in mnpk and swlpk.

NO.	GENE NAME	FAMILY
1	katG	Catalase_peroxidase
2	inhA	Enoyl-ACP_Rdtase_NADH
3	ahpC	AhpC-type
4	fabG	SDR_fam
5	rpoA	DNA-dir_RpoA
6	rpoC	DNA-dir_RpoC
7	embA	Arabinose_trans_C
8	embB	Arabinose_trans_C
9	embC	Arabinose_trans_C
10	ethA	Flavin_mOase-like
11	gyrA	DNA gyrase subunit A
12	gyrB	Topo_IIA
13	tlyA	Haemolysin_A/TlyA
14	gidB	rRNA_ssu_MeTfrase_G

Genome-wide analysis of virulent genes

Despite the high conservation in genome of *M tuberculosis*, the various lineages have diverse degrees of virulence. In our study, we looked into the genome of *M tuberculosis* strains mnpk and swlpk for virulence factors using 257 virulence factor obtained from VFDB as bait (Table S1). Among these, about 86 are experimentally verified as described by Jia et al[27] and VFDB database,[25] whereas the rests are putative ones. The results showed that all the verified virulence factors existed not only in strain mnpk but also in swlpk except hspR (Rv0353), fbpD (Rv 3803c), and devS (Rv3132c) described in Table S1. Moreover, all these verified encoded genes are conserved in 3 strains. We also found that several virulent factors were absent in strains mnpk and swlpk such as mce2E and mce1E putative virulent genes. This may be important to note that these strains may be not highly virulent and verification of these missing drug resistances via RNA sequencing will be highly vital for the ecology of these strains and its association with global strains.

Table 4. Comparative gene expression profile of type 7 secretion systems of *Mycobacterium tuberculosis* mnpk and swlpk.

NAME OF GENE	*M TUBERCULOSIS* MNPK	*M TUBERCULOSIS* SWLPK
esxA (Rv3875)	10.177	12.254
esxB (Rv3874)	13.321	15.531
eccA1 (Rv3868)	12.047	10.294
eccB1 (Rv3869)	21.265	24.661
eccCa1 (Rv3870)	26.509	20.465
eccCb1 (Rv3871)	11.598	9.365
PE35 (Rv3872)	12.210	12.402
eccD1 (Rv3877)	14.955	10.910
espK (Rv3879c)	0.375	0.568
eccE1 (Rv3882c)	0.794	0.594
mycP1 (Rv3883c)	9.537	8.075
espD (Rv3614c)	7.537	5.075
espC (Rv3615c)	3.401	21.585
espA (Rv3616c)	21.549	2.115
espB (Rv3881c)	33.063	27.371
eccA5 (Rv1798)	6.353	2.305
eccE5 (Rv1797)	0.338	0.006
eccD5 (Rv1795)	0.194	0.004
Rv1794	32.194	06.104
esxN (Rv1793)	6.809	8.627
esxM (Rv1792)	2.492	17.757
eccCb5 (Rv1784)	13.501	6.493
eccCa5 (Rv1783)	11.809	13.627
eccB5 (Rv1782)	16.492	3.757
PPE41 (Rv2430c)	03.501	16.493

Quantitative expression of T7SS genes

The quantification results of the expression levels of 25 T7SS genes in 2 strains are shown in Table 4 and the detailed set of primers for all target genes for qRT-PCR are listed in Table S3. In strain mnpk as well as in swlpk, genes espk, *eccE1, eccE5, eccD5* showed the lowest expression levels from 0.177 to 0.004. The T7SS genes esxA, esxB, eccB1, eccCa1, eccb1, PE35, eccD1, espK, *eccE1, eccCb5, and eccCa5* showed higher than 5.0. The genes espB, eccB1, and eccCa1 were exceptional which showed the highest expression levels from 21.265 to 33.06.

We also examined the expressional differences between 2 strains such as the expression levels of espA and Rv1794 genes which were significantly higher in mnpk and ($P < .05$) than in swlpk (Table 4). We also found that the expression of 3 genes (espC, esxM, and PPE41) was significantly lower ($P < .05$) in mnpk isolates than in swlpk isolates.

