Complete Genome Sequence of a New Zealand *Mycobacterium tuberculosis* Strain Responsible for Ongoing Transmission over the Past 30 Years

Claire V. Mulholland,[a,g] Gregory Gimenez,[b] Adele Williamson,[a] Mackenzie Steele,[a] Duncan Thorpe,[d] Noel Karalus,[d] Ray T. Cursons,[a] Veronica M. Playle,[e,f,g] Sally A. Roberts,[f,g] Gregory M. Cook,[c,g] Vickery L. Arcus,[a,g] Htin Lin Aung[c,g]

aSchool of Science, University of Waikato, Hamilton, New Zealand
bDepartment of Pathology, Dunedin School of Medicine, Dunedin, New Zealand
cDepartment of Microbiology and Immunology, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
dDepartment of Pathology, Waikato Hospital, Waikato District Health Board, Hamilton, New Zealand
eDepartment of Molecular Medicine and Pathology, University of Auckland, Auckland, New Zealand
fLabPlus, Auckland City Hospital, Auckland District Health Board, Auckland, New Zealand
gMaurice Wilkins Centre for Molecular Biodiscovery, University of Auckland, Auckland, New Zealand

Claire V. Mulholland and Gregory Gimenez contributed equally to this work. Author order was determined alphabetically.

ABSTRACT We report here the complete genome sequence of *Mycobacterium tuberculosis* strain Colonial S-type 1 (CS1), which has been responsible for ongoing outbreaks of tuberculosis in New Zealand over the past 30 years. CS1 appears to be highly transmissible, with greater rates of progression to active disease, compared to other circulating *M. tuberculosis* strains; therefore, comparison of its genomic content is of interest.

Tuberculosis (TB) is a treatable disease, but it still claims over 1.5 million lives annually (1). Molecular typing of *Mycobacterium tuberculosis* isolates circulating in New Zealand shows that the Colonial S-type 1 (CS1) strain, formerly known as the Rangipo strain, represents the largest cluster of isolates (2, 3). Phylogenetically, CS1 belongs to sublineage L4.4.1.1 of the Euro-American lineage of *M. tuberculosis*, and it has been the cause of ongoing outbreaks in the New Zealand population for at least the past 30 years (4, 5). This warrants further studies to characterize its genetics in order to understand its spread and virulence. Here, we report the complete genome sequence of the New Zealand strain CS1, which was isolated in 1996 from a cerebrospinal fluid sample as part of routine clinical diagnosis and strain typing. Ethical approval was not required for this purpose or sequencing of New Zealand clinical *M. tuberculosis* isolates.

This isolate was recovered from a frozen culture stock and cultured in modified 7H9 medium with Middlebrook oleic acid-albumin-dextrose-catalase (OADC) growth supplement (10%) using the BACTEC mycobacterial growth indicator tube (MGIT) system. DNA was extracted using a cetyltrimethylammonium bromide (CTAB) extraction protocol adapted from reference 6. Briefly, the culture was heat inactivated in Tris-EDTA (TE) buffer, and the pellet was then treated for 2 h at 37°C with lysozyme and RNase A, followed by overnight digestion with proteinase K at 56°C. DNA was then extracted by phenol-chloroform extraction, followed by CTAB-chloroform extraction and isopropanol precipitation (6).

The full-length CS1 genome was sequenced with an RS II (P6-C4 chemistry) instrument (Pacific Biosciences [PacBio], Menlo Park, CA, USA) using SMRTbell libraries with an additional size selection step with the 20-kb-template BluePippin size selection system (Sage Science, Beverly, MA, USA). Paired-end 250-bp short-read sequencing was performed on a MiSeq platform (Illumina, Inc., San Diego, CA, USA) using the Nextera XT DNA library preparation kit (Illumina).

1

We obtained 157,670 long reads and 1,280,644 paired-end short reads with the PacBio and Illumina systems, respectively. Long-read lengths ranged from 35 to 51,700 bp, with an N_{50} value of 7.852 kb. Read quality control was performed using FastQC. Long reads were first trimmed and error corrected (17,152 reads, yielding 153,575,068 cleaned bases) and then assembled using Canu (version 1.8) (7), which resulted in the assembly of a single large circular contig. This contig was polished and error corrected with Pilon (version 1.23) (8) using Illumina short reads, with a mean coverage of $>50\times$. Canu was run with the following options: genomeSize=4.4M, minReadLength=500, and minOverlapLength=200; Pilon was run with default parameters.

The resulting circular genome has a length of 4,416,671 nucleotides, with a GC content of 65.6%, and was annotated using Prokka (version 1.14) (9) with H37Rv as a reference genome (GenBank accession number NC_000962.3). A total of 4,135 coding sequences were found, together with 56 tRNAs and 3 rRNAs. The CS1 strain and H37Rv proteomes share 2,918 proteins with strong homology ($>80\%$). As a reference genome, the CS1 sequence reported here will be helpful in further studies to elucidate the genetic bases for the transmission of this New Zealand strain and its virulence characteristics and host specificity.

Data availability. The whole-genome sequence of New Zealand *M. tuberculosis* strain CS1 has been deposited in DDBJ/ENA/GenBank under the accession number CP044345. The raw reads were deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA573545 and the SRA accession numbers SRR20995191, SRR20998128, SRR20998236, and SRR20998249.

REFERENCES

1. World Health Organization. 2021. Global tuberculosis report 2021. https://www.who.int/publications/i/item/9789240037021.
2. Sexton KS, Petera S, Pandey S. 2008. Five years of molecular typing of Mycobacterium tuberculosis isolates in New Zealand, 2003 to 2007. Institute of Environmental Science and Research Ltd., Porirua, New Zealand. https://surv.esr.cri.nz/PDF_surveillance/TbMolTyp/TBMolecTypReport.pdf?m=1221688560.
3. Aung HL, Devine TJ, Mulholland CV, Arcus VL, Cook GM. 2019. Tackling tuberculosis in the indigenous people of New Zealand. Lancet Public Health 4:e496. https://doi.org/10.1016/S2468-2667(19)30180-X.
4. Mulholland CV, Ruthe A, Cursons RT, Durrant R, Karalus N, Coley K, Bower J, Permina E, Coleman MJ, Roberts SA, Arcus VL, Cook GM, Aung HL. 2017. Rapid molecular diagnosis of the *Mycobacterium tuberculosis* Rangipo strain responsible for the largest recurring TB cluster in New Zealand. Diagn Microbiol Infect Dis 88:138–140. https://doi.org/10.1016/j.diagmicrobio.2017.03.012.
5. Mulholland CV, Shockey AC, Aung HL, Cursons RT, O'Toole RF, Gautam SS, Brites D, Gagneux S, Roberts SA, Karalus N, Cook GM, Pepperell CS, Arcus VL. 2019. Dispersal of *Mycobacterium tuberculosis* driven by historical European trade in the South Pacific. Front Microbiol 10:2778. https://doi.org/10.3389/fmicb.2019.02778.
6. van Helden PD, Victor TC, Warren RM, van Helden EG. 2001. Isolation of DNA from *Mycobacterium tuberculosis*. Methods Mol Med 54:19–30. https://doi.org/10.1385/1-59259-147-7:019.
7. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
8. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
9. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.