AMERICAN SOCIETY FOR MICROBIOLOGY

Microbiology® Resource Announcements

GENOME SEQUENCES

Annotated Genome Sequence of *Aspergillus tanneri* NIH1004

Stephanie Mounaud,[a] Pratap Venepally,[a] Indresh Singh,[a] Liliana Losada,[a]* Seyedmojtaba Seyedmousavi,[b]* Kyung J. Kwon-Chung,[b] William C. Nierman[a]

[a]Department of Infectious Disease, J. Craig Venter Institute, Rockville, Maryland, USA

[b]Molecular Microbiology Section, Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

Kyung J. Kwon-Chung and William C. Nierman contributed equally to this work.

ABSTRACT The annotated genome of *Aspergillus tanneri*, a recently discovered drug-resistant pathogen, was determined by employing the Oxford Nanopore MinION platform and the Funannotate pipeline. The genome size and the number of protein-coding genes are notably larger than those of the most common etiological agent of aspergillosis, *Aspergillus fumigatus*.

Two fatal invasive aspergillosis (IA) cases in chronic granulomatous disease (CGD) patients who failed to respond to aggressive antifungal therapies were caused by a newly discovered *Aspergillus* species, *Aspergillus tanneri* (1), in the *Aspergillus* taxonomic section *Tanneri* (2). Mycological characterization and targeted gene identification of the clinical *A. tanneri* strains were performed (1).

A. tanneri strain NIH1004 was isolated from a 19-year-old CGD patient who suffered from fatal aspergillosis. Primary clinical cultures of *A. tanneri* NIH1004 were grown on Sabouraud dextrose agar. Isolates were subcultured and incubated for 2 to 3 weeks at 37°C. DNA was isolated from lyophilized mycelium that had been grown in liquid yeast glucose medium for 36 h at 37°C and was extracted using the cetyltrimethylammonium bromide (CTAB) method (3) after vigorous mixing with glass beads. Genomic DNA was sheared and sized for preparation of libraries to sequence on three sequencing platforms and was quality checked using an Agilent 2100 Bioanalyzer (Santa Clara, CA), as well as by quantitative PCR (catalog number KK4835; Kapa library quantification kit). Initial sequencing was performed on the Illumina HiSeq 2500 platform using 100-bp paired-end reads and 8-kb paired-end 454 reads. The Illumina sequence reads provided 30× genome coverage, consisting of 37,410,025 bp (G+C content, 47.4%). Reads were assembled *de novo* using the Celera Assembler (4), which resulted in 870 contigs with an N_{50} of 134,193 bp. An improved assembly was obtained using a long-read sequencing technology. An *A. tanneri* library was prepared using a ligation sequencing kit (product number SQK-LSK108; Oxford Nanopore) and was analyzed in an Oxford 9.4.1 flow cell with a MinION device. Assembly was performed using Minimap2 (5) and miniasm (6), with default parameters. The sequences within the assembled contigs were error corrected using Racon v1.3.1 (7) and Pilon (v1.22; four rounds) (8). The reads used for error correcting were generated using wgsim (https://github.com/lh3/wgsim) to simulate reads based on the Celera Assembler-assembled contigs. We generated 4 million simulated 150-bp paired-end reads with a quality score of 40. The wgsim tool was modified from the MAQ read simulator by dropping dependencies; wgsim was originally released in the SAMtools software package. The resulting MinION assembly consisted of 38,719,388 bp (G+C content, 47.3%). This improved *A. tanneri* assembly resulted in 14 contigs, with an N_{50} of 4,499,170 bp, and is the first published *A. tanneri* sequence.

Whole-genome annotation of the *A. tanneri* assembly was performed using the

Citation Mounaud S, Venepally P, Singh I, Losada L, Seyedmousavi S, Kwon-Chung KJ, Nierman WC. 2020. Annotated genome sequence of *Aspergillus tanneri* NIH1004. Microbiol Resour Announc 9:e01374-19. https://doi.org/10.1128/MRA.01374-19.

Editor Jason E. Stajich, University of California, Riverside

Address correspondence to Kyung J. Kwon-Chung, jkchung@niaid.nih.gov, or William C. Nierman, wnierman@jcvi.org.

* Present address: Liliana Losada, Division of Microbiology and Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA; Seyedmojtaba Seyedmousavi, Microbiology Service, Department of Laboratory Medicine, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA.

