ORIGINAL ARTICLE

https://doi.org/10.4046/trd.2017.0091
ISSN: 1738-3536(Print)/2005-6184(Online) • Tuberc Respir Dis 2018;81:216-221

trd Tuberculosis & Respiratory Diseases

Comparing Two *Mycobacterium tuberculosis* Genomes from Chinese Immigrants with Native Genomes Using Mauve Alignments

Sungweon Ryoo, Ph.D.[1], Jeongsoo Lee, Ph.D.[2], Jee Youn Oh, M.D.[3], Byeong Ki Kim, M.D.[4], Young Kim, M.D.[4], Je Hyeong Kim, M.D.[4], Chol Shin, M.D.[4] and Seung Heon Lee, M.D., Ph.D.[4]

[1]Clinical Research Centre, Masan National Tuberculosis Hospital, Changwon, [2]LAS Inc., Gimpo, [3]Division of Pulmonary, Sleep, and Critical Care Medicine, Department of Internal Medicine, Korea University Guro Hospital, Korea University College of Medicine, Seoul, [4]Division of Pulmonary, Sleep, and Critical Care Medicine, Department of Internal Medicine, Korea University Ansan Hospital, Korea University College of Medicine, Ansan, Korea

Background: The number of immigrants with tuberculosis (TB) increases each year in South Korea. Determining the transmission dynamics based on whole genome sequencing (WGS) to cluster the strains has been challenging.

Methods: WGS, annotation refinement, and orthology assignment for the GenBank accession number acquisition were performed on two clinical isolates from Chinese immigrants. In addition, the genomes of the two isolates were compared with the genomes of *Mycobacterium tuberculosis* isolates, from two native Korean and five native Chinese individuals using a phylogenetic topology tree based on the Multiple Alignment of Conserved Genomic Sequence with Rearrangements (Mauve) package.

Results: The newly assigned accession numbers for two clinical isolates were CP020381.2 (a Korean-Chinese from Yanbian Province) and CP022014.1 (a Chinese from Shandong Province), respectively. Mauve alignment classified all nine TB isolates into a discriminative collinear set with matched regions. The phylogenetic analysis revealed a rooted phylogenetic tree grouping the nine strains into two lineages: strains from Chinese individuals and strains from Korean individuals.

Conclusion: Phylogenetic trees based on the Mauve alignments were supposed to be useful in revealing the dynamics of TB transmission from immigrants in South Korea, which can provide valuable information for scaling up the TB screening policy for immigrants.

Keywords: Immigrants; *Mycobacterium tuberculosis*; Strains; Whole Genome Sequencing

Address for correspondence: Seung Heon Lee, M.D., Ph.D.
Division of Pulmonary, Sleep and Critical Care Medicine, Department of Internal Medicine, Korea University Ansan Hospital, Korea University College of Medicine, 123 Jeokgeum-ro, Danwon-gu, Ansan 15355, Korea
Phone: 82-31-4124973, **Fax:** 82-31-4125604
E-mail: lee-sh@korea.ac.kr

Received: Aug. 3, 2017
Revised: Oct. 1, 2017
Accepted: Nov. 1, 2017
Published online: Jun. 19, 2018

Introduction

One-third of the world's population is infected with *Mycobacterium tuberculosis*. In high tuberculosis (TB) burden countries, the rapid diagnosis and treatment of infectious TB are major concerns. However, in low TB burden countries, the main TB control policy is to screen for active and latent TB in selected risk groups, including immigrants from TB-endemic areas.

The TB burden in South Korea is intermediate, with an annual incidence of 80/100,000 in 2015[1]. To achieve the Sustainable Development Goals for 2030 proposed by the World Health Organization, the Korean government launched a comprehensive plan for TB control in 2013[2], and it is scaling up the policy to include intensive TB screening for immigrants from TB endemic areas because the number of immigrants with TB has increased each year in Korea[3].

