Improved Reference Genome Sequence of *Coccidioides immitis* Strain WA_211, Isolated in Washington State

Marcus de Melo Teixeira,[a,b] Bridget Marie Barker,[b] Jason E. Stajich[c]

aFaculty of Medicine, University of Brasília, Brasília, Federal District, Brazil
bThe Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, USA
cDepartment of Microbiology & Plant Pathology and Institute for Integrative Genome Biology, University of California, Riverside, Riverside, California, USA

ABSTRACT *Coccidioides* fungi are widely distributed in the American continents, with an expanding western range documented by a recently discovered cryptic population of *Coccidioides immitis* in Washington State. The assembled and annotated reference genome sequence of the soil-derived *C. immitis* strain WA_211 will support population and functional genomics studies.

*C*occidioides immitis and *Coccidioides posadasii* are fungal species found in desert-like areas of the American continents (1) and are the causative agents of coccidioidomycosis or "valley fever." While *C. posadasii* is broadly distributed, *C. immitis* is restricted to southern California and northern Mexico. Infections due to *C. immitis* in California are increasing and reached alarming rates in 2017 (2). The disease range appears to be expanding, with uncommon autochthonous infections reported in Washington State (3). Washington *C. immitis* isolates from the soil and clinic are reciprocally monophyletic within *C. immitis* (4, 5) and rarely hybridize with *C. posadasii* (6). As most assembled genomes of *C. immitis* are patient-derived isolates, we annotated and assembled the genome of the soil-derived strain WA_211, as it represents a unique emerging lineage of *C. immitis*, for comparative, population, and functional genomics research.

The WA_211 strain was cultured from soil on yeast extract medium at 37°C for 5 days (5), and its DNA was isolated after growth on Sabouraud's medium using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany). The sequencing library was prepared with a Kapa Biosystems (Woburn, MA) kit (catalog number kk8201) and sequenced with an Illumina MiSeq platform (2 × 300 bp) (4, 5). The 2.8 million read pairs (1.7 Gbp) were obtained from the SRA and processed with shovill v.1.0.4 (https://github.com/tseemann/shovill) using "–minlen 500 –trim" and shovill defaults for downstream tools. The shovill pipeline trimmed reads for adaptors and low quality using Trimmomatic v.0.39 (7), corrected bases using Lighter v.1.1.2 (8), and merged overlapping read pairs using FLASH v.1.2.11 (9) to produce 1.6 million merged reads and 1 million unmerged read pairs. Contigs were assembled with SPAdes v.3.10.1 (10), polished with one round of Pilon v.1.22 (11), and cleaned of vector sequence and redundant contigs by AAFTF v.0.2.1 (12), using default parameters. The 297 contigs were scaffolded to the *C. immitis* RS genome (GenBank accession number AAEC00000000) (13) with RagOO v.1.1 (14), with default parameters, to produce a 27.4-Mb assembly of 62 scaffolds (N_{50}, 3.79 Mb; longest scaffold length, 8.24 Mb; G+C content, 46.4%). This scaffolding assumes the colinearity of WA_211 and RS, but no breakpoints were observed within contigs in a comparison of a dotplot by D-GENIES (15).

The genome contains 15% repetitive sequences masked by RepeatMasker v.open-4.0.7 using a *Coccidioides* repeat element library (16). Genes were predicted in the masked genome with Funannotate v.1.5.2 (17). Gene prediction parameters were generated by "funannotate train" using alignments of *C. immitis* spherule and hyphal

Address correspondence to Jason E. Stajich, jason.stajich@ucr.edu.

RNA sequencing (RNA-seq) (18). Gene prediction ("funannotate predict") was performed running *ab initio* predictors Augustus (19), SNAP v.2013-11-29 (20), Coding-Quarry v.2.0 (21), and GeneMark-ES v.4.33 (22) using exon hints from spliced alignments of transcripts, Onygenales proteins (13, 23, 24), and the Swiss-Prot database (25). Consensus gene models were generated by Funannotate running EVidenceModeler v.1.1.1 (26). Funannotate assigned putative gene products by searches to the Swiss-Prot, InterPro (27), eggNOG (28), MEROPS (29), and dbCAN (30) databases, using default parameters. A total of 7,815 protein-coding gene models were predicted, of which 5,477 had InterPro domains. Twenty-one secondary metabolite clusters were predicted by antiSMASH 4.0 (31), comprising 8 polyketide synthases (PKS), 4 nonribosomal peptide synthetases (NRPS), 2 hybrid PKS-NRPS, 1 indole-NRPS, 3 terpene, and 4 "other" type clusters. The annotation and assembly pipeline steps, full parameters, and log files are archived in the GitHub repository (32).

Data availability. This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number RHJW00000000. The version described in this paper is version RHJW02000000. The genomic sequence reads used in this assembly were previously deposited under SRA project accession number SRR1292227, and the RNA-seq reads used had been deposited under BioProject number PRJNA169242 and SRA accession number SRP013923.

REFERENCES

1. Lewis ERG, Bowers JR, Barker BM. 2015. Dust devil: the life and times of the fungus that causes valley fever. PLoS Pathog 11:e1004762. https://doi.org/10.1371/journal.ppat.1004762.

2. California Department of Public Health. 2019. Epidemiologic summary of coccidioidomycosis in California, 2017. Infectious Diseases Branch, Division of Communicable Disease Control, Center for Infectious Diseases, California Department of Public Health, Sacramento, CA. https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/CocciEpiSummary2017.pdf.

3. James AE, Pastenkos G, Bradway D, Baszler T. 2019. Autochthonous transmission of *Coccidioides* in animals, Washington, USA. Emerg Infect Dis J 25:123. https://doi.org/10.3201/eid2501.180411.

