genomeA
Journals.ASM.org

Complete Genome Sequence of Steroid-Transforming *Nocardioides simplex* VKM Ac-2033D

Victoriya Y. Shtratnikova,[a] Mikhail I. Schelkunov,[a] Yury A. Pekov,[a] Victoria V. Fokina,[b] Mariya D. Logacheva,[a,c] Sergey L. Sokolov,[b] Eugeny Y. Bragin,[b] Vasily V. Ashapkin,[c] Marina V. Donova[b]

Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia[a]; G.K. Skryabin Institute of Biochemistry & Physiology of Microorganisms, Russian Academy of Sciences, Pushchino, Russia[b]; A.N. Belozersky Department of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia[c]

***Nocardioides simplex* VKM Ac-2033D is an effective microbial catalyst for 3-ketosteroid 1(2)-dehydrogenation, and it is capable of effective reduction of carbonyl groups at C-17 and C-20, hydrolysis of acetylated steroids, and utilization of natural sterols. Here, the complete genome sequence is reported. An array of genes related to steroid metabolic pathways have been identified.**

Received 25 November 2014 **Accepted** 1 December 2014 **Published** 8 January 2015

Citation Shtratnikova VY, Schelkunov MI, Pekov YA, Fokina VV, Logacheva MD, Sokolov SL, Bragin EY, Ashapkin VV, and Donova MV. 2015. Complete genome sequence of steroid-transforming *Nocardioides simplex* VKM Ac-2033D. Genome Announc 3(1):e01406-14. doi:10.1128/genomeA.01406-14.

Address correspondence to Marina V. Donova, donova@ibpm.pushchino.ru.

Nocardioides simplex VKM Ac-2033D (synonyms, *Arthrobacter simplex* [basonym] and *Pimelobacter simplex* [senior homotypic synonym] [1]) effectively introduces a 1(2)-double bond in various 1(2)-saturated 3-ketosteroids, thus enabling the production of valuable pharmaceuticals and immediate precursors for the steroid industry (2-4). The strain is also capable of effective hydrolysis of acetylated steroids (3), utilization of natural sterols, and the reduction of carbonyl groups at C-17 and C-20 of androstanes and pregnanes, respectively. This bacterium of soil origin was first classified as *Arthrobacter globiformis* 193 and then reclassified as *N. simplex* VKM Ac-2033D based on a complex analysis using a polyphase taxonomic approach (2).

The short-read library containing DNA fragments of 226 ± 33-bp insert length was prepared with a TruSeq DNA sample preparation kit (Illumina) after digestion of the genomic DNA with NEBNext double-stranded DNA (dsDNA) fragmentase. The library was read on a HiSeq 2000 (with paired-end 100-nucleotide reads). The mate-pair libraries with 3,222 ± 251-bp-long to 9,992 ± 2,172-bp-long fragments were created with the Nextera mate-pair sample preparation kit (Illumina) and were sequenced on a MiSeq. NextClip 0.8 (5) was used to remove possible paired-end contaminations. Both the paired-end and mate-pair reads were adapter and quality trimmed by Trimmomatic 0.32 (6). The mean coverage of the genome by three libraries was 1,989×. *De novo* genome assembly was performed with Velvet 1.2 (7) and SPAdes 2.5 (8) using paired-end reads and with SPAdes 3.1.0, CLC Genomics Workbench 6.0, and MaSuRCA 2.3.2 (9) using both paired-end and mate-pair reads. The produced contigs were manually combined into a single circular contig in BioEdit (10). The quality of the resulting contig was assessed by REAPR 1.0.17 (11). The contig was also checked by mapping reads in CLC Genomics Workbench and by a visual inspection of putatively ambiguous places.

The length of the genome is 5,637,355 nucleotides (nt), and the G+C content is 72.66%. Annotation of the genome was carried out with the service RAST (http://rast.nmpdr.org/) and with GenBank tools. The RAST annotation revealed 5,421 protein-coding sequences, and the GenBank annotation revealed 4,633 coding sequences (CDS) and 816 pseudogenes; both annotations show 46 tRNAs (44 of which were unique), one pseudo-tRNA and 6 rRNAs. A preliminary analysis of the sequences showed several clusters of genes involved in cholesterol metabolism (side chain degradation, steroid core degradation, and transport).

The reported complete genome sequence will contribute to the elucidation of the range of the steroid substrates that may be metabolized by this organism and the revelation of the scope of its potential application in pharmaceutical steroid production.

Nucleotide sequence accession number. The complete genome sequence has been deposited in GenBank under the accession no. CP009896.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (grant no. 14-04-31089).

We acknowledge A. S. Kondrashov for providing sequencing on HiSeq 2000 and MiSeq.

REFERENCES

1. **Garrity GM, Bell GA, Lilburn TG.** 2004. Taxonomic outline of the prokaryotes. In Bergey's manual of systematic bacteriology, 2nd ed, release 5.0. Springer-Verlag, New York, NY.
2. **Fokina VV, Sukhodol'skaya GV, Gulevskaya SA, Gavrish EY, Evtushenko LI, Donova MV.** 2003. The 1(2)-dehydrogenation of steroid substrates by *Nocardioides simplex* VKMAc-2033D. Microbiology **72:**24–29. http://dx.doi.org/10.1023/A:1022265720470.
3. **Fokina VV, Sukhodolskaya GV, Baskunov BP, Turchin KF, Grinenko GS, Donova MV.** 2003. Microbial conversion of pregna-4,9(11)-diene-17alpha,21-diol-3,20-dione acetates by *Nocardioides simplex* VKM Ac-2033D. Steroids **68:**415–421. http://dx.doi.org/10.1016/S0039-128X(03)00043–6.
4. **Fokina VV, Donova MV.** 2003. 21-Acetoxypregna-4(5),9(11),16(17)-triene-21-ol-3,20-dione conversion by *Nocardioides simplex* VKM Ac-

2033D. J Steroid Biochem Mol Biol **87:**319–325. http://dx.doi.org/10.1016/j.jsbmb.2003.10.002.

5. **Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M.** 2014. NextClip: an analysis and read preparation tool for Nextera long mate pair libraries. Bioinformatics **30:**566–568. http://dx.doi.org/10.1093/bioinformatics/btt702.
6. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.
7. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
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. http://dx.doi.org/10.1089/cmb.2012.0021.
9. **Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA.** 2013. The MaSuRCA genome assembler. Bioinformatics **29:**2669–2677. http://dx.doi.org/10.1093/bioinformatics/btt476.
10. **Hall TA.** 1999. BioEdit: a user-friendly biological sequence alignment Editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser **41:**952669–98.
11. **Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD.** 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol **14:**R47. http://dx.doi.org/10.1186/gb-2013-14-5-r47.