Discussion

Although many research works had been conducted and still continue with several efforts to get insight into the nature and origin of MDR-TB, dilemma remains obscure. In our study, we have conducted the extensive genomics study to reveal the multidrug resistance as well as the virulence features. By this study, we wished to get closer in the understanding of *multidrug-resistant tuberculosis*. Moreover, we believe that understanding the molecular basis of TB drug resistance using next-generation sequencing technologies could pave the way to fight against this old foe by developing of new diagnosis approaches. The prevalence and rapid dissemination of *Mycobacterium tuberculosis* Beijing strains are reported around the globe and makes it an important issue of public health. By this study, we could not draw any concrete conclusion that about the close association of *M tuberculosis* strains mnpk and swlpk with Beijing genotypes but presence of drug resistance and virulent genes has been shown in many settings of strains mnpk and swlpk such as Beijing genotypes. It showed that these strains could have high level of virulence and multidrug resistance, resulting in rapid progression from infection to active disease and increased transmissibility in *M tuberculosis* H37Rv which is highly virulent, drug resistant, and endemic over Asia.[28]

The difference in genome size with reference strain H37Rv may lead to the prediction that strains mnpk and swlpk may have some missing genes or could be an overestimation due to the incomplete nature of genomes and scaffolds quality could be lower with stretches of N's. The high GC contents in these strains depict the stability of the genome as characteristically high GC% age also verifies the hypothesis that horizontal gene transfer events are nearly missing in *M tuberculosis* which is consistent as described by Šmarda et al.[28]

The *M tuberculosis* strains mnpk and swlpk were highly syntenic with H37Rv genomes verifying their close relationship when compared using Mauve 2.3.1. The composition of bacterial genomes is highly polarized nucleotide in the 2 replichores and this asymmetry of genomic strand could be visualized by GC skew graph. Most prokaryotes and archaea contain only 1 DNA replication origin.[29,30] There will be a GC skew when composition of guanine and cytosine nucleotides is over- or underabundant in a particular region of DNA. The GC skew in strain mnpk and swlpk represents that nucleotide composition is asymmetric between the lagging strand and leading strand which is consistent to *Escherichia coli*.[31]

The presence of key drug-resistant genes and their belonging to PPE and PE family protein indicate that patterns of these drug resistances may differ widely not only in transmissibility of drug resistance but also from single-drug to multiple-drug resistance and which are allied with numerous genetic

mutations.[32] RpoB gene identified in mnpk as well as swlpk is notorious to drug resistance by mutation such as 95% of RIF resistance in *M tuberculosis* strains which is caused by mutation in rpoB genes. It is known that most of the mutations took place in the 81-bp core region of the rpoB gene which encodes the β-subunit of the RNA polymerase and is known as RIF-resistance-determining region.[33] Similarly, other RIF-associated genes also go to various mutations and developing drug resistance such as *rpoC* and *rpoA* in mnpk and swlpk strains may provide preliminary data for RIF-resistant patterns and more importantly these are also known as a surrogate marker of multidrug-resistant genes in *M tuberculosis*.[34] Similarly, resistance to isoniazid has been linked not only with mutations in genes such as *inhA* and *KatG* but also with 2 different genes such as *ahpC*, and oxyR.[7]

The development of resistance against SM drugs in *M tuberculosis* is known as linked with mutations rpsL which code for ribosomal protein S12 and in rrs gene which codes for 16S rRNA.[35] The data regarding these mutations are in a limited proportion for SM-resistant *M tuberculosis* clinically isolated. Both of these genes are not retrieved in *M tuberculosis* strains mnpk and swlpk. It shows that these strains may depend on *gidB* which has been reported that development of resistance is caused by mutations within the *gidB* gene at conserved 7-methylguanosine (m7G) methyltransferase position.[36] The *embA*, *embB*, and *embC* genes were identified in both strains and it has been reported that the cause of resistance to EMB drugs is mutation in these genes which is consequently making *M tuberculosis* more drug resistant.[37]

The existence of these well-known drug-resistant genes in Pakistani strains will be beneficial for further analysis on identification and characterizations and will enrich the TB genetic data sources. The missing of *faDE24*, *iniA*, *rpoR*, *rpsL*, *rrs* genes indicates that further analyses are required and these genes may not be used as a marker when identifying samples from Pakistan-resourced strains. In silico identification of protein family significantly highlights the functional characterizations of proteins or interaction contributing to the overall role of a protein.[38] The results of our study showed that all proteins identified in strain mnpk as well as swlpk encode similar proteins which either is not new and showed the close relation of these proteins with Beijing genotype.

