Data in Brief 33 (2020) 106388

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# The whole genome sequence data analyses of a *Mycobacterium tuberculosis* strain SBH321 isolated in Sabah, Malaysia, belongs to Ural family of Lineage 4

Jaeyres Jani[a], Zainal Arifin Mustapha[b], Chin Kai Ling[c], Amabel Seow Ming Hui[d], Roddy Teo[d], Kamruddin Ahmed[a,e,*]

[a] *Borneo Medical and Health Research Centre, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*
[b] *Department of Medical Education, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*
[c] *Department of Biomedical Sciences and Therapeutics, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*
[d] *Tuberculosis and Leprosy Control Unit, Sabah State Health Department, Kota Kinabalu, Sabah, Malaysia*
[e] *Department of Pathobiology and Medical Diagnostics, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*

ARTICLE INFO

*Article history:*
Received 2 July 2020
Revised 29 September 2020
Accepted 30 September 2020
Available online 8 October 2020

*Keywords:*
*Mycobacterium tuberculosis*
Whole genome sequencing
Next generation sequencing
Ural family
Sabah
Malaysia

ABSTRACT

In 2019, 10 million new cases of tuberculosis have been reported worldwide. Our data reports genetic analyses of a *Mycobacterium tuberculosis* strain SBH321 isolated from a 31-year-old female with pulmonary tuberculosis. The genomic DNA of the strain was extracted from pure culture and subjected to sequencing using Illumina platform. *M. tuberculosis* strain SBH321 consists of 4,374,895 bp with G+C content of 65.59%. The comparative analysis by SNP-based phylogenetic analysis using maximum-likelihood method showed that our strain belonging to sublineage of the Ural family of Europe–America–Africa lineage (Lineage 4) and clustered with *M. tuberculosis* strain OFXR-4 from Taiwan. The whole genome

* Corresponding author at: Department of Pathobiology and Medical Diagnostics, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia.
*E-mail address:* ahmed@ums.edu.my (K. Ahmed).

https://doi.org/10.1016/j.dib.2020.106388

sequence is deposited at DDBJ/ENA/GenBank under the accession WCJH00000000 (SRR10230353).



## Specifications Table

| | |
|---|---|
| Subject | Medicine and Public Health |
| Specific subject area | Microbiology |
| Type of data | Whole genome sequence with gene annotation and comparative genomic of *Mycobacterium tuberculosis* strain SBH321 Ural strain from Sabah, Malaysia. |
| How data were acquired | *M. tuberculosis* cultured in BACTEC MGIT system and the gDNA extracted and sequenced. *de novo* whole genome sequencing, phylogenetic and variant calling data analysis subsequently conducted. |
| Data format | Raw data and whole genome sequence data analysis. |
| Parameters of data collection | Genomic DNA from pure culture. |
| Description of data collection | *M. tuberculosis* identified by Genexpert MTB/RIF cultured in 7H9 Middlebrook liquid media and incubated in BACTEC MGIT system. Genomic DNA isolated using Masterpure Complete DNA and RNA purification kit. Whole genome sequencing performed by Illumina HiSeq 4000 system. Genome assembled through SPAdes version 3.11, variant calling by Genome Analysis Toolkit (GATK), and annotation by NCBI Prokaryotic Genome Annotation Pipeline (PGAP). |
| Data source location | Lahad Datu, Sabah, Malaysia |
| Data accessibility | Repository name: Mendeley data (https://data.mendeley.com/datasets/yw63xz9rgm/1) Data is publicly available at NCBI Genbank from the following links: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA575111 https://www.ncbi.nlm.nih.gov/biosample/SAMN12877714 https://www.ncbi.nlm.nih.gov/nuccore/SRR10230353 |

## Value of the Data

- Since *Mycobacterium tuberculosis* of the Ural family has never been reported in Malaysia, the whole genome sequence of this strain could provide fundamental knowledge and insight towards understanding its microbial activities.
- The data could be used in the examination of the molecular characteristics and genetic variability of the *M. tuberculosis* strain which would benefit molecular epidemiologists.
- The data, an important source towards understanding the relationship between *M. tuberculosis* strains from Sabah and other regions, could assist in developing informed policy in the design and implementation of tuberculosis control programme.

