Contents lists available at ScienceDirect

Data in Brief

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

Data Article

First whole genome sequencing data of a *Mycobacterium tuberculosis* STB-T1A strain isolated from a spinal tuberculosis patient in Sabah, Malaysia

Kai Ling Chin[a,b,*], Eraniyah Jastan Suing[a], Ruhini Andong[a], Choong Hoon Foo[c], Sook Kwan Chan[c], Jaeyres Jani[b], Kamruddin Ahmed[b,d], Zainal Arifin Mustapha[e]

[a] *Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*
[b] *Borneo Medical and Health Research Centre, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*
[c] *Department of Orthopaedics, Queen Elizabeth Hospital, Ministry of Health Malaysia, Kota Kinabalu, Sabah, Malaysia*
[d] *Department of Pathology and Microbiology, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*
[e] *Department of Medical Education, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*

ARTICLE INFO

Article history:
Received 9 April 2024
Accepted 22 April 2024
Available online 25 April 2024

Dataset link: Mycobacterium tuberculosis strain STB-T1A, whole genome shotgun sequencing project (Original data)

Keywords:
Pott's disease
Spine
Mycobacterium tuberculosis
Next generation sequencing
Sabah
Malaysia

ABSTRACT

Spinal tuberculosis, also referred to as Pott's disease, presents a significant risk of severe paralysis if not promptly detected and treated, owing to complications such as spinal cord compression and deformity. This article presents the genetic analysis of a *Mycobacterium tuberculosis* STB-T1A strain, isolated from the spine of a 29-year-old female diagnosed with spinal tuberculosis. Genomic DNA was extracted from pure culture and subjected to sequencing using the Illumina NovaSeq 6000 sequencing system. The genome of the *M. tuberculosis* STB-T1A strain spans 4,367,616 base pairs with a G+C content of 65.56 % and 4174 protein-coding genes. Comparative genomic analysis, conducted via single nucleotide polymorphism (SNP)-based phylogenetic analysis using the

* Corresponding author at: Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia.

E-mail address: chinkl@ums.edu.my (K.L. Chin).

https://doi.org/10.1016/j.dib.2024.110476

Maximum Likelihood method, revealed that the strain falls within the Indo-Oceanic lineage (Lineage 1). It clusters with the *M. tuberculosis* 43-16836 strain, which was isolated from the cerebrospinal fluid of a patient with tuberculous meningitis in Thailand. The complete genome sequence has been deposited at the National Center for Biotechnology Information (NCBI) GenBank database with the accession number JBBMVZ000000000.

Specifications Table

Subject	Health and Medical Sciences
Specific subject area	Infectious diseases
Type of data	Raw data, whole genome sequencing, gene annotation, variant calling, and comparative genomic analysis of a *Mycobacterium tuberculosis* strain
Data collection	Bone tissue was taken from the spine of a patient suspected to have spinal tuberculosis, and it was tested using Xpert® MTB/RIF Ultra for tuberculosis diagnosis. Subsequently, the sample was cultured, bacterial genomic DNA was extracted, and whole-genome sequencing was performed. The raw sequencing data was then utilized for *de novo* assembly and phylogenetic analysis.
Data source location	Queen Elizabeth Hospital, Kota Kinabalu, Sabah
Data accessibility	Repository name: National Center for Biotechnology Information (NCBI) Data identification number: BioProject: PRJNA1091826, BioSample: SAMN40613452, Sequence Read Archive (SRA): SRR28465663, GenBank: JBBMVZ000000000 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1091826 https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN40613452 https://www.ncbi.nlm.nih.gov/sra/?term=SRR28465663 https://www.ncbi.nlm.nih.gov/nuccore/JBBMVZ000000000

1. Value of the Data

- This is the first report of a *M. tuberculosis* strain isolated from a spinal tuberculosis patient in Sabah, Malaysia, and its whole-genome sequence could provide fundamental insights into its microbial activities, facilitating a deeper understanding of its characteristics.
- The data are crucial for comprehending the genetic characteristics of the *M. tuberculosis* strain by providing detailed information about gene content, function, and genomic organization.
- The data are important for gaining crucial insights into the genetic diversity and evolutionary dynamics of *M. tuberculosis* strains from Sabah and other regions, facilitating the understanding of transmission patterns across geographical areas.