The most common race of TB patients among immigrants is Korean-Chinese (Chosun race) from northeastern China, including the Yanbian Korean Autonomous Prefecture[4]. However, it is essential to determine the molecular epidemiology of *Mycobacterium tuberculosis* isolates and identify the transmission dynamics in the Korean-Chinese population to define a strategy for immigration TB screening. A clustering of *M. tuberculosis* strains from native Koreans (K family) was reported[5], and it was different from that of foreign-born patients according to restriction fragment length polymorphism profile analyses[6].

Recently, genome organization studies have become possible because whole genome sequences, including the genomes of microorganisms, can be sequenced using next-generation sequencing (NGS)[7]. Moreover, phylogenetic relationship studies using the Multiple Alignment of Conserved Genomic Sequence with Rearrangements (Mauve) package based on genomic bioinformatics are promising, because Mauve package is suitable for sequence comparison not posed by short sequences in the presence of rearrangements and horizontal transfer[8].

The aims of this study are to characterize TB strains isolated from immigrants from China based on molecular and bioinformatics as well as to compare them with clinical isolates from native Chinese and Korean individuals.

Materials and Methods

1. Bacterial culture conditions and DNA extraction

Two new clinical isolates from immigrants who were treated in Korea University Guro Hospital and Korea University Ansan Hospital were used for whole genome sequencing. The mycobacterial cultures were grown at 37°C on Löwenstein-Jensen medium. Strains harvested at the early exponential phase were used for DNA extraction. All DNA extraction was performed as previously described[9]. At least one loop-full of cells (100 mg wet weight) were washed twice with TE (Tris-HCl pH 8.0, 10 mM; EDTA 1 mM) and an equal volume of saturated cesium chloride solution containing 1% Triton-X was added to it. Further purification was performed by the usual phenol-chloroform extraction and DNA was spooled out after adding 0.5 vol. of 5 M ammonium acetate and 0.75 vol. of isopropanol; it was washed with alcohol, partially dried and suspended in TE buffer. It was then treated with RNase A (50 μg/mL) for 30 minutes at 37°C, re-extracted with chloroform-isoamyl alcohol, precipitated and finally re-suspended in TE buffer.

2. *M. tuberculosis* genome sequences assembly

Whole-genome shotgun sequencing of the two strains was carried out using PacBio SMRT sequencing technology[10] with ~150× depth. For assembly of the genomes, we applied the recently described hierarchical genome assembly process (HGAP) to the SMRT cells of sequencing data generated from an 8- to 10-kb insert library[11]. Pre-assembly error correction was performed with HGAP of SMRT analysis version 2.3.0 (Pacific Biosciences, Menlo Park, CA, USA) using default parameters. Error-corrected reads were then assembled using Celera Assembler[12]. To produce circular full-length sequence, each end section was resolved manually.

3. Bioinformatics study and Mauve alignments for *M. tuberculosis* strains

We compared the genome topology network (GTN) of each strain using a bootstrap topology tree and detected the locally collinear blocks of the conserved segments using a modified MUSCLE (multiple sequence alignment method with reduced time and space complexity) global alignment algorithm. We used the Mauve package for the identification and alignment of the conserved genomic DNA sequences of the nine *M. tuberculosis* strains[8]. Initially, local multiple alignments that had unique subsequences shared by two or more input genomes were selected, and ungapped extension was performed until the seed pattern no longer matched. Then, a progressive genome alignment according to a guide tree was built up after computing a pairwise genome content distance using the neighbor joining method and a pairwise breakpoint distance matrix. During progressive alignment, the breakpoint penalty according to the expected level of sequence divergence and the number of well-supported genomic rearrangements among the pair of input genomes were scaled. Anchor alignment using a global genome alignment algorithm was performed, and alignments that had unrelated sequences were ultimately rejected. The algorithm began with the initial set of matching regions (multiple maximal unique matches) represented as connected blocks. The matches were partitioned into a minimum set of collinear blocks. Each sequence of identically colored blocks represents a collinear set of matching regions. One connecting line is drawn per collinear block.

This study was approved by the Institutional Review Board (IRB) of Korea University Ansan Hospital (KUAS15157-001) and Korea University Guro Hospital (2017GR0301) with waivers of informed consents.