Received 31 October 2019
Accepted 6 December 2019
Published 16 January 2020

Funannotate pipeline (v1.5.1-93c317b) (9). Initially, following the masking of repeats identified by RepeatMasker (v1.332) (10) and RepeatModeler (v1.0.11) (10), *ab initio* gene models for the contig sequences were predicted using the GeneMark-ES (v4.36) (11) and AUGUSTUS (v3.2.3) (12) programs. Evidence-based gene models were generated by aligning the contig sequences from the *A. tanneri* genome with the combined protein sequence (UniProtKB) database using DIAMOND (v0.9.21.122) (13) and later polishing using Exonerate (v2.4.0) (14). EVidenceModeler (v0.1.30) (15) with its weighting algorithm, as implemented in the Funannotate pipeline, was used to select the consensus models from among the *ab initio* and evidence-based gene models. Functional annotation of the consensus models was performed after removal of those with short lengths, gaps, and transposable elements. A total of 11,846 genes were associated with 64,436 annotations by performing sequence similarity searches against the Pfam (v32.0) (16), InterPro (v71.0) (17), BUSCO (v2.0) (18), EggNOG (v4.5) (19), MEROPS (v12.0) (19), and CAZyme (v7.0) (20) databases and using the SignalP secretome prediction program (v4.1) (21). The tRNA genes were identified by using tRNAscan-SE (v1.23) (22).

The biosynthetic gene cluster (BGC) mining program antiSMASH (v4.1.0) (23), with its Minimum Information on Biosynthetic Gene cluster (MIBiG) repository of experimentally characterized BGCs (24), was utilized to identify 95 distinct secondary metabolite BGCs. This number of clusters is considerably higher than those in a set of eight related aspergilli, with a range from the highest at 68 for *A. niger* to the lowest at 39 for *A. fumigatus* at 39 (25).

Data availability. This whole-genome shotgun project has been deposited in DDBJ/ENA/GenBank under the accession number QUQM00000000. Raw sequence reads have been deposited in the SRA under accession number SRX4502713.

ACKNOWLEDGMENTS

This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract HHSN272200900007C. K.J.K.-C. is supported by a research fund from the intramural program of the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

REFERENCES

1. Sugui JA, Peterson SW, Clark LP, Nardone G, Folio L, Riedlinger G, Zerbe CS, Shea Y, Henderson CM, Zelazny AM, Holland SM, Kwon-Chung KJ. 2012. *Aspergillus tanneri* sp. nov., a new pathogen that causes invasive disease refractory to antifungal therapy. J Clin Microbiol 50:3309–3317. https://doi.org/10.1128/JCM.01509-12.
2. Jurjević Ž, Kubátová A, Kolařík M, Hubka V. 2015. Taxonomy of *Aspergillus* section *Petersonii* sect. nov. encompassing indoor and soil-borne species with predominant tropical distribution. Plant Syst Evol 301: 2441–2462. https://doi.org/10.1007/s00606-015-1248-4.
3. Jurjevic Z, Peterson SW, Horn BW. 2012. *Aspergillus* section *Versicolores*: nine new species and multilocus DNA sequence based phylogeny. IMA Fungus 3:59–59. https://doi.org/10.5598/imafungus.2012.03.01.07.
4. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC. 2000. A whole-genome assembly of *Drosophila*. Science 287:2196–2204. https://doi.org/10.1126/science.287.5461.2196.
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, 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.
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. Palmer J. 2016. Funannotate: pipeline for genome annotation. https://funannotate.readthedocs.io/en/latest/index.html.
10. Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4:Unit 4.10. https://doi.org/10.1002/0471250953.bi0410s05.
11. Mattupalli C, Glasner JD, Charkowski AO. 2014. A draft genome sequence reveals the *Helminthosporium solani* arsenal for cell wall degradation. Am J Potato Res 91:517–524. https://doi.org/10.1007/s12230-014-9382-z.
12. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439. https://doi.org/10.1093/nar/gkl200.
13. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
14. Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31. https://doi.org/10.1186/1471-2105-6-31.
15. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9:R7. https://doi.org/10.1186/gb-2008-9-1-r7.
16. 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.

17. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong S-Y, Finn RD. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351–D360. https://doi.org/10.1093/nar/gky1100.
18. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548. https://doi.org/10.1093/molbev/msx319.
19. Ferrés I, Iraola G. 2018. Phylen: automatic phylogenetic reconstruction using the EggNOG database. J Open Source Softw 3:593. https://doi.org/10.21105/joss.00593.
20. Terrapon N, Lombard V, Drula E, Coutinho PM, Henrissat B. 2017. The CAZy database/the Carbohydrate-Active Enzyme (CAZy) database: principles and usage guidelines, p 117–131. *In* Aoki-Kinoshita K (ed), A practical guide to using glycomics databases. Springer, Tokyo, Japan.
21. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423. https://doi.org/10.1038/s41587-019-0036-z.
22. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.955.
23. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87. https://doi.org/10.1093/nar/gkz310.
24. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Düsterhus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJN, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann H-W, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, et al. 2015. Minimum information about a biosynthetic gene cluster. Nat Chem Biol 11:625–631. https://doi.org/10.1038/nchembio.1890.
25. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, Fedorova ND. 2010. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol 47:736–741. https://doi.org/10.1016/j.fgb.2010.06.003.