Several studies have identified components of T7SS in *M tuberculosis* and their role in pathogenicity.[39] Although very important proteins such as *EccD1* (Rv3877), *EccCb1* (Rv3871), and *EccCa1* (Rv3870) which are indispensable for secretion of ESX-1 in *M tuberculosis* expression level were significantly better among both strains, but *EccE1* (Rv3882c) is the core membrane protein and espK is an ESX-1 secretion–associated protein. In our study, expression of *espK* is very low and *EccC1a* and *EccC1b* showed higher gene expression. The ESX-1 which is a part of T7SS is the most important virulent component in *M tuberculosis*, but the role of *espK* in this system is not clear. Similarly, McLaughlin et al[40] reported that espB interacts with espK (which showed low level of expression); the interaction of these proteins leads to the interaction of membrane-associated ATPases $EccC_{1a}$ and $EccC_{1b}$. It is well-documented that virulence features of Beijing strains such as H37Rv are more than the modern isolates. There were no appreciable differences except the absence of few genes mentioned above between Beijing genotype strain (H37Rv) and other strains such as mnpk and swlpk (Table S1). We may infer that these missing genes might have been lost through IS6110.

Similarly, some differential gene expression among strain to strain were also noted such as *espA*, *Rv1794* were highly expressed in mnpk and low expressed in swlpk, esxM, espC, PPE41 were highly expressed in swlpk and low expressed in mnpk. The findings of gene expression profile lead toward 2 major key information in gene expression profile such as low-expression *espK* and *eccE1*, and the possible effect of these variations could impact on mechanism of virulence which needs to be disclosed and is conceivable that these various differences noted in our study might have a functional impact.

Conclusions

This study of comparative genomics analysis of 2 *Mycobacterium* strains resourced from Pakistan allowed in-depth genome-wide comparison and covered important genetic heterogeneity. The high similarity with H37Rv but still missing of key drug resistance genes, variation in the expression profile of already verified T7SS proteins, particularly low expression level of espK, which is one of the important virulent factors and associated with membrane-associated ATPase $EccC_{1a}$-$EccC_{1b}$, raises concern that the mechanism of virulence of mnpk and swlpk strains via RNA sequence as well as gene manipulation will be important future prospective.

Acknowledgements

Nucleotide sequence accession number: This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NCKT00000000 for *Mycobacterium tuberculosis* strain MNPK and SWLPK data were available under the accession NYMU00000000.

Author Contributions

AMY performed the experiment. AMY, GZ, AH, YC, AR, AH, and ZB contributed significantly to analysis and manuscript preparation. HB and MI conceived and designed the work, helped in preparation the manuscript that led to the submission. All the authors reviewed and approved the final manuscript.

REFERENCES

1. Fitzgerald DW, Sterling TR, Haas DW. *Mycobacterium tuberculosis*. In: Mandell GL, Bennett JE, Dolin R, eds. *Principle and Practice of Infectious Diseases*. 7th ed. Philadelphia, PA: Churchill Livingstone; 2010:3129–3163.
2. WHO. Tuberculosis fact sheet, 2017, http://www.who.int/en/news-room/fact-sheets/detail/tuberculosis.