2. Background

A 29-year-old female presented with clinical symptoms indicative of spinal tuberculosis, including gibbous deformity, cold abscess, paradiscal lesion, anterior vertebral loss, narrowed disc space, and paravertebral shadows. She also exhibited tuberculosis (TB)-related symptoms such as loss of appetite, weight loss, and malnutrition, with a body mass index (BMI) below 18.5, a high-risk factor for TB infection. A bone tissue sample was obtained from the spine and the patient was diagnosed with tuberculosis using Xpert® MTB/RIF Ultra. The bacterial isolate was obtained using the BD BACTEC™ MGIT™ culture system. Bacterial DNA was extracted and whole genome sequencing (WGS) was conducted.

3. Data Description

This article presents the data analysis of the WGS of *M. tuberculosis* STB-T1A strain from Sabah, Malaysia. A total of 18,097,866 paired reads at 150 bp read length were generated from the Illumina NovaSeq 6000 sequencing system, with a sequencing coverage of 615X. *De novo* assembly of the genome generated 146 contigs with N50 of 161,185 bp and the largest contig observed was 303,749 bp. The whole genome size was 4,367,616 bases with G+C content of 65.56 %. The genetic makeup comprises 4174 coding sequences (CDS), 45 tRNAs, one 5S, one 16S, and one 23S rRNAs, and three ncRNAs. Statistical report of variant calling showed that 99.58 % of the reads were mapped to the *M. tuberculosis* H37Rv reference genome. Within this dataset, 2193 single nucleotide polymorphisms (SNPs), 192 insertions, and 166 deletions were identified. Comparative genomic analysis with *M. tuberculosis* strains from different lineages revealed that the *M. tuberculosis* STB-T1A strain belongs to the Indo-Oceanic lineage (Lineage 1) and has similar characteristics with the *M. tuberculosis* 43-16836 isolated from a tuberculous meningitis patient in Thailand [1] (Fig. 1). The *M. tuberculosis* STB-T1A strain is predicted to be drug susceptible based on analysis using the Mykrobe software.

4. Experimental Design, Materials and Methods

4.1. Sample Collection and Tuberculosis Detection

A 29-year-old female presented with clinical symptoms indicative of spinal tuberculosis at Queen Elizabeth Hospital in Kota Kinabalu, Sabah. Bone tissue sample was collected via biopsy method from the spine by an orthopedic surgeon. The sample was subjected to tuberculosis (TB) detection with Xpert® MTB/RIF Ultra (Cepheid, Sunnyvale, CA, USA) following the manufacturer’s protocol. The processed sample was transferred to a cartridge and inserted into a GeneXpert machine for automated DNA extraction and real-time polymerase chain reaction (qPCR) for qualitative detection of *Mycobacterium tuberculosis* Complex (MTBC) and rifampicin (RIF) resistance [2]. Based on the cycle threshold (Ct) value, the semi-quantitative bacterial load was reported.

4.2. Bacterial Culture and DNA Extraction

The bone tissue was decontaminated with BBL™ MycoPrep™ (Becton, Dickinson, NJ, USA). The processed sample was cultured in a Mycobacterium Growth Indicator Tube (MGIT) tube containing 7H9 Middlebrook broth with PANTA (polymyxin-B, Amphotericin-B, nalidixic acid, trimethoprim, azilocillin) antibiotic and OADC (oleic acid, albumin, dextrose, catalase) supplement mixture. The tube was loaded into the BD BACTEC™ MGIT™ 320 system (Becton, Dickinson, NJ, USA), and incubated at 37 °C until bacterial growth was detected by the system [2]. DNA was extracted using the Masterpure™ Complete DNA and RNA Purification kit (Epicentre Biotechnologies, Madison, WI, USA) according to the manufacturer’s instruction, with an extended lysis protocol for 16 h with Proteinase K. The quality of the extracted DNA was determined by Nanodrop 2000c spectrophotometer (ThermoFisher Scientific, USA) and gel electrophoresis [3].

