Data in brief 26 (2019) 104445

Contents lists available at ScienceDirect

Data in brief

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

Data Article

# Whole genome sequencing data and analysis of a rifampicin-resistant *Mycobacterium tuberculosis* strain SBH162 from Sabah, Malaysia

Jaeyres Jani [a], Zainal Arifin Mustapha [b], Norfazirah Binti Jamal [d], Cheronie Shely Stanis [d], Chin Kai Ling [c], Richard Avoi [e], Naing Oo Tha [e], Valentine Gantul [f], Daisuke Mori [d], Kamruddin Ahmed [a, d, *]

[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 Therapeutic, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*
[d] *Department of Pathobiology and Medical Diagnostics, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*
[e] *Department of Community and Family Medicine, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Sabah, Malaysia*
[f] *Tuberculosis and Leprosy Control Unit, Sabah State Health Department, Kota Kinabalu, Sabah, Malaysia*

## ARTICLE INFO

*Article history:*
Received 22 May 2019
Received in revised form 3 August 2019
Accepted 20 August 2019
Available online 28 August 2019

*Keywords:*
*M. tuberculosis*
Whole genome sequencing
Next generation sequencing
Rifampicin resistant
Sabah
Malaysia

## ABSTRACT

A *Mycobacterium tuberculosis* strain SBH162 was isolated from a 49-year-old male with pulmonary tuberculosis. GeneXpert MDR/RIF identified the strain as rifampicin-resistant *M. tuberculosis*. The whole genome sequencing was performed using Illumina HiSeq 4000 system to further investigate and verify the mutation sites of the strain through genetic analyses namely variant calling using bioinformatics tools. The *de novo* assembly of genome generated 100 contigs with N50 of 156,381bp. The whole genome size was 4,343,911 bp with G + C content of 65.58% and consisted of 4,306 predicted genes. The mutation site, S450L, for rifampicin resistance was detected in the *rpoB* gene. Based on the phylogenetic analysis using the Maximum Likelihood method, the strain was identified

* Corresponding author. Borneo Medical and Health Research Centre, 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.2019.104445

as belonging to the Europe America Africa lineage (Lineage 4). The genome dataset has been deposited at DDBJ/ENA/GenBank under the accession number SMOE00000000.

Specifications Table

| | |
|---|---|
| Subject area | Environmental Science |
| Specific subject area | Immunology and Microbiology |
| Type of data | Whole genome sequence with gene annotation and comparative genomic of *Mycobacterium tuberculosis* strain SBH162. The strain is also resistant to rifampicin drug. |
| Data acquisition | *De novo* whole genome sequencing, phylogenetic and variant calling with Illumina HiSeq 4000 system |
| Data format | Raw and analyzed data of whole genome sequences |
| Experimental factors | Isolated and cultured in 7H9 middlebrook medium, and incubated at BACTEC MGIT 320, Extraction of genomic DNA from a pure culture, library preparation for sequencing, Illumina sequencing, *de novo* assembly, annotation, variant calling and comparative genomic analyses |
| Experimental features | DNA extraction was performed using Masterpure Complete DNA and RNA purification kit; library was prepared using NEBNext® Ultra™ DNA Library Prep Kit for Illumina®; sequencing was performed using Illumina Hiseq 4000 system. The genome was assembled using SPAdes, variant calling by GATK tools, annotated with NCBI Prokaryotic Genome Annotation Pipeline and comparative genomic through kSNP3. |
| Data source location | Kota Kinabalu, Sabah, Malaysia |
| Data accessibility | Data is publicly available at NCBI Genbank from the following links: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA524470<br>https://www.ncbi.nlm.nih.gov/biosample/SAMN11026786<br>https://www.ncbi.nlm.nih.gov/nuccore/SMOE00000000 |

**Value of the data**

- The data will shed light on the molecular biology of a *Mycobacterium tuberculosis* strain, which will be beneficial to researchers working on tuberculosis.
- The data will give insight into drug resistance in *M. tuberculosis*, which will benefit clinicians and patients.
- The data will help to understand the relation between *M. tuberculosis* strains from Sabah and other areas, which will contribute to policy making for the control of tuberculosis.