## 1. Data Description

*Mycobacterium tuberculosis* is divided into seven lineages, among these, Lineage 4 is highly diversified. Several families of strains are found in this lineage and Ural amongst them [1]. The Ural family, first reported in 2005, constituted 15% of the researched strains in the Middle Ural area of Russia [2,3]. Besides Russia, these strains are mainly found in Iran, Afghanistan, Pakistan, Turkey, Kyrgyzstan, Ukraine, Abkhazia, Kazakhstan, Georgia and Armenia [1]. Apart from only one report from northern India and north eastern China, these satins have not yet been reported from any South, East, and Southeast Asia countries [1]. The Malaysian Borneo state of Sabah, located in the region of Southeast Asia where tuberculosis cases are increasing, has one of the highest cases of tuberculous in Malaysia [4]. The exact reason for this high incidence of tuberculosis is unknown; it was due to this gap of data that we undertook a project to perform

whole genome sequence (WGS) analysis of *M. tuberculosis* strains from Sabah to determine their genetic characterizations.

Data analysis of the WGS of *M. tuberculosis* strain SBH321 from Sabah, Malaysia is documented in this paper. *M. tuberculosis* strain SBH321 was isolated from a 31-year-old Filipino female patient from Lahad Datu, Sabah. She was confirmed to be tuberculosis-positive by GeneXpert MTB/RIF. The strain was cultured using BACTEC MGIT system. WGS was performed by Illumina HiSeq 4000 system. The *de novo* assembly of genome generated 114 contigs with N50 of 193,257 bp. The genome size was 4,374,895 bp with 4059 predicted genes and 65.59% of G+C content. The comparative genomic analysis using the WGS of 78 strains revealed that SBH321 strain belonged to the Ural family of Lineage 4 and resembled the strains of OFXR-4 from Taiwan (Fig. 1).

## 2. Experimental Design, Materials and Methods

### 2.1. Bacterial culture and DNA extraction

The *M. tuberculosis* strain SBH321 was isolated from the sputum of a patient with tuberculosis diagnosed by GeneXpert MTB/RIF. The strain was grown in 7H9 Middlebrook medium, and incubated at 37 °C in a BACTEC MGIT 320 system (Becton-Dickinson, Oxford, United Kingdom). The genomic DNA extraction was performed using Masterpure Complete DNA and RNA purification kit (Epicenter Inc., Madison, Wisconsin, USA) according to the manufacturer's instructions but with modification in the lysis step by extending the lysis duration to 16 h. The quality of the extracted DNA was determined by Nanodrop 2000c spectrophotometer (ThermoFisher Scientific, USA).

### 2.2. Whole genome sequencing and bioinformatic analyses

99% of the genome was completely sequenced using 386 × sequencing coverage, generated a total of 11,355,058 paired reads of a 150-bp paired-end library via NEB next Ultra kit (Illumina, San Diego, CA). The sequencing data was deposited in the Sequence Read Archive (SRA) (Bio-sample accession number of SAMN12878104) and under the bio-project accession number PRJNA575111. SPAdes version 3.11.1 [5] software was used for *de novo* assembly, and NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [6] software utilized to annotate the generated contigs.

### 2.3. Assembly statistic

| | |
|---|---|
| Sequencing depth | 386 × |
| Total length of sequences (bp) | 4,374,895 |
| Total number of contigs | 114 |
| N50 (bp) | 193,257 |
| GC (%) | 65.59 |
| CDSs | 4059 |
| tRNAs | 52 |
| 5s, 16s, 23s rRNA | 1, 1, 1 |

### 2.4. Variant calling

For the variant calling analysis, the raw sequence reads were first aligned to a reference genome, *M. tuberculosis* H37Rv (GenBank accession number NC_000962.3) using BWA MEM