3. Gandhi N, Moll A, Sturm A, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet*. 2006;368:1575–1580.
4. Chaovavanich A, Chottanapand S, Manosuthi W, et al. Survival rate and risk factors of mortality among HIV/tuberculosis co-infected patients with and without antiretroviral therapy. *J Acquir Immune Defici Syndr*. 2006;6:42–43.
5. Dagnra A, Adjoh K, Heunda S, et al. Prevalence of HIV-TB co-infection and impact of HIV infection on pulmonary tuberculosis outcome in Togo. *Bulletin de la Société de pathologie exotique*. 2010;3:342–346.
6. Müller B, Borrell S, Rose G, Gagneux S. The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. *Trends Genet*. 2013;29:160–169.
7. Hershberg R, Lipatov M, Small P, et al. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol*. 2008;12:e311.
8. Bishai W. The *Mycobacterium tuberculosis* genomic sequence: anatomy of a master adaptor. *Trends Microbiol*. 1998;6:464–465.
9. Roetzer A, Diel R, Kohl TA, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med*. 2013;2:e1001387.
10. El Amin NM, Hanson HS, Pettersson B, Petrini B, Von Stedingk LV. Identification of non-tuberculous mycobacteria: 16S rRNA gene sequence analysis vs. conventional methods. *Scand J Infect Dis*. 2000;32:47–50.
11. Lawn SD, Mwaba P, Bates M, et al. Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. *Lancet Infect Dis*. 2013;4:349–361.
12. Gholoobi A, Masoudi-Kazemabad A, Meshkat A, Meshkat Z. Comparison of culture and PCR methods for diagnosis of *Mycobacterium tuberculosis* in different clinical specimens. *Jundishapur J Microbiol*. 2014;7:e8939.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–410.
14. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;7:1870–1874.
15. Kearse M, Moir R, Wilson A, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–1649.
16. Aziz RK, Bartel D, Best AA, et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*. 2008;9:75.
17. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of tRNA genes in genomic sequence. *Nucleic Acids Res*. 1997;259:55–964.
18. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007;35:3100–3108.
19. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*. 2004;7:1394–1403.
20. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. *Bioinformatics*. 2005;21:537–539.
21. Ze-Jia C, Qing Y, Hong-Yu Z, Qiang Z, Qing-Ye Z. Bioinformatics identification of drug resistance-associated gene pairs in *Mycobacterium tuberculosis*. *Int J Mol Sci*. 2016;17:1417.
22. Marchler-Bauer A, Derbyshire MK, Gonzales NR, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res*. 2015;43:D222–D226.
23. Hunter S, Apweiler R, Attwood TK, et al. InterPro: the integrative protein signature database. *Nucleic Acids Res*. 2009;37:D211–D215.
24. Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. *Nucleic Acids Res*. 2004;32:D138–D141.
25. Chen L, Yang J, Yu J, et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*. 2005; 33:D325–D328.
26. Houben ENG, Korotkov KV, Bitter W. Take five—Type VII secretion systems of Mycobacteria. *Biochim Biophys Acta (BBA)*. 2014;1843:1707–1716.
27. Jia X, Yang L, Dong M, Chen S, Lv L, et al. The bioinformatics analysis of comparative genomics of *Mycobacterium tuberculosis* Complex (MTBC) provides insight into dissimilarities between intraspecific groups differing in host association, virulence, and epitope diversity. *Front Cell Infect Microbiol*. 2017;21:788.
28. Wang S, Dong X, Zhu Y, et al. Revealing of *Mycobacterium marinum* transcriptome by RNA-seq. *PLoS ONE*. 2013;8:e75828.
29. Liu Y, Wang S, Lu H, Chen W, Wang W. Genetic diversity of the *Mycobacterium tuberculosis* Beijing family based on multiple genotyping profiles. *Epidemiol Infect*. 2016;144:1728–1735.
30. Šmarda P, Bureš P, Horová L, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci USA*. 2014;111:E4096–E4102.
31. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*. 1996;13:660–665.
32. Stephen HG. Evolution of drug resistance in *Mycobacterium tuberculosis*: clinical and molecular perspective. *Antimicrob Agents Chemoth*. 2002;46:267–274.
33. Bishai W. The *Mycobacterium tuberculosis* genomic sequence: anatomy of a master adaptor. *Trends Microbiol*. 2011;6:464–465.
34. Dalla Costa ER, Ribeiro MO, Silva MSN, et al. Correlations of mutations in katG, oxyR-ahpC and inhA genes and in vitro susceptibility in *Mycobacterium tuberculosis* clinical strains segregated by spoligotype families from tuberculosis prevalent countries in South America. *BMC Microbiol*. 2009;9:39.
35. Trauner A, Borrell S, Reither K, Gagneux S. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. *Drugs*. 2014;10:1063–1072.
36. Wong SY, Lee JS, Kwak HK, Via LE, Boshoff HI, Barry CE. Mutations in gidB confer low-level streptomycin resistance in *Mycobacterium tuberculosis*. *Antimicrob Agent Chemother*. 2011;6:2515–2522.
37. Cui Z, Li Y, Cheng S, et al. Mutations in the embC-embA intergenic region contribute to *Mycobacterium tuberculosis* resistance to ethambutol. *Antimicrob Agents Chemother*. 2014;58:6837–6843.
38. Abdallah AM, Bestebroer J, Savage ND, et al. Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation. *J Immunol*. 2011;187:4744–4753.
39. Daleke MH, Ummels R, Bawono P, et al. General secretion signal for the Mycobacterial type VII secretion pathway. *Proc Natl Acad Sci USA*. 2012;109:11342–11347.
40. McLaughlin B, Chon JS, MacGurn JA, Carlsson F, Cheng TL, Cox JS. A mycobacterium ESX-1-secreted virulence factor with unique requirements for export. *PLoS Pathog*. 2007;3:e10.