Table 1. Description of the two complete whole genome sequence from immigrants

Strain	*Mycobacterium tuberculosis* strain MTB1	*Mycobacterium tuberculosis* strain MTB2
GenBank accession No.	CP020381.2	CP022014.1
Chromosome size (bp)	4,433,542	4,417,716
Genes, coding sequences (CDS)	4,306 Genes genes comprising 4,255 CDS (total)	4,290 Genes comprising 4,239 CDS (total)
RNAs*	3 rRNA genes (5S, 16S, and 23S), 45 tRNAs, and 3 ncRNAs were annotated	In 117 pseudo genes, 3 rRNA genes (5S, 16S, and 23S), 45 tRNAs, and 3 ncRNAs were annotated

*Annotated using NCBI Prokaryotic Genome Annotation Pipeline (PGAP; http://www.ncbi.nlm.nih.gov/genome/annotation_prok).

Table 2. Genome sequences and origin of TB strains in comparative groups

Identification	Access number	DST	Origin	Reference
1_CP007809_ER17_ko	CP007809	Sensitive	KIT, South Korea	13
2_NZ_CP007803_ko	CP007803	Sensitive	Masan, South Korea	14
3_CP009100_IZ84	CP009100	XDR	Zunyi, China	15
4_CP009101_JE53	CP009101	XDR	Zunyi, China	15
5_CP001641_CCDC5079.fna	CP001641	Sensitive	Fujian Province, China	16
6_CP001642_CCDC5180.fna	CP001642	MDR	Fujian Province, China	16
7_NZ_CP009426.fna	CP009426	Sensitive	Beijing, China	17
MTB_1_RevComp_circular	CP020381.2	RIF resistant	Similar to China	N/A
MTB_2_RevComp_circular	CP022014.1	MDR	Similar to South Korea	N/A

TB: tuberculosis; DST: drug sensitivity test; KIT: Korean Institute of Tuberculosis; XDR: extensively drug-resistant; MDR: multi-drug resistant; RIF: rifampin; N/A: non-applicable.

Results

1. Whole-genome sequences of clinical isolates of two immigrants

Table 1 presents the characteristics of two complete whole genome sequences from immigrants. The complete genome sequence of *M. tuberculosis* MTB_1 has been assigned the GenBank accession number CP020381.2. The *M. tuberculosis* MTB1 genome comprises one chromosome of 4,433,542 bp. In total, 4,306 genes comprising 4,255 coding sequences (CDS), three rRNA genes (5S, 16S, and 23S), 45 tRNAs, and three non-coding RNAs (ncRNAs) were annotated using NCBI Prokaryotic Genome Annotation Pipeline (PGAP; http://www.ncbi.nlm.nih.gov/ genome/annotation_prok). The patient with the MTB_1 strain was a 39-year-old woman. She had recently emigrated from Shandong Province in China to South Korea to marry a Korean. Her chest computed tomogram showed a cavity in the left lung apex representing the reactivation of a previous TB infection.

The complete genome sequence of *M. tuberculosis* MTB_2 has been assigned the GenBank accession number CP022014.1. The *M. tuberculosis* MTB2 genome comprises one chromosome of 4,417,716 bp. In total, 4,290 genes comprising 4,239 CDS, 117 pseudo genes, three rRNA genes (5S, 16S, and 23S), 45 tRNAs, and three ncRNAs were annotated. The patient with MTB_2 strain was a 33-year-old man. He had moved to South Korea from the Yanbian Korean autonomous prefecture where most Korean-Chinese (Chosun race) reside, 2 years before the TB diagnosis was confirmed. His chest computed tomogram showed centrilobular nodules and branching opacities in both upper lung fields.