## 1. Data

In this paper, we present the data and analysis of the whole genome sequence (WGS) of *M. tuberculosis* strain SBH162 from Sabah, Malaysia. Tuberculosis was newly detected in a 49-year-old male patient using GeneXpert MDR/RIF. The whole genome was sequenced and *de novo* assembly, variant calling and comparative genomic of strain were performed. The *de novo* assembly of genome generated 100 contigs with N50 of 156,381bp. The whole genome size was 4,343,911 bp with G + C content of 65.58% and consisted of 4,306 predicted genes. In addition, the variant calling verified the mutation site in the *rpoB* gene, locus S450L. Based on the comparative genomics analysis using WGS of 77 strains, we determined that our strain belongs to the LAM family of Lineage 4 and is similar to the strains from South Africa [9] and Gambia [10] (see Fig. 1).

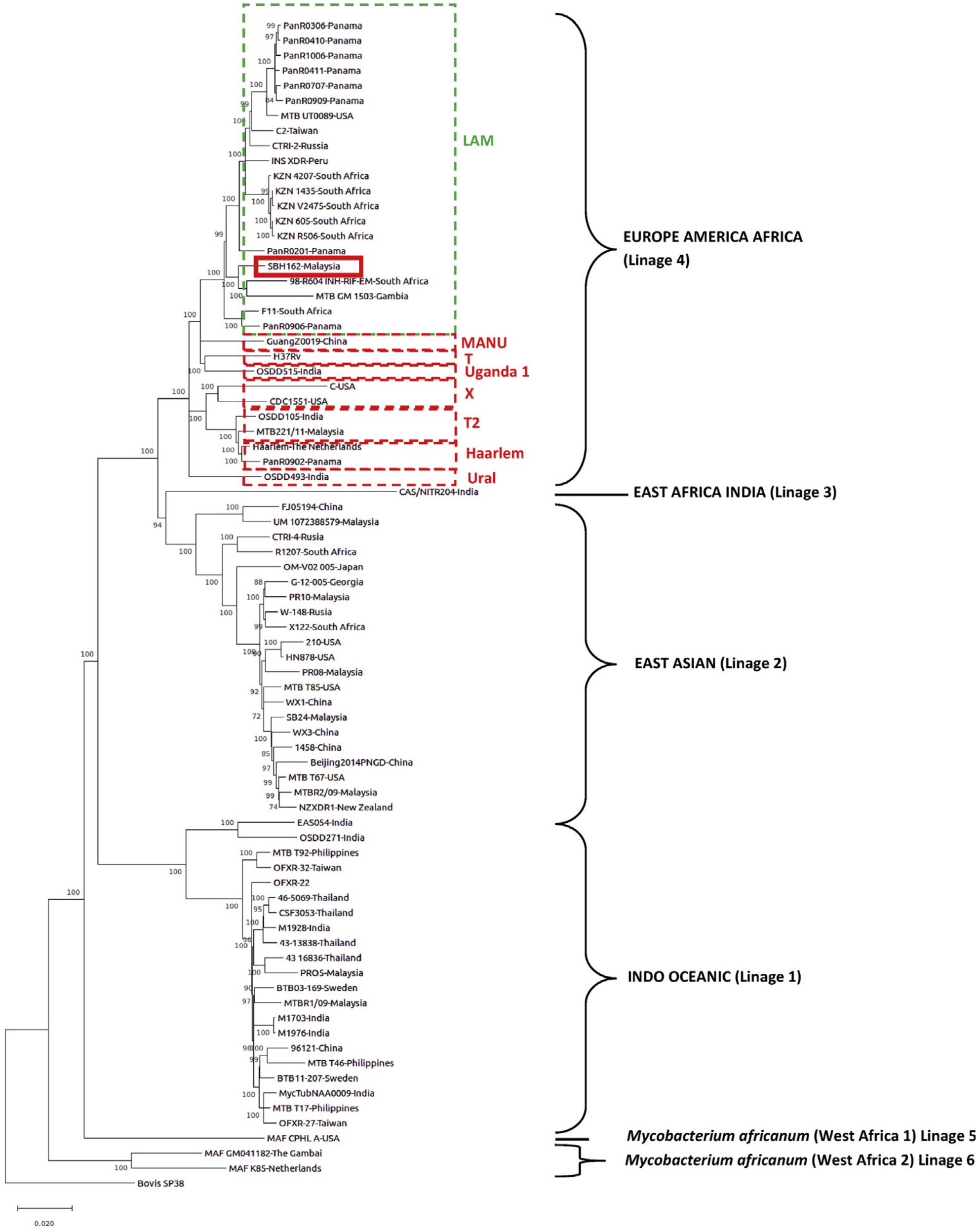

**Fig. 1.** Comparative phylogenetic analysis of strain SBH162. This strain belongs to Lineage 4 and is clustered with other strains from the LAM family. The Malaysian strains are also in Lineage 4 and belong to T2 family while other Malaysian strains belong to Lineages 1 and 2. The phylogenetic tree was constructed using SNPs data extracted from the genome sequence. The phylogenetic tree was inferred using the Maximum Likelihood method and General Time Reversible model. The tree is rooted with *M. bovis* SP38 as outgroup.

## 2. Experimental design, materials and methods

### 2.1. Isolation, culture, DNA extraction, library preparation and sequencing

The *M. tuberculosis* strain SBH162 was isolated from the sputum of a 49-year-old male from Kota Kinabalu, Sabah, Malaysia, who was newly diagnosed with tuberculosis in April 2017. The sputum was analyzed using GeneXpert MDR/RIF and cultured in 7H9 middlebrook medium using BACTEC MGIT 320 (Becton-Dickinson, Oxford, United Kingdom). Genomic DNA was extracted using Masterpure Complete DNA and RNA purification kit (Epicenter, Inc., Madison, Wisconsin, USA) according to the manufacturer's instructions. The quality of the extracted DNA was determined by Nanodrop 2000c spectrophotometer (ThermoFisher Scientific, USA). In addition, the concentration was determined using Qubit® 2.0 fluorometer (Invitrogen, ThermoFisher Scientific, USA).