4. Engelthaler DM, Roe CC, Hepp CM, Teixeira M, Driebe EM, Schupp JM, Gade L, Waddell V, Komatsu K, Arathoon E, Logemann H, Thompson GR, III, Chiller T, Barker B, Keim P, Litvintseva AP. 2016. Local population structure and patterns of Western Hemisphere dispersal for *Coccidioides* spp., the fungal cause of valley fever. mBio 7:e00550-16. https://doi.org/10.1128/mBio.00550-16.

5. Litvintseva AP, Marsden-Haug N, Hurst S, Hill H, Gade L, Driebe EM, Ralston C, Roe C, Barker BM, Goldoft M, Keim P, Wohrle R, Thompson GR, III, Engelthaler DM, Brandt ME, Chiller T. 2015. Valley fever: finding new places for an old disease: *Coccidioides immitis* found in Washington State soil associated with recent human infection. Clin Infect Dis 60:e1–e3. https://doi.org/10.1093/cid/ciu681.

6. Maxwell CS, Mattox K, Turissini DA, Teixeira MM, Barker BM, Matute DR. 2019. Gene exchange between two divergent species of the fungal human pathogen, Coccidioides. Evolution 73:42–58. https://doi.org/10.1111/evo.13643.

7. 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.

8. Song L, Florea L, Langmead B. 2014. Lighter: fast and memory-efficient sequencing error correction without counting. Genome Biol 15:509. https://doi.org/10.1186/s13059-014-0509-9.

9. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. https://doi.org/10.1093/bioinformatics/btr507.

10. 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.

11. 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.

12. Stajich JE, Palmer J. 2018. AAFTF: Automated Assembly for the Fungi, v0.2.1. https://doi.org/10.5281/zenodo.1658103.

13. Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, Jordar VS, Maiti R, Kodira CD, Neafsey DE, Zeng Q, Hung CY, McMahan C, Muszewska A, Grynberg M, Mandel MA, Kellner EM, Barker BM, Galgiani JN, Orbach MJ, Kirkland TN, Cole GT, Henn MR, Birren BW, Taylor JW. 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. Genome Res 19:1722–1731. https://doi.org/10.1101/gr.087551.108.

14. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. Fast and accurate reference-guided scaffolding of draft genomes. bioRxiv. https://doi.org/10.1101/519637.

15. Cabanettes F, Klopp C. 2018. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958. https://doi.org/10.7717/peerj.4958.

16. Kirkland TN, Muszewska A, Stajich JE. 2018. Analysis of transposable elements in *Coccidioides* species. J Fungi 4:13. https://doi.org/10.3390/jof4010013.

17. Palmer JM, Stajich JE. 2018. Funannotate: eukaryotic genome annotation pipeline. http://funannotate.readthedocs.io.

18. Whiston E, Zhang Wise H, Sharpton TJ, Jui G, Cole GT, Taylor JW. 2012.

Comparative transcriptomics of the saprobic and parasitic growth phases in *Coccidioides* spp. PLoS One 7:e41034. https://doi.org/10.1371/journal.pone.0041034.

19. Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080.

20. Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59. https://doi.org/10.1186/1471-2105-5-59.

21. Testa AC, Hane JK, Ellwood SR, Oliver RP. 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16:170. https://doi.org/10.1186/s12864-015-1344-4.

22. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18:1979–1990. https://doi.org/10.1101/gr.081612.108.

23. Neafsey DE, Barker BM, Sharpton TJ, Stajich JE, Park DJ, Whiston E, Hung CY, McMahan C, White J, Sykes S, Heiman D, Young S, Zeng Q, Abouelleil A, Aftuck L, Bessette D, Brown A, FitzGerald M, Lui A, Macdonald JP, Priest M, Orbach MJ, Galgiani JN, Kirkland TN, Cole GT, Birren BW, Henn MR, Taylor JW, Rounsley SD. 2010. Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. Genome Res 20:938–946. https://doi.org/10.1101/gr.103911.109.

24. Mitchell NM, Sherrard AL, Dasari S, Magee DM, Grys TE, Lake DF. 2018. Proteogenomic re-annotation of Coccidioides posadasii strain Silveira. Proteomics 18:1700173. https://doi.org/10.1002/pmic.201700173.

25. Stockinger H, Altenhoff AM, Arnold K, Bairoch A, Bastian F, Bergmann S, Bougueleret L, Bucher P, Delorenzi M, Lane L, Le Mercier P, Lisacek F, Michielin O, Palagi PM, Rougemont J, Schwede T, von Mering C, van Nimwegen E, Walther D, Xenarios I, Zavolan M, Zdobnov EM, Zoete V, Appel RD. 2014. Fifteen years SIB Swiss Institute of Bioinformatics: life science databases, tools and support. Nucleic Acids Res 42:W436–W441. https://doi.org/10.1093/nar/gku380.

26. 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.

27. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

28. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. https://doi.org/10.1093/nar/gky1085.

29. Rawlings ND, Waller M, Barrett AJ, Bateman A. 2014. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 42:D503–D509. https://doi.org/10.1093/nar/gkt953.

30. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: a Web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445–W451. https://doi.org/10.1093/nar/gks479.

31. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W, Breitling R, Takano E, Medema MH. 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43:W237–W243. https://doi.org/10.1093/nar/gkv437.

32. Stajich JE. 2019. stajichlab/Coccidioides_immitis_211: Coccidioides immitis 211 annotation for MRA. https://zenodo.org/record/3300712#.XTemWB1Ki1s.