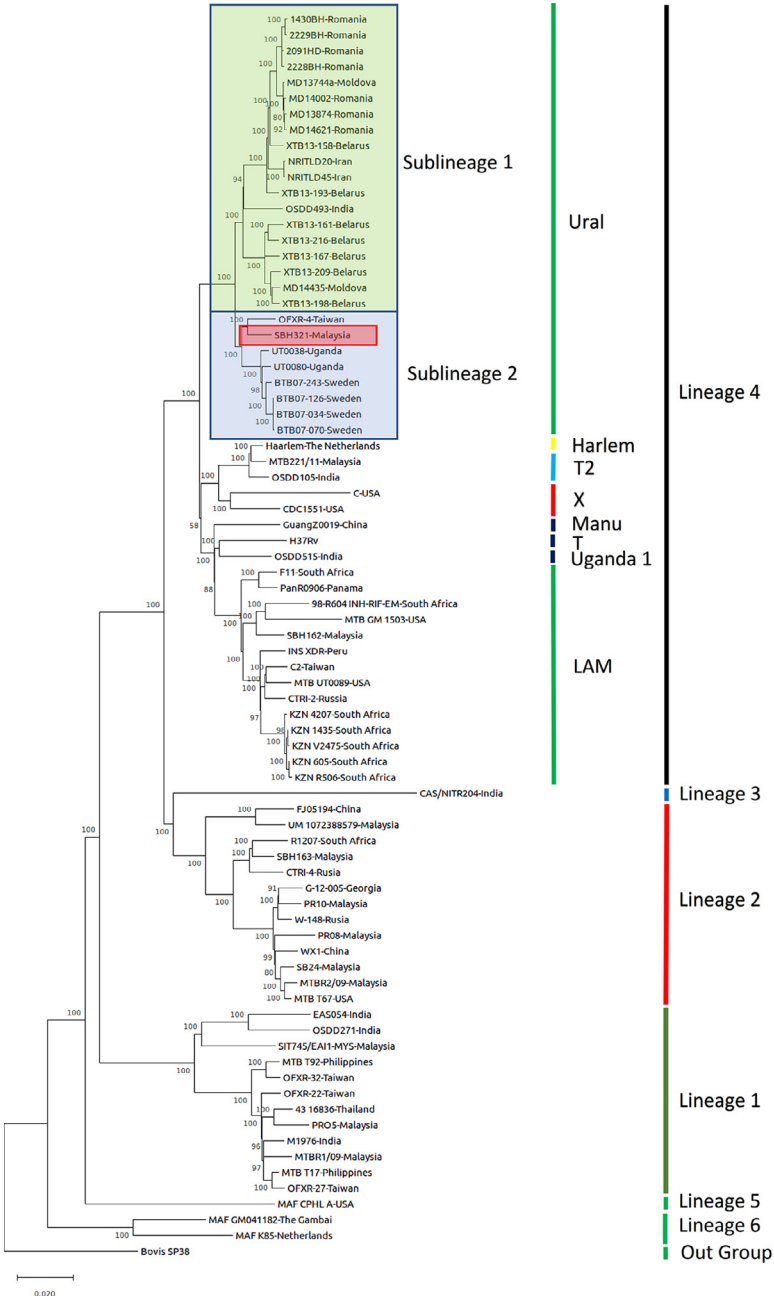


**Fig. 1.** The phylogenetic tree shows that *Mycobacterium tuberculosis* strain SBH321 belongs to sublineage 2 of the Ural family in Lineage 4. The phylogenetic tree is inferred using the maximum likelihood method and General Time Reversible model by using MEGA X. The tree is rooted with Mycobacterium bovis SP38 as an outgroup.

version 0.7.1231 [7] in SAM-BAM format. In order to convert the format into readable sequences and sort the alignments, Samtools version 0.1.1932 [8] was used. Next, Genome Analysis Toolkit (GATK) version 3.4.033 [9] performed local realignment of the sequence reads and generated the reports on variant calling analysis of the *M. tuberculosis* strain SBH321. SnpEff version 4.134 [10] was utilized for the annotation of single nucleotide polymorphism (SNP).

### 2.5. SNP-based phylogenetic genotype data of SBH321

The entire SNP matrix used in the phylogenetic analysis was performed by the Maximum Likelihood method using MEGA (Molecular Evolutionary Genetic Analysis) X [11] after aligning the nucleotide sequences using CLUSTALW [11]. The significance of the branching patterns was evaluated by bootstrap analysis of 1000 replicates. The whole genome sequence of 78 strains of *M. tuberculosis* were extracted from GenBank and were used in phylogenetic analysis [12–14] which showed that our strain belonged to the Ural family of Europe-America-Africa lineage (Lineage 4) and clustered with ofloxacin-resistant *M. tuberculosis* strain OFXR-4 from Taiwan [15].

### 2.6. Nucleotide sequence accession number

The whole genome sequence has been deposited at DDBJ/ENA/GenBank under the accession number WCJH00000000.

## Ethics Statement

This data was approved by the Ethics Committee of the Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah [JKEtika 2/16 (6)].