### 2.2. *Quality trim,* de novo *assembly and annotation*

The genome was sequenced until 99% completion using 332X sequencing coverage. A total of 9,773,850 paired reads (~1GB) of a 300-bp insert-size library by NEBnext Ultra kit (Illumina, San Diego, CA) were generated from Illumina HiSeq 4000. The data sequence was deposited in the Sequence Read Archive (SRA) (biosample accession number SAMN11026786) under the bioproject accession number PRJNA524470. For the purpose of analysis, the quality of the sequence read was checked using FastQC. All of the raw reads were pre-processed using BBMap version 38.43 tools [1], whereby the adapters were trimmed and the reads with less than 50bp were removed, based on the phred with a quality below Q30 using BBDuk.sh [1]. *De novo* assembly was performed using SPAdes version 3.11.1 [2]. The generated contigs were annotated using NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [3].

### 2.3. Assembly statistic

| | |
|---|---|
| Sequencing depth | 332X |
| Total length of sequences (bp) | 4,343,911 |
| Total number of contigs | 100 |
| N50 (bp) | 156,381 |
| GC (%) | 65.58 |
| CDSs | 4,306 |
| tRNAs | 45 |
| 5s,16s,23s rRNA | 1, 1, 1 |

Supplementary data 1, 2 and 3.

.

## 3. Variant calling

In the variant calling, sequence reads were trimmed with a phred score above Q20. Reads shorter than 50bp and possible contaminating adaptor sequences were excluded using BBMap version 38.43 tools [1]. Paired-end raw reads were mapped to the *M. tuberculosis* H37Rv reference genome (GenBank accession number NC_000962.3) using BWA MEM version 0.7.1231 [4]. Samtools version 0.1.1932 [5] was used to convert the SAM-BAM format and to sort the mapped sequences. Local realignment of the mapped reads was performed using GATK version 3.4.033 [6]. The statistic reports for the variant calling were generated using GATK and Samtools, whereby the average mapping rate of the sequences was 99.47% to the reference genome. Variant sites were filtered based on the following criteria: mapping quality greater than 50bp; base quality or base alignment quality greater than 20bp; and more than 10 covering each site. The SnpEff version 4.134 [7] was used for single nucleotide polymorphism (SNP) annotation. The list of SNPs (novel and previously reported) is provided as supplementary data 4.

