Complete Genome Sequence of *Nocardia farcinica* W6977^T Obtained by Combining Illumina and PacBio Reads

Christopher A. Gulvik,[a] Robert A. Arthur,[a] Ben W. Humrighouse,[a] Dhwani Batra,[b] Lori A. Rowe,[b] Brent A. Lasker,[a] John R. McQuiston[a]

[a]Bacterial Special Pathogens Branch, Division of High-Consequence Pathogens and Pathology, National Center for Emerging and Zoonotic Infectious Disease, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[b]Biotechnology Core Facility Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

ABSTRACT The complete genome sequence of the *Nocardia farcinica* type strain was obtained by combining Illumina HiSeq and PacBio reads, producing a single 6.29-Mb chromosome and 2 circular plasmids. Bioinformatic analysis identified 5,991 coding sequences, including putative genes for virulence, microbial resistance, transposons, and biosynthesis gene clusters.

*N*ocardia farcinica, first isolated in 1888 (1), is an opportunistic bacterial pathogen of clinical relevance due to high levels of morbidity and mortality and a high intrinsic degree of antimicrobial resistance (2, 3). We report here the completed genome sequence of the *Nocardia farcinica* type strain obtained from the ATCC and identify potential markers for virulence, antimicrobial resistance, transposons, and the production of secondary biosynthesis metabolites.

A single colony was inoculated into Trypticase soy broth and grown at 35°C for 5 days. Genomic DNA was extracted using the MasterPure DNA purification kit (Epicentre, Madison, WI, USA). A 20-kb library was prepared with the SMRTbell template prep kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA). The library was bound to polymerase using the DNA/polymerase binding kit P6v2 (PacBio), loaded on a single-molecule real-time (SMRT) cell (PacBio), and then sequenced with C4v2 chemistry (PacBio) on an RS II (PacBio) instrument. An aliquot of the same DNA preparation was sheared on an M220 Focused-ultrasonicator (Covaris, Inc., Woburn, MA) to generate DNA fragments averaging 500 bp in length. The NEBNext Ultra DNA kit (New England BioLabs, Ipswich, MA, USA) was used to create libraries, and paired-end sequencing (2 × 250 bp) was performed on a MiSeq version 2 500-cycle reagent kit (Illumina, San Diego, CA). PacBio raw reads of ≥1 kbp were converted into FastQ format with bash5tools 0.8.0. These 77,719 long reads (643,112,291 bp) were scrubbed using DASCRUBBER-wrapper (4) and Gene Myers's Dazzler utilities (DALIGNER, datander, DAScover, DASedit, DASpatch, DASqv, DAStrim, REPmask, and TANmask) (5). Using default parameters, reads were self-aligned with DALIGNER in order to mask interspersed repeats with REPmask, and tandem repeats (3.3% of reads, 0.6% of bases) were identified with datander and masked with TANmask. The reads were then realigned with repeat sites masked to find overlaps and estimate coverage with DAScover. The intrinsic qualities were plotted with DASqv to identify thresholds (best 80% were ≥Q22, worst 7% were ≤Q29), which were used in read trimming and patching with DAStrim and DASpatch. Finally, 72,526 scrubbed reads (516,421,081 bp) were captured with DASedit and DB2fasta. This scrubbing process discarded 106.7 Mbp (16.6%) and repaired 48.8 Mbp (7.5%) low-quality nucleotides, removed 110.9 Mbp of chimeras, and clipped off 8.9 Mbp of missed adaptamers. Paired-end Illumina reads were cleaned with BBDUK 37.77 to remove PhiX, and Trimmomatic 0.36 (6) was used to remove adapters and discard sequences with a Phred score of less than 30. Illumina and PacBio reads

were then assembled with Unicycler 0.4.6 (7), which depended on SPAdes 3.12.0 (8), minimap2 (9), miniasm (10), racon 1.3.1 (11), and BLAST 2.7.1+ (12) for assembly. Short-read polishing of the assembly was also accomplished in Unicycler, which used ALE 4aec46e (13), Bowtie 2.3.4.1 (14), SAMtools 1.8 (15), and Pilon 1.22 (16) and fixed 24 assembly errors. CheckM 1.0.11 (17) indicated that the genome is 99.8% complete (two missing markers) when using 799 *Nocardia* genus reference markers.

The genome was annotated with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP**)** (18). The W6799^T genome contains 6,291,633 bp, with a G+C content of 70.0%, 5,991 coding sequences (CDSs), 3 complete rRNA operons, 57 tRNAs, 5 CRISPR arrays, and two circular plasmids of 96.7 and 41.2 kb, respectively. A total of 129 virulence-related genes and 11 antimicrobial resistance-related genes, many previously described for *Nocardia farcinica* IFM 10152 (19), were identified. Thirteen transposable elements, 85 genomic biosynthesis gene clusters (BGCs), and 1 plasmid BGC were identified using antiSMASH 4.0 (20), suggesting that the genome shows great potential for the discovery of new natural products.

Data availability. The whole-genome sequence of *Nocardia farcinica* W6977^T has been deposited at the DDBJ/ENA/GenBank database under the accession number CP031418 for the chromosome and numbers CP031419 and CP031420 for plasmids 1 and 2, respectively. Illumina and PacBio raw reads have been submitted to the SRA under BioSample number SAMN09723209.

ACKNOWLEDGMENTS

REFERENCES

1. Goodfellow M, Maldonado LA. 2012. Genus *Nocardia* Trevisan 1889, p 376–419. *In* Goodfellow M, Kämpfer P, Busse HJ, Trujillo ME, Suzuki K, Ludwig W, Whitman W (ed), Bergey's manual of systematic bacteriology, 2nd ed, vol 5. Springer, New York, NY.

2. Brown-Elliott BA, Brown JM, Conville PS, Wallace RJ, Jr. 2006. Clinical and laboratory features of the *Nocardia* spp. based on current molecular taxonomy. Clin Microbiol Rev 19:259–282. https://doi.org/10.1128/CMR.19.2.259-282.2006.

3. Zhao P, Zhang X, Du P, Li G, Li L, Li Z. 2017. Susceptibility profiles of Nocardia spp. to antimicrobial and antituberculotic agents detected by a microplate Alamar Blue assay. Sci Rep 7:43660. https://doi.org/10.1038/srep43660.

4. Wick RR. 2018. DASCRUBBER. https://github.com/rrwick/DASCRUBBER-wrapper.

5. Myers EW. 2018. DALIGNER. https://github.com/thegenemyers.

6. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

7. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.

9. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

10. 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.

11. Vaser R, Sović I, Nagarajan N, Šikić 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.

12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K,

Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 15:421. https://doi.org/10.1186/1471-2105-10-421.

13. Clark SC, Egan R, Frazier PI, Wang Z. 2013. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics 29:435–443. https://doi.org/10.1093/bioinformatics/bts723.

14. Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.

15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

16. 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.

17. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114.

18. Tatusova T, DiCuccio M, Badretdin A, Chetvernin 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.

19. Ishikawa J, Yamashita A, Mikami Y, Hoshino Y, Kurita H, Hotta K, Shiba T, Hattori M. 2004. The complete genome sequence of *Nocardia farcinica* IFM 10152. Proc Natl Acad Sci U S A 101:14925–14930. https://doi.org/10.1073/pnas.0406410101.

20. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de Los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T, Medema MH. 2017. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res 45:W36–W41. https://doi.org/10.1093/nar/gkx319.