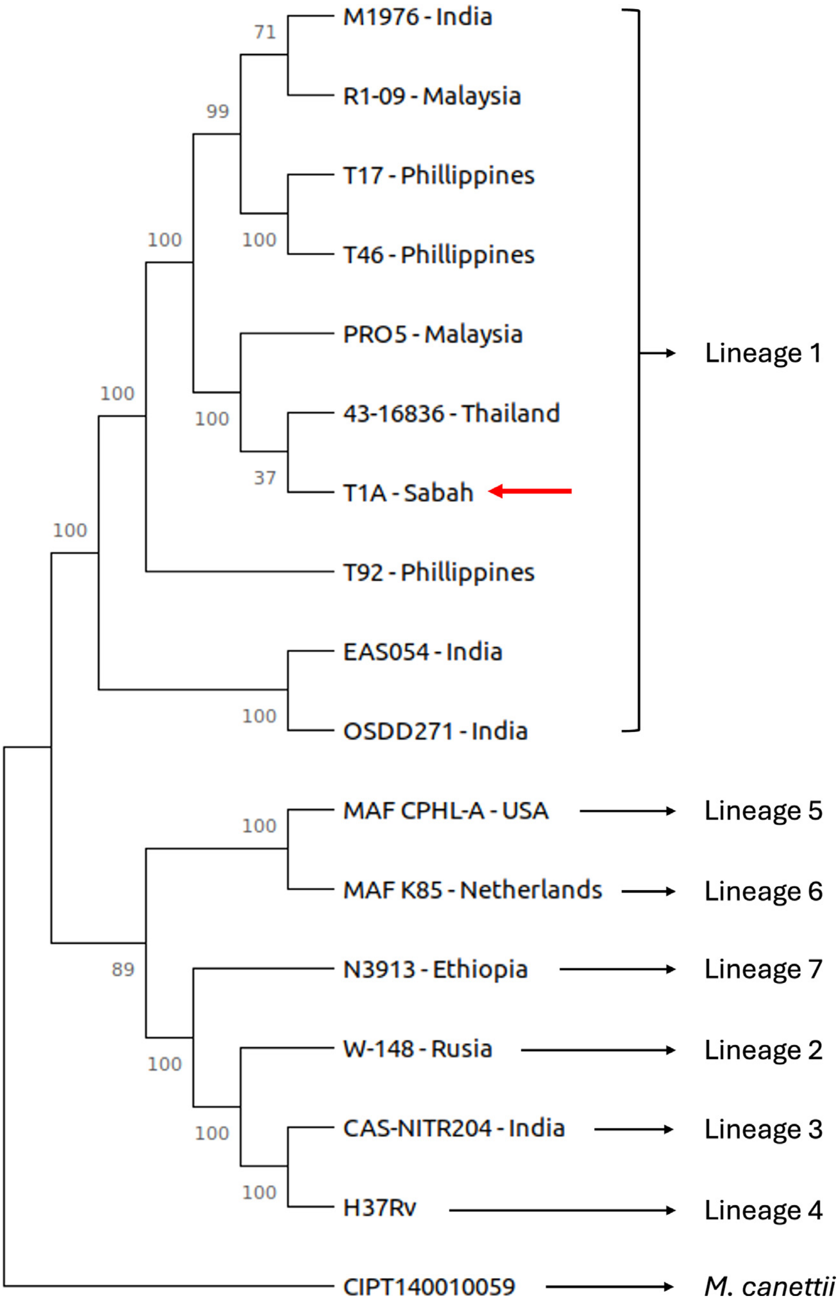

Fig. 1. Comparative phylogenetic analysis of *M. tuberculosis* STB-T1A strain (red arrow). This strain belongs to Lineage 1 and clusters with the *M. tuberculosis* 43-16836 strain from Thailand. The phylogenetic tree was constructed using SNP data, utilizing the Maximum Likelihood method and the General Time Reversible model. The tree was rooted with *M. canettii* serving as the outgroup.