This study was approved by the ethics committee at the Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah (JKEtika 2/16 (6)).

## 4. SNP-based phylogenetic genotype study of SBH162

The genotype of our isolate was determined by the whole genome SNP. We identified that SBH162 belongs to Lineage 4 (LAM family) of the *M. tuberculosis* complex, where the sample was clustered with *M. tuberculosis* 98-R604 INH-RIF-EM and GM 1053 [8,10,14]. A mutation, S450L, was detected in the *rpoB* gene of our strain, which is responsible for resistant to rifampicin [13,15]. Strain 98-R604 is from South Africa [9] and is resistant to isoniazid, rifampicin and ethambutol. On the other hand strain MTB GM1503 is from Gambia [14], is not rifampicin resistant *M. tuberculosis*.

Core-SNP was identified using kSNP3 package [11]. The entire SNP matrix was used in the phylogenetic analysis, which was performed with the Maximum Likelihood method using MEGA (Molecular Evolutionary Genetic Analysis) Software 6.0 [12] after aligning the nucleotide sequences using CLUSTAL W [12]. The significance of the branching patterns was evaluated through bootstrap analysis of 1,000 replicates. The whole genome sequence of 77 strains of *M. tuberculosis* were extracted from GenBank and used in the phylogenetic analysis [9,10,14].

## 5. Nucleotide sequence accession number

The whole genome sequence has been deposited at DDBJ/ENA/GenBank under the accession number SMOE00000000.

## Acknowledgments

This work was supported by the Trans-Disciplinary Research Grant Scheme (TRGS) (grant number TRG009-2016) under the Ministry of Education Malaysia.

## Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104445.

## References

[1] B. Bushnell, J. Rood, E. Singer, BBMerge – accurate paired shotgun read merging via overlap, PLoS One 12 (2017) 1–15. https://doi.org/10.1371/journal.pone.0185056.

[2] 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). https://doi.org/10.1089/cmb.2012.0021.

[3] 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. https://doi.org/10.1093/nar/gkw569.

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

[5] H. Li, The sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

[6] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, A. Mark, DePristo, the genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303. https://doi.org/10.1093/bioinformatics/btp352.

[7] P. Cingolani, A. Platts, le L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3, Fly 6 (2012) 80–92. https://doi.org/10.4161/fly.19695.

[8] S. Feuerriegel, C. Ko, L. Tru, J. Archer, S. Ru, E. Richter, S. Niemann, Thr202Ala in thyA is a marker for the Latin American Mediterranean Lineage of the *Mycobacterium tuberculosis* complex rather than para-aminosalicylic acid resistance, Antimicrob. Agents Chemother. 54 (2010) 4794–4798. https://doi.org/10.1128/AAC.00738-10.

[9] 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). https://doi.org/10.1371/journal.pone.0131694.

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

[11] S.N. Gardner, T. Slezak, B.G. Hall, kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome, Bioinformatics 31 (2015) 2877–2878. https://doi.org/10.1093/bioinformatics/btv271.

[12] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729. https://doi.org/10.1093/molbev/mst197.

[13] E. Andre, L. Goeminne, A. Cabibbe, P. Beckert, B.K. Mukadi, V. Mathys, E. Cambau, Consensus numbering system for the rifampicin resistance-associated *rpoB* gene mutations in pathogenic mycobacteria, Clin. Microbiol. Infect. 23 (2017) 167–172. https://doi.org/10.1016/j.cmi.2016.09.006.

[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. 7 (2012). https://doi.org/10.1371/journal.pone.0052841.

[15] S.M. Gygli, S. Borrell, A. Trauner, S. Gagneux, Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives, FEMS Microbiol. Rev. 41 (2017) 354–373. https://doi.org/10.1093/femsre/fux011.