2. Epidemiologic characteristics of TB strains for comparison

As shown in Table 2, the genomic sequences of seven strains (from Nos. 1 to 7) from native Chinese and South Korean individuals, which were obtained from a public website, were compared with MTB_1 and MTB_2 from the immigrants to South Korea. The Nos. 1 and 2 strains, which are known to be strains from native South Koreans, were isolated during an outbreak of pulmonary TB in high schools[13,14]. The Nos. 3, 4, 5, and 6 strains were reported as strains from native China. The Nos. 3 and 4 strains were isolated from Chinese individuals living in Zunyi, which is located in southwest China[15]; the Nos. 5 and 6 strains were isolated from Chinese individuals living in Fujian Province, which is located in southeast China[16]; and

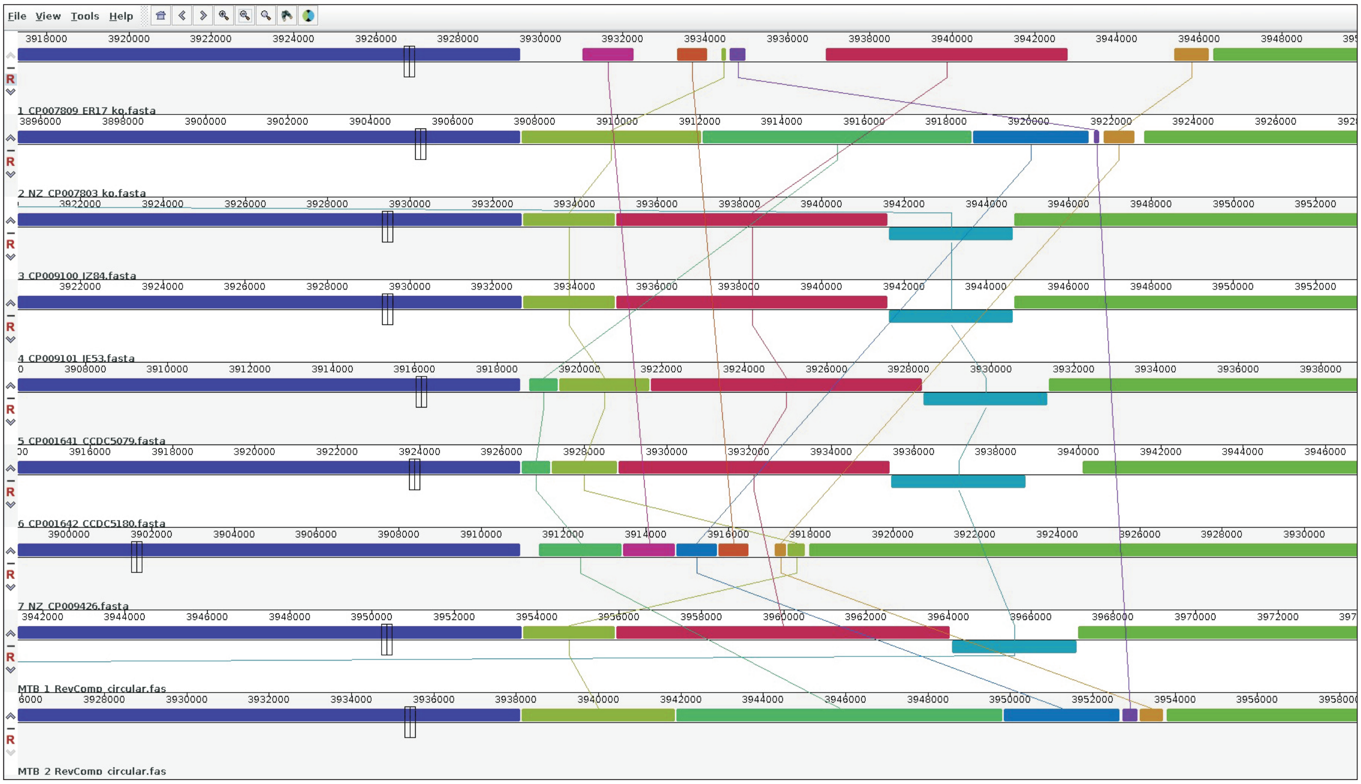

Figure 1. Mauve alignment of nine *Mycobacterium tuberculosis* strains.

the No. 7 strain was isolated from Chinese individuals living in Beijing, which is located in northwest China[17]. When the strains from the immigrants were analyzed, the MTB_2 strain was similar to the No. 1 and No. 2 strains isolated from native Koreans. On the other hand, the MTB_1 strain was similar to the strains isolated from the native Chinese individuals.