## Declaration of Competing Interest

No competing interest.

## Acknowledgment

This work was supported by the Trans-Disciplinary Research Grant Scheme (TRGS) (Grant No. TRG009-2016) from the Ministry of Education Malaysia and UMSGreat Research Grant (Grant No. GUG0288-2/2018) from Universiti Malaysia Sabah.

## References

[1] D. Brites, S. Gagneux, The nature and evolution of genomic diversity in the Mycobacterium tuberculosis complex. Strain variation in the *Mycobacterium Tuberculosis* complex: its role in biology, Adv. Exp. Med. Biol. 1019 (2017) 1–26, doi:10.1007/978-3-319-64371-7_1.

[2] I. Mokrousov, The quiet and controversial: Ural family of *Mycobacterium tuberculosis*, Infect. Genet. Evol. 12 (2012) 619–629, doi:10.1016/j.meegid.2011.09.026.

[3] V. Sinkov, O. Ogarkov, I. Mokrousov, B. Igor, Z. Yuri, H. Svetlana, K. Scott, New epidemic cluster of pre-extensively drug resistant isolates of *Mycobacterium tuberculosis* Ural family emerging in Eastern Europe, BMC Genom. 19 (2018) 762, doi:10.1186/s12864-018-5162-3.

[4] M.M.D. Goroh, G.S. Rajahram, R. Avoi, C.H.A.V.D. Boogaard, T. William, A.P. Ralph, C. Lowbridge, Epidemiology of tuberculosis in Sabah, Malaysia, 2012-2018, Infect. Dis. Poverty 9 (2020) 119, doi:10.1186/s40249-020-00739-7.

[5] A. Bankevich, S. Nurk, D. Antipov, A.A Gurevich, M. Dvorkin, A.S Kulikov, V.M. Lesin, S.I Nikolenko, S. Pham, A.D Prjibelski, A.V Pyshkin, A.V Sirotkin, N. Vyahhi, G. Tesler, M.A Alekseyev, P.A Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477, doi:10.1089/cmb.2012.0021.

[6] T. Tatusova, M. DiCuccio., A. Badretdin, V. Chetvernin, E.P Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M. Borodovsky, J. Ostell, NCBI prokaryotic genome annotation pipeline, Nucleic Acids Res. 44 (2016) 6614–6624, doi:10.1093/nar/gkw569.

[7] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760, doi:10.1093/bioinformatics/btp324.

[8] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, doi:10.1093/bioinformatics/btp352.

[9] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303, doi:10.1101/gr.107524.110.

[10] P. Cingolani, A. Platts, L.L Wang, M. Coon, T. Nguyen, L. Wang, J.L Susan, X. Lu, D.M Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff, Fly 6 (2012) 80–92, doi:10.4161/fly.19695.

[11] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 6 (2018) 1547–1549, doi:10.1093/molbev/msy096.

[12] J. Chong, S.M. Yew, Y.-C. Tan, K.P. Ng, Y.F. Toh, J.-S. Khoo, W.-Y. Yee, Genome analysis of the first extensively drug-resistant (XDR) *Mycobacterium tuberculosis* in Malaysia provides insights into the genetic basis of its biology and drug resistance, PLoS One 10 (2015) e0131694, doi:10.1371/journal.pone.0131694.

[13] E. Natalya, M.V.Z. Mikheecheva, A.V. Melerzanov, V.N. Danilenko, A Nonsynonymous SNP catalog of *Mycobacterium*, Genome Biol. Evol. 9 (2017) 887–899, doi:10.1093/gbe/evx053.

[14] Y. Blouin, Y. Hauck, C. Soler, M. Fabre, R. Vong, P. Massoure, E. Garnotel, Significance of the identification in the horn of Africa of an exceptionally deep branching *Mycobacterium tuberculosis* clade, PLoS One 7 (2012) e52841, doi:10.1371/journal.pone.0052841.

[15] D. Zhang, J.E Gomez, J-Y. Chien, N. Haseley, C.A Desjardins, A.M Earl, P-R Hsueh, DT Hung, Genomic analysis of the evolution of fluoroquinolone resistance in 2 *Mycobacterium tuberculosis* prior to tuberculosis diagnosis, Antimicrob. Agents Chemother. 60 (2016) 6600–6608, doi:10.1128/AAC.00664-16.