4.3. Whole Genome Sequencing and Data cleaning

The genomic DNA was sent to Apical Scientific Sdn. Bhd., Malaysia for library preparation, followed by whole genome sequencing by Illumina NovaSeq 6000 platform. The sequencing data has been submitted to the National Center for Biotechnology Information (NCBI) and can be accessed under the following accession numbers: BioProject PRJNA1091826, BioSample SAMN40613452, Sequence Read Archive (SRA) SRR28465663, and GenBank JBBMVZ000000000.

The output of the sequencing was in FastQ format file. FastQC version 0.12.1 was used for assessing the quality of raw sequencing reads [4], and fastp version 0.23.4 was used for trimming adapter sequences and filtering out reads with less than 50 bp [5].

4.4. De novo *Assembly, Variant Calling, and Phylogenetic Analysis*

The *de novo* assembly process began with KmerGenie version 1.7051 to determine the optimal k-mer for assembly, utilizing the processed reads [6]. Subsequently, a draft genome was generated using SPAdes version 3.15.4 to assemble the processed reads into contigs [7]. Following assembly, the quality of the resulting contigs was assessed with QUAST version 5.0.2 [8]. Finally, functional annotation of the assembled contigs was performed using NCBI Prokaryotic Genome Annotation Pipeline (PGAP) to identify genes and annotate their functions [9].

The variant calling process began by aligning the processed reads with the *M. tuberculosis* H37Rv reference genome (GenBank accession number: NC_000962.3) using Burrows-Wheeler Aligner (BWA) version 0.7.17 [10]. The mapped reads in Sequence Alignment/Mat (SAM) format were converted to Binary Alignment Map (BAM) format and sorted using Samtools version 1.19.2 [11]. Following alignment, variant calling was performed using Genome Analysis Toolkit (GATK), which employs HaplotypeCaller to identify differences (variants) between the sample genome and the reference genome, including SNPs, insertions, deletions, and other genomic variations [12]. After initial variant calling, BCFtools version 1.19 was used for further refining the variants [11]. The functional effects of the variants were annotated using SnpEff version 5.0 to gain insights into the potential functional consequences of these genetic alterations on genes [13].

kSNP3 was used to detect SNPs and obtain a SNP matrix representing genetic variations among strains [14], including the draft genome of *M. tuberculosis* STB-T1A strain generated by SPAdes, and whole genome sequences from other Lineages obtained from NCBI GenBank, i.e., L1: *M. tuberculosis* T92 (NZ_JLDA00000000.1), *M. tuberculosis* MTBR1/09 (LATN00000000.1), *M. tuberculosis* T46 (ACHO00000000.1), *M. tuberculosis* EAI/OSDD271 (AQQC00000000.1), *M. tuberculosis* T17 (JLCV00000000.1), *M. tuberculosis* 43-16836 (ATNF00000000.1), *M. tuberculosis* PR05 (AOMG00000000.2), *M. tuberculosis* M1976 (KK331618.1), and *M. tuberculosis* EAS054 (ABOV00000000.1); L2: *M. tuberculosis* W-148 (NZ_CP012090.1); L3: *M. tuberculosis* CAS/NITR204 (CP005386.1); L4: *M. tuberculosis* H37Rv (NC_000962.3); L5: *M. africanum* CPHL_A (ACHP00000000.1); L6: *M. africanum* K85 (ACHQ00000000.1); L7: *M. tuberculosis* N3913 (NZ_CP069063.1); and *M. canettii* CIPT 140010059 (NC_015848.1). The resulting SNP matrix was used for downstream phylogenetic analysis with Molecular Evolutionary Genetics Analysis version 11 (MEGA 11) [15]. After alignment of the nucleotide sequences using ClustalW, the most appropriate evolutionary model for the dataset was predicted and Maximum Likelihood analysis with bootstrapping (1000 replicates) was performed to infer the phylogenetic relationships among the strains. The drug susceptibility of the strain was predicted using Mykrobe Predictor TB version 0.1.0, utilizing raw sequencing reads as the input data [16].