3. Mauve alignment and phylogenetic trees of nine *M. tuberculosis* strains

Figure 1 represents the Mauve alignment for nine *M. tuberculosis* strains, which shows collinear set of matched colored regions. And, in Figure 2, the phylogenetic trees exhibit similarities and differences in topology. The phylogenetic analysis of the GTN resulted in a rooted phylogenetic tree grouping the nine strains into two lineages: (1) strains from Chinese individuals and (2) strains from Korean individuals. However, even though the MTB_1 strain was grouped to a lineage from Chinese individuals (from the Nos. 3 to 7 strains), the MTB_1 strain was more similar to the Nos. 3 and 4 strains (Zunyi, China) than to the Nos. 5, 6, and 7 strains (Fujian Province and Beijing, China). On contrast, MTB_2 was grouped to a lineage from the Korean individuals (Nos. 1 and 2) that was different from the strains from the Chinese individuals, closer to the No. 2 strain than to the No. 1 strain.

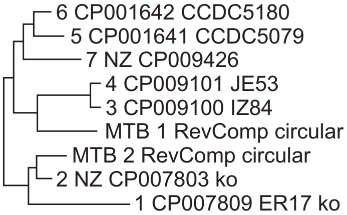

Figure 2. Phylogenetic tree showing similarities and differences in topology.

Discussion

The whole-genome sequences of two clinical isolates from the immigrants were deposited in the GenBank under accession numbers CP020381.2 and CP022014.1. Furthermore, after classifying the alignments of the conserved genomic DNA sequences from nine strains into two lineages using phylogenetic trees based on Mauve alignments, we were able to infer the origin of the strains from the immigrants.

Considering the similarities of the MTB_2 strain isolated from a male Korean-Chinese (Chosun race) individual who emigrated from Yanbian Province, China, with the native Korean K1 strains (Nos. 1 and 2), he may have been infected with a strain from native Koreans after immigration, as Jeon et al.[6]

reported, because the Korean K1 strains have been reported as the most widely distributed characteristic strains in South Korea[5]. Otherwise, another plausible explanation is that the strains from Yanbian Province, where Chosun race live because of the geographical proximity to the Korean Peninsula, may have the same lineage as the native Korean K strains. Therefore, additional molecular epidemiology studies with large numbers of samples must be conducted to characterize the strains from Korean-Chinese (Chosun race) individuals.

However, considering the similarities of the MTB_1 strain isolated from a Chinese woman who emigrated from Shandong Province in China with the native Chinese strains (Nos. 3, 4, 5, 6, and 7), she is thought to have settled in South Korea without a detailed proper screening for active or latent TB during the immigration process. Therefore, to prevent the inflow of TB by immigrants, a more aggressive and thorough TB screening process for immigrants must be performed to reach the goals of the 2030 TB elimination project in South Korea.

Because the distribution of specific resistance-conferring mutations is fairly constant worldwide, suggesting that drug resistance has arisen via common mechanisms[18], further bioinformatic analyses of multi-drug resistant TB or extensively drug-resistant TB isolates from China and Korea are needed, which should include a large number samples with accurate information including phenotypic resistance, to determine the phylogenetic relationship with drug resistance.

This study had several limitations. First, we could not analyze a sufficient number of isolates because it was very difficult to extract and isolate mycobacterial DNA from stock strains, as a time-consuming and tedious process was necessary to disrupt the thick and lipopolysaccharide-rich cell wall without causing damage to the genomic DNA, even though a recently approved method was used[9]. Second, it seems too early to consider NGS a generalized genome analysis method for obtaining full *M. tuberculosis* sequences, annotation refinement, and orthology assignment because it is very expensive. Finally, there have been no reports on the differences between the Chinese strains and the Korean strains using bio-informatics analyses; therefore, the phylogenetic trees for the strains constructed using the small number of samples were likely not definite.

