AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

High-Quality Sequencing, Assembly, and Annotation of the *Streptomyces griseofuscus* DSM 40191 Genome

Tetiana Gren,[a] Tue S. Jørgensen,[a] Christopher M. Whitford,[a] Tilmann Weber[a]

[a]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark

Tetiana Gren and Tue S. Jørgensen contributed equally to this work; the author order was determined alphabetically.

ABSTRACT Here, we report the sequencing, assembly, and annotation of the genome of *Streptomyces griseofuscus* DSM 40191. The genome of *S. griseofuscus* was sequenced using PacBio and Illumina technologies. It consists of a linear 8,721,740-bp chromosome and three plasmids, pSGRIFU1 (220 kb), pSGRIFU2 (88 kb), and pSGRIFU3 (86 kb).

Streptomyces griseofuscus DSM 40191 (NRRL B-5429) was originally isolated from a soil sample collected near Dake Hot Spring, Fukushima Prefecture, Japan (1), as a producer of the antibiotics bundlin A and B (lankacidins C and A) and later submitted to the DSMZ collection. We believe that this strain is an interesting producer of various secondary metabolites and a promising heterologous expression host. Its current RefSeq genome assembly (accession number GCF_000718315.1) is fragmented into 244 contigs and is not suited for genome mining (2). The strain was grown in yeast extract-malt extract (YEME) medium according to reference 3, and genomic DNA was purified using the Genomic Tip 100 kit (Qiagen, Venlo, Netherlands). An Illumina (San Diego, CA) whole-genome sequencing (WGS) library was constructed using the KAPA (St. Louis, MO, USA) HYPRplus kit, and 3,535,535 read clusters were generated on an Illumina MiSeq sequencer with a 2× 150-nucleotide (nt) kit. Macrogen, Inc. (Seoul, South Korea), generated the PacBio (Menlo Park, CA, USA) RS II data (191,463 subreads; N_{50}, 15,731 nt) using the 8PAC V3, DNA polymerase binding kit P6, and 2 single-molecule real-time (SMRT) cells; a g-TUBE (Covaris, Inc., Woburn, MA, USA) was used for DNA shearing followed by BluePippin size selection (Sage Science, MA, USA). Default software parameters were used except where otherwise noted. The Illumina data were adapter trimmed using AdapterRemoval v2 (v.2.2.2) (4) with the switches "–trimns –trimqualities." The PacBio subreads were assembled with Flye (v.2.4.1-geb89c9e) (5) with the switches "–genome-size 8m –iterations 5." Then, the assembly was polished with the polishing module of Unicycler using the Illumina data (Unicycler v. 0.4.8-beta) (6). The inverted repeats, which are typically found at the ends of the linear *Streptomyces* chromosomes, were manually added according to the assembly graph, and the genome was polished again. We confirmed that the inverted repeat ends of the chromosome were correctly attached by identifying reads overlapping the junctions using minimap2 (v. 2.16-r922) (7), Bowtie 2 (v.2.3.5) (8), and Artemis (v.18.0.2) (9). The telomeric sequence from the related strain *Streptomyces rochei* 7434AN4 (accession number AB905441.1) was found within the inverted repeat but not at the absolute terminus. The coverage of PacBio data was found to be uniform and ca. 150× in the chromosome. The assembled genome was annotated using Prokka (v.1.14.0) with the default databases as well as the PFAM-A (v.32.0) (10) database and six actinobacterial genomes as reference data sets (NC_003155.5, NC_004719.1, NC_010572.1, NC_008596.1, NC_014318.1, NC_003888.3, NC_003903.1, NC_003904.1, NC_021985.1, NC_022001.1, and NC_021986.1). RNAmmer (v.1.2) (11) was used for rRNA predictions. The polished

Citation Gren T, Jørgensen TS, Whitford CM, Weber T. 2020. High-quality sequencing, assembly, and annotation of the *Streptomyces griseofuscus* DSM 40191 genome. Microbiol Resour Announc 9:e01100-20. https://doi.org/10.1128/MRA.01100-20.

Editor Julie C. Dunning Hotopp, University of Maryland School of Medicine

Address correspondence to Tilmann Weber, tiwe@biosustain.dtu.dk.

Received 23 September 2020
Accepted 29 October 2020
Published 19 November 2020

genome of *S. griseofuscus* consists of a chromosome of 8,721,740 bp with inverted repeat ends of ca. 84 kb and the following three plasmids: pSGRIFU1 (220 kb), pSGRIFU2 (88 kb), and pSGRIFU3 (86 kb). The assembly graph reveals the architecture of the linear genome with inverted repeat ends and two linear plasmids (pSGRIFU1 and pSGRIFU2) and a third (pSGRIFU3) that is highly repetitive and has a complex assembly path that is not resolved. The average GC content of the genome is 71.63%. Of the 8,114 coding DNA sequences (CDSs), 5,893 (73%) were functionally annotated. A total of 292/292 (100%) BUSCO (v.4.0.5) (12) actinobacterial core genes (actinobacteria_class_odb10 ortholog set) were found in single copy. Comparing the final assembly to the existing draft RefSeq version (GCF_000718315.1) of *S. griseofuscus*, the main difference is the continuity, with the final assembly being complete. Furthermore, mapping the existing draft RefSeq version (GCF_000718315.1) to the finished version and calculating the differences between them with minimap2 (with the switch "-N 1000") and SAMtools (v.1.7, using htslib 1.7–2) (13), we found a total of 186,462 nt in the present assembly which are not covered by the draft version (GCF_000718315.1), including the majority of the inverted repeat chromosome ends and several gaps between the contigs, as well as 44 single nucleotide variations and six insertions and deletions.

Data availability. All data are available as BioProject PRJNA622435. Raw reads are deposited at SRA with accession numbers SRR11462380 (Illumina) and SRR11462379 (PacBio). The GenBank accession numbers are CP051006, CP051007, CP051008, and CP051009.

ACKNOWLEDGMENTS

We thank Alexandra Hoffmeyer for her invaluable help with sequencing.

This work was funded by grants of the Novo Nordisk Foundation, Denmark (NNF10CC1016517 and NNF16OC0021746).

REFERENCES

1. Sakamoto JMJ, Konōd S-I, Yumoto H, Arishima M. 1962. Bundlins A and B, two antibiotics produced by Streptomyces griseofuscus nov. sp. J Antibiot Ser A 15:98–102.
2. Blin K, Kim HU, Medema MH, Weber T. 2019. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. Brief Bioinform 20:1103–1113. https://doi.org/10.1093/bib/bbx146.
3. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA. 2000. Practical Streptomyces genetics. John Innes Foundation, Norwich, United Kingdom.
4. Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes 9:88. https://doi.org/10.1186/s13104-016-1900-2.
5. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8.
6. 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.
7. Li H. 2018. minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
8. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
9. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28:464–469. https://doi.org/10.1093/bioinformatics/btr703.
10. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995.
11. Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160.
12. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness, p 227–245. *In* Kollmar M (ed), Gene prediction: methods and protocols. Springer, New York, NY.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.