Genomics Data 7 (2016) 119–120

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Genome Sequencing and Annotation of *Mycobacterium tuberculosis* PR08 strain

Mohammad Maaruf Jaafar [a], Mohd Zakihalani A. Halim [a], Mohamad Izwan Ismail [a], Lee Lian Shien [a], Teh Lay Kek [a,b], Ngeow Yun Fong [c], Norazmi Mohd Nor [d,e], Zainul Fadziruddin Zainuddin [d], Tang Thean Hock [f], Mohd Nazalan Mohd Najimudin [g], Mohd Zaki Salleh [a,*]

[a] Integrative Pharmacogenomics Institute (iPROMISE), Universiti Teknologi MARA (UiTM), Malaysia
[b] Faculty of Pharmacy, Universiti Teknologi MARA (UiTM), Malaysia
[c] Faculty of Medicine, University of Malaya, Malaysia
[d] School of Health Sciences, Health Campus, Universiti Sains Malaysia, Malaysia
[e] Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia, Malaysia
[f] Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia, Malaysia
[g] School of Biological Sciences, Universiti Sains Malaysia, Malaysia

ARTICLE INFO

Article history:
Received 26 November 2015
Received in revised form 30 December 2015
Accepted 30 December 2015
Available online 31 December 2015

Keywords:
Mycobacterium tuberculosis
Genome
Extrapulmonary
CSF

ABSTRACT

Mycobacterium tuberculosis is an acid fast bacterial species in the family Mycobacteriaceae and is the causative agent of most cases of tuberculosis. Here, we report the genomic features of *Mycobacterium tuberculosis* isolated from the cerebrospinal fluid (CSF) of a patient diagnosed with both pulmonary and extrapulmonary tuberculosis (TB). The isolated strain was identified as *Mycobacterium tuberculosis* PR08 (MTB PR08). Genomic DNA of the MTB PR08 strain was extracted and subjected to whole genome sequencing using MiSeq (Illumina, CA,USA). The draft genome size of MTB PR08 strain is 4,292,364 bp with a G + C content of 65.2%. This strain was annotated to have 4723 genes and 48 RNAs. This whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number CP010895.

Specifications	
Organism/cell line/tissue	*Mycobacterium tuberculosis*
Strain	PR08
Sequencer or array type	Illumina MiSeq sequencer
Data format	Processed
Experimental factors	Microbial strain
Experimental features	Draft genome sequence of *Mycobacterium tuberculosis* PR08 assembly and annotation
Consent	N/A
Sample source location	Kuala Lumpur, Malaysia 4.1936°N 103.7249°E

1. Direct link to deposited data [provide URL below]

http://www.ncbi.nlm.nih.gov/bioproject/PRJNA196391. (Biosample: SAMN03290698).

2. Experimental design, materials and methods

Mycobacterium tuberculosis PR08 (MTB PR08) was isolated from the cerebrospinal fluid (CSF) of a patient diagnosed with both pulmonary and extrapulmonary tuberculosis at a local hospital. The sample was cultured in BBL™ MGIT™ Mycobacterial Growth Indicator Tube supplemented with BBL™ MGIT™ OADC enrichment and BBL™ MGIT™ PANTA™ antibiotic mixture (Becton–Dickinson, Oxford, United Kingdom).

Genomic DNA was extracted from MTB PR08 and was sequenced using MiSeq (Illumina, CA, USA), generating a total of 46,013,686 reads in a 300-cycle run. Raw reads were trimmed and assembled *de novo* using CLCbio (CLC Genomics Workbench version 7.0.3) (CLCbio, Aarhus, Denmark), producing an average coverage of 378×. Annotation was performed using the Bacterial Annotation System (BASys) [1] and Rapid Annotation using Subsystem Technology (RAST) [2] online services, and the pathogenicity and virulence genes were determined. The genes were validated using the following external gene annotation databases: TubercuList (http://tuberculist.epfl.ch), UniProtKB (http://www.ebi.ac.uk/uniprot), Virulence Factor Database (VFDB) (http://www.mgc.ac.cn), and TBDatabase (TBDB) (http://www.tbdb.org).

