AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Genome Sequence of Flavor-Producing Yeast *Saprochaete suaveolens* NRRL Y-17571

Hana Lichancová,[a] Viktória Hodorová,[a] Karolina Sienkiewicz,[b] Sarah Mae U. Penir,[c] Philipp Afanasyev,[d] Dominic Boceck,[e] Sarah Bonnin,[f] Siras Hakobyan,[g] Pawel S. Krawczyk,[h] Urszula Smyczynska,[i] Erik Zhivkoplias,[j] Maryna Zlatohurska,[k] Adrian Odrzywolski,[l] Eugeniusz Tralle,[m] Alina Frolova,[n] Leszek P. Pryszcz,[m] Broňa Brejová,[o] Tomáš Vinař,[p] Jozef Nosek[a]

[a]Department of Biochemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovak Republic
[b]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
[c]Philippine Genome Center, National Science Complex, University of the Philippines Diliman, Quezon City, Philippines
[d]Laboratory of Evolutionary Genomics, Vavilov Institute of General Genetics, Moscow, Russia
[e]Algorithms in Bioinformatics, ZBIT Center for Bioinformatics, University of Tübingen, Tübingen, Germany
[f]Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain
[g]Institute of Molecular Biology NAS RA, Yerevan, Armenia
[h]Laboratory of RNA Biology and Functional Genomics, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
[i]Department of Biostatistics and Translational Medicine, Medical University of Lodz, Lodz, Poland
[j]Biology Education Centre, Uppsala University, Uppsala, Sweden
[k]Institute of Microbiology and Virology, National Academy of Science of Ukraine, Kyiv, Ukraine
[l]Medical University in Lublin, Lublin, Poland
[m]International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland
[n]Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine
[o]Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovak Republic
[p]Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovak Republic

ABSTRACT *Saprochaete suaveolens* is an ascomycetous yeast that produces a range of fruity flavors and fragrances. Here, we report the high-contiguity genome sequence of the ex-holotype strain, NRRL Y-17571 (CBS 152.25). The nuclear genome sequence contains 24.4 Mbp and codes for 8,119 predicted proteins.

Saprochaete suaveolens is a fermentative yeast from the *Magnusiomyces*/*Saprochaete* clade (phylum *Ascomycota*, subphylum *Saccharomycotina*). It has been isolated from nutrient-rich sources, including industrial wastes, brewery water, process water from wheat-starch production plants, effluent milk, maize mash, soybean flakes, figs, and dragon fruits, and some strains were isolated from patients with pulmonary infections (1–3). It produces large amounts of volatile organic compounds with an intensive fruity odor (3–5).

The *S. suaveolens* strain NRRL Y-17571 was originally isolated from water in a brewery (1). Its genome was assembled by the combination of long reads (MinION, Oxford Nanopore Technologies) and short reads (HiSeq 2000, Illumina). DNA was isolated from a culture grown overnight in yeast extract-peptone-dextrose (YPD) medium (1% [wt/vol] yeast extract, 2% [wt/vol] peptone, 1% [wt/vol] glucose) at 28°C using a standard protocol and purified using the DNeasy mini spin column (Qiagen) for HiSeq 2000 analysis or Genomic-tip 100/G (Qiagen) for MinION analysis (6). Total cellular RNA from the midexponential phase culture grown in yeast extract-peptone-galactose (YPGal) medium (1% [wt/vol] yeast extract, 2% [wt/vol] peptone, 2% [wt/vol] galactose) at 28°C was extracted with hot acidic phenol (7) and purified with the RNeasy minikit (Qiagen).

Citation Lichancová H, Hodorová V, Sienkiewicz K, Penir SMU, Afanasyev P, Boceck D, Bonnin S, Hakobyan S, Krawczyk PS, Smyczynska U, Zhivkoplias E, Zlatohurska M, Odrzywolski A, Tralle E, Frolova A, Pryszcz LP, Brejová B, Vinař T, Nosek J. 2019. Genome sequence of flavor-producing yeast *Saprochaete suaveolens* NRRL Y-17571. Microbiol Resour Announc 8:e00094-19. https://doi.org/10.1128/MRA.00094-19.

Editor Christina Cuomo, Broad Institute of MIT and Harvard University

Address correspondence to Broňa Brejová, brejova@dcs.fmph.uniba.sk, or Jozef Nosek, jozef.nosek@uniba.sk.

Received 25 January 2019
Accepted 4 February 2019
Published 28 February 2019

TABLE 1 Candidate genome assemblies[a]

Assembly software	Software version	Polishing procedure	Length of assembly (Mbp)	No. of contigs	No. of contigs >50 kbp	Longest contig	N_{50} value	No. of mismatches per 100 kbp	No. of indels per 100 kbp
SPAdes (18)	3.12.0		24.2	2,224	137	640 kbp	173 kbp		
Canu (19)	1.7.1	Pilon (2×)	24.9	26	24	4.2 Mbp	1.7 Mbp	41.7	13.9
MaSuRCA (20)	3.2.8	Pilon (1×)	25.4	29	23	6.8 Mbp	2.7 Mbp	45.7	7.6
Miniasm (21)/minimap2 (22)	0.3/2.12	Racon (2×)	24.5	15	13	3.7 Mbp	2.8 Mbp	81.5	367.9
DBG2OLC (23)/platanus (24)	1.2.4	Racon (1×), Pilon (1×)	25.8