Limitations

None.

Ethics Statement

The ethics approval for this study was obtained from the National Medical Research Register (NMRR) and the Medical Research Ethics Committee (MREC) (NMRR ID-22-02464-T2O). Informed consent for sample collection was obtained from the participant. The authors kept the ethical concerns into consideration when gathering data and ensured that the information obtained from the respondent was only utilized for research purposes.

Data Availability

Mycobacterium tuberculosis strain STB-T1A, whole genome shotgun sequencing project (Original data) (NCBI).

CRediT Author Statement

Kai Ling Chin: Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Investigation, Data curation, Supervision; **Eraniyah Jastan Suing:** Visualization, Investigation, Data curation; **Ruhini Andong:** Visualization, Investigation, Data curation; **Choong Hoon Foo:** Conceptualization, Supervision, Data curation; **Sook Kwan Chan:** Data curation; **Jaeyres Jani:** Visualization, Investigation, Data curation; **Kamruddin Ahmed:** Conceptualization, Supervision; **Zainal Arifin Mustapha:** Conceptualization, Supervision, Funding acquisition.

Acknowledgments

The authors acknowledge the Ministry of Higher Education (MOHE) for funding under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2022/SKK12/UMS/01/2).

Declaration of Competing 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.

References

[1] W. Viratyosin, et al., Draft genome sequence of the *Mycobacterium tuberculosis* strain 43-16836, belonging to the indo-oceanic lineage, isolated from tuberculous meningitis in Thailand, Genome Announc. 1 (5) (2013), doi:10.1128/genomea.00801-13.

[2] Z. Li, et al., Evaluation of different diagnostic methods for spinal tuberculosis infection, BMC Infect. Dis. 23 (1) (2023) 695.

[3] J. Jani, et al., The whole genome sequence data analyses of a *Mycobacterium tuberculosis* strain SBH321 isolated in Sabah, Malaysia, belongs to Ural family of Lineage 4, Data Br. 33 (2020) 106388.

[4] S. Andrews, FastQC: a quality control tool for high throughput sequence data. Available online at: https://qubeshub.org/resources/fastqc (2010).

[5] S. Chen, et al., fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) i884–i890.

[6] R. Chikhi, P. Medvedev, Informed and automated k-mer size selection for genome assembly, Bioinformatics 30 (1) (2013) 31–37.

[7] A. Prjibelski, et al., Using SPAdes De novo assembler, Curr. Protoc. Bioinform. 70 (1) (2020) e102.

[8] A. Mikheenko, et al., Versatile genome assembly evaluation with QUAST-LG, Bioinformatics 34 (13) (2018) i142–i150.

[9] T. Tatusova, et al., NCBI prokaryotic genome annotation pipeline, Nucleic Acids Res. 44 (14) (2016) 6614–6624.

[10] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.

[11] P. Danecek, et al., Twelve years of SAMtools and BCFtools, Gigascience 10 (2) (2021) 1–4.

[12] M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (5) (2011) 491–498.

[13] P. Cingolani, et al., 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 (2) (2012) 80–92 (Austin).

[14] 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 (17) (2015) 2877–2878.

[15] K. Tamura, G. Stecher, S. Kumar, MEGA11: molecular evolutionary genetics analysis version 11, Mol. Biol. Evol. 38 (7) (2021) 3022–3027.

[16] M. Hunt, et al., Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe, Wellcome Open. Res. 4 (2019) 191.