Complete Genome Sequence of *Francisella tularensis* Live Vaccine Strain NR-28537 (BEI Master Cell Bank)

Robert A. Player, a Kathleen J. Verratti, a Sarah L. Grady, a Linda C. Beck, b Bruce G. Goodwin, c Christopher G. Earnhart, c Shanmuga Sozhamannan, e

a Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, USA
b Joint Research and Development, Inc., Stafford, Virginia, USA
c Joint Program Executive Office, Joint Project Lead, Chemical, Biological, Radiological, and Nuclear Defense, Enabling Biotechnologies, Frederick, Maryland, USA
d Defense Biological Product Assurance Office, Chemical, Biological, Radiological, and Nuclear Defense, Enabling Biotechnologies, Frederick, Maryland, USA
e Logistics Management Institute, Tysons, Virginia, USA

**ABSTRACT** The genome of *Francisella tularensis* live vaccine strain NR-28537 was sequenced by a hybrid approach utilizing an Oxford Nanopore Technologies R9 flow cell and an Illumina MiSeq platform. *De novo* assembly of the resulting long and short reads produced a single-contig whole-genome sequence.

*Francisella tularensis* is the causative agent of the disease tularemia and is a CDC category A biothreat agent. While tularemia causes significant morbidity and death, no *F. tularensis* vaccine is currently licensed in the United States due to incomplete understanding of the mechanism of live vaccine strain (LVS) attenuation.

A live culture of Master Cell Bank *Francisella tularensis* LVS strain NR-28537 was obtained from BEI Resources. A bacterial inoculum was streaked onto a chocolate agar plate and incubated overnight at 37°C. The entire culture growth was scraped off the plate and used for nucleic acid extraction with the Lucigen MasterPure complete DNA and RNA purification kit (catalog number MC85200). Genomic DNA was quantified using a Qubit v4.0 fluorometer and quality checked on the TapeStation 2200 (DNA integrity number [DIN] 6.9). Sequencing libraries were prepared using an Oxford Nanopore Technologies (ONT) ligation kit (SQK-LSK109) and an Illumina Nextera XT kit (catalog number FC-131-1024). The Illumina library was sequenced using a 600-cycle reagent kit (catalog number MS-102-3003) on a MiSeq system to produce 2 × 300-bp paired-end reads. The ONT library was sequenced on an R9 flow cell (FLO-MIN106D) using the GridION platform. Illumina reads were filtered with a minimum length of 35 bp and a minimum Phred quality score of 20 using Cutadapt v2.5 with Python v3.6.10 (1). Raw (24.07 million reads) and trimmed (24.04 million reads) Illumina reads were quality checked using FastQC v0.11.5 (2). Trimmed reads were joined using PEAR v0.9.10 for assembly polishing (21.63 million reads; average read length, 341 bp) (3); 603,833 base-called ONT reads (Guppy server v3.2.10+aabd4ec) were filtered with an average base Phred quality score of >7 to 511,393 pass reads (read N50, 3,951 bp) and then quality checked using LRPlot v0.5.2 (4).

Genome scaffolding was accomplished by generating an overlapping read file by mapping ONT pass reads against themselves using minimap2 v2.15.911-dirty and then running miniasm (with default settings) on the overlap file (v0.3-r179) (5, 6). The resulting single contig was polished three times with ONT pass reads and then two times with Illumina joined reads using Racon v1.4.3 (7). A measure of assembly completeness was calculated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.0 with the order taxonomic rank database thiotrichales_odb10 (8). Statistics reported in Table 1 were derived from ONT (minimap2 v2.15.911-dirty) and Illumina (HISAT2 v2.0.0-beta) alignments (9).
The single polished contig has a GC content of 32.15% and a length of 1,890,193 bp, and >99.98% identity to *F. tularensis* subsp. *holarctica* LVS (GenBank accession number GCF_000833335.1) was determined via alignment using minimap2 (with the -x parameter set to asm5 [5, 6]). Two high-quality single-nucleotide polymorphisms were identified between the two sequences, both involving transposase genes, i.e., (i) ATT-to-AT deletion at position 259332, causing a frame shift in an IS5 family transposase, and (ii) G-to-A transition at position 825679, causing an upstream gene variant in an IS630 family transposase. Of 491 complete BUSCOs expected to be found in a genome of the order *Thiotrichales*, 439 (89.4%) complete, 26 (5.3%) fragmented, and no duplicated sequences were identified, a high-quality assembly with a low duplication rate. Additionally, upon mapping of both read sets to the reference genome sequences were identified, indicating a high-quality assembly with a low duplication rate. The single polished contig has a GC content of 32.15% and a length of 1,890,193 bp, and >99.98% identity to *F. tularensis* subsp. *holarctica* LVS (GenBank accession number GCF_000833335.1), deletions of the pilA and FTT0918 genes implicated in virulence attenuation of LVSs were confirmed in the NR-28537 strain (10, 11).

ONT and Illumina reads were used for *de novo* assembly and polishing of the assembly, respectively. Depth and breadth of coverage were estimated based on read mapping to the polished assembly. The number of coding sequences (CDSs) was determined by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (12).

**Data availability.** The whole-genome sequence has been deposited in NCBI under BioProject and BioSample accession numbers PRJNA668347 and SAMN16400345, respectively. The corresponding raw sequencing data from Illumina MiSeq and ONT GridION sequencing have been deposited in the SRA under accession numbers SRR12802542 and SRR12802916, respectively. The assembly has been deposited in NCBI GenBank under accession number CP063128.

**ACKNOWLEDGMENT**

Funding for the current study was provided by the Joint Program Executive Office for Chemical, Biological, Radiological, and Nuclear Defense (CBRND) Joint Project Lead CBRND Enabling Biotechnologies-Special Immunization Program through NAVSEA contract N00024-13-D-6400.

**REFERENCES**

1. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMNet J 17:10–12. https://doi.org/10.14806/ej.17.1.200.
2. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
3. Zhang J, Robert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End read merger. Bioinformatics 30:614–620. https://doi.org/10.1093/bioinformatics/btt593.
4. Hauser A. Longread_plots: long read plotting package, version 0.5.2. https://github.com/ahcm/longread_plot.
5. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
6. Li H. 2016. Minimap and Miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32:2103–2110. https://doi.org/10.1093/bioinformatics/btw152.
7. Vaser R, Sovic I, Nagarajan N, Sišč M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116.
8. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
9. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915. https://doi.org/10.1038/s41587-019-0201-4.
10. Forlund A-L, Kuoppa K, Svensson K, Salomonsson I, Johansson A, Byström M, Oyston PCT, Mitchell SL, Titball RW, Noppa L, Frithz-Lindsten E, Forsman M, Forsberg A. 2006. Direct repeat-mediated deletion of a type IV pilin gene results in major virulence attenuation of *Francisella tularensis*. Mol Microbiol 59:1818–1830. https://doi.org/10.1111/j.1365-2958.2006.05061.x.
11. Twine S, Byström M, Chen W, Forsman M, Golovllov I, Johansson A, Kelly J, Lindgren H, Svensson K, Zingmark C, Conlan W, Sjöstedt A. 2005. A mutant of *Francisella tularensis* strain SCHU S4 lacking the ability to express a 58-kilodalton protein is attenuated for virulence and is an effective live vaccine. Infect Immun 73:8345–8352. https://doi.org/10.1128/IAI.73.12.8345-8352.2005.
12. Tatusova T, DiCuccio M, Badretdin A, Chevertvin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.