* Corresponding author.

http://dx.doi.org/10.1016/j.gdata.2015.12.030

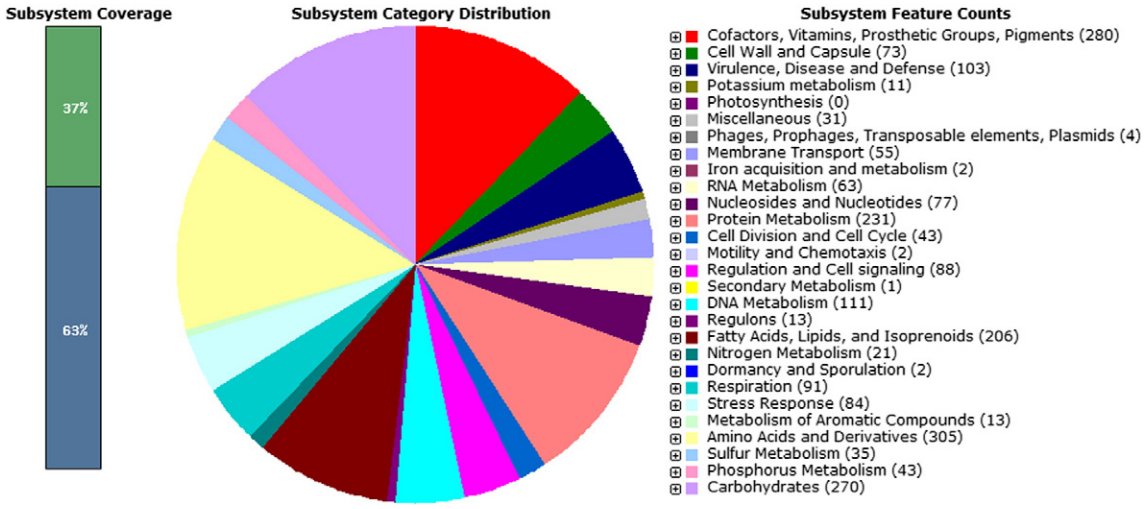

Fig 1. Subsystem distribution of *Mycobacterium tuberculosis* strain PR08 (based on RAST annotation server).

Table 1
Comparative analysis between MTB PR05, MTB PR08 and the reference genome H37Rv.

Genome	PR08	PR05	H37Rv (reference)
Genome size (bp)	4,292,364	4,419,501	4,411,532
Number of subsystems	393	403	390
Number of coding sequences	4203	4437	4360
Number of genes	4723	4739	4644
Number of RNAs	48	48	48

The size of the draft genome of MTB PR08 is 4,292,364 bp with a G + C content of 65.2%. It is composed of 214 contigs with 4723 predicted genes of which 4203 were protein coding genes and 48 RNA-encoding genes. A total of 2295 (54.6%) of the protein coding genes were assigned into the Cluster of Orthologous Group (COG) [2]. Using RAST, a total of 393 subsystems were annotated in the MTB PR08 genome (Fig. 1).

Comparative analysis of MTB PR08 was performed against two other genomes; PR05 [3] and the reference genome H37Rv. Annotation and comparative genomics analysis of MTB PR08 and the selected reference genomes were carried out using RAST as shown in Table 1. In order to identify the functions of the genes that contributed to extrapulmonary TB, the genes were annotated using BASys. Based on the analysis, a putative gene (opcA gene) which may have been involved in extrapulmonary infection was identified. It has been reported to play a role in meningococcal adhesion, invasion of epithelial and endothelial cells and in assembly of Glucose-6-Phosphate-Dehydrogenase (G6PD) [4,5].

Comparison of genome sequences using RAST revealed that the closest strains of MTB PR08 are *Mycobacterium tuberculosis* NCGM2209 (score 521), *Mycobacterium tuberculosis* UM 1072388579 (score 473) and *Mycobacterium tuberculosis* NA-A0008 (score 454).

This Whole Genome Shotgun project has been deposited at GenBank under the accession number CP010895.

Nucleotide sequence accession number

The whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number CP010895.

Conflict of interest

The authors declare that there is no conflict of interests with respect to the work published in this paper.

Acknowledgments

This study was supported by a grant from the Ministry of Education, Malaysia [600-RMI/LRGS 5/3 (6/2011)].

References

[1] G.H. Van Domselaar, P. Stothard, S. Shrivastava, J.A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner, D.S. Wishart, BASys: a web server for automated bacterial genome annotation. Nucleic Acids Res. 33 (Suppl. 2) (2005) W455–W459.
[2] R.K. Aziz, D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, F. Meyer, The RAST server: rapid annotations using subsystems technology. BMC Genomics 9 (1) (2008) 75.
[3] A. Ismail, L.K. Teh, Y.F. Ngeow, M.N. Norazmi, Z.F. Zainul, T.H. Tang, N. Najimudin, M.Z. Salleh, Draft genome sequence of a clinical isolate of *Mycobacterium tuberculosis* strain PR05. Genome Announcements 1 (3) (2013).
[4] J. Cowan, S. Pandey, L.G. Filion, J.B. Angel, A. Kumar, D.W. Cameron, Comparison of interferon-γ-, interleukin (IL)-17- and IL-22-expressing CD4 T cells, IL-22-expressing granulocytes and proinflammatory cytokines during latent and active tuberculosis infection. Clin. Exp. Immunol. 167 (2) (2012) 317–329.
[5] S. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, B. C.3., T. F., Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393 (6685) (1998) 537–544.