In conclusion, phylogenetic trees based on Mauve alignments supposed to reveal the dynamics of TB transmission from immigrants to South Korea, providing important information for the scaling up of the TB screening policy for immigrants, especially from China.

Authors' Contributions

Conceptualization: Ryu S, Lee SH. Formal analysis: Lee J, Lee SH. Data curation: Ryu S, Oh JY, Kim BK, Kim Y, Lee SH. Supervision: Kim JH, Shin C. Validation: Ryu S, Lee J, Lee SH. Writing - original draft preparation: Ryu S, Lee SH. Writing - review and editing: all authors. Approval of final manuscript: all authors.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This study was supported by a 2016 grant from the Korean Academy of Tuberculosis and Respiratory Disease.

References

1. World Health Organization. Global tuberculosis report 2016 [Internet]. Geneva: World Health Organization; 2016 [cited 2018 Jan 5]. Available from: http://apps.who.int/iris/bitstream/10665/ 250441/1/9789241565394-eng.pdf?ua=1.
2. Lee YK, Kwon YH, Lee SC, Sohn HJ, Koh YW. Overview of tuberculosis control and prevention polices in Korea. Public Health Wkly Rep 2015;8:651-6.
3. Korea Centers for Disease Control and Prevention. Annual report on the notified tuberculosis in Korea. Cheongwon: Korea Centers for Disease Control and Prevention; 2015.
4. Min GH, Kim Y, Lee JS, Oh JY, Hur GY, Lee YS, et al. Social and clinical characteristics of Immigrants with tuberculosis in South Korea. Yonsei Med J 2017;58:592-7.
5. Kim SJ, Bai GH, Lee H, Kim HJ, Lew WJ, Park YK, et al. Transmission of *Mycobacterium tuberculosis* among high school students in Korea. Int J Tuberc Lung Dis 2001;5:824-30.
6. Jeon CY, Kang H, Kim M, Murray MB, Kim H, Cho EH, et al. Clustering of *Mycobacterium tuberculosis* strains from foreign-born patients in Korea. J Med Microbiol 2011;60(Pt 12):1835-40.
7. Jiang J, Gu J, Zhang L, Zhang C, Deng X, Dou T, et al. Comparing *Mycobacterium tuberculosis* genomes using genome topology networks. BMC Genomics 2015;16:85.
8. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 2004;14:1394-403.
9. Bose M, Chander A, Das RH. A rapid and gentle method for the isolation of genomic DNA from mycobacteria. Nucleic Acids Res 1993;21:2529-30.
10. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science 2009;323:133-8.
11. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assem-

blies from long-read SMRT sequencing data. Nat Methods 2013;10:563-9.

12. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 2012;30:693-700.
13. Park YK, Kang H, Yoo H, Lee SH, Roh H, Kim HJ, et al. Whole-genome sequence of *Mycobacterium tuberculosis* Korean strain KIT87190. Genome Announc 2014;2:e01103-14.
14. Han SJ, Song T, Cho YJ, Kim JS, Choi SY, Bang HE, et al. Complete genome sequence of *Mycobacterium tuberculosis* K from a Korean high school outbreak, belonging to the Beijing family. Stand Genomic Sci 2015;10:78.
15. Chen L, Zhang DT, Zhang J, Su YA, Zhang H. Whole-genome sequences of two clinical isolates of extensively drug-resistant *Mycobacterium tuberculosis* from Zunyi, China. Genome Announc 2014;2:e00910-14.
16. Zhang Y, Chen C, Liu J, Deng H, Pan A, Zhang L, et al. Complete genome sequences of *Mycobacterium tuberculosis* strains CCDC5079 and CCDC5080, which belong to the Beijing family. J Bacteriol 2011;193:5591-2.
17. Wan X, Qian L, Hou S, Drees KP, Foster JT, Douglas JT. Complete genome sequences of Beijing and Manila family strains of *Mycobacterium tuberculosis*. Genome Announc 2014;2:e01135-14.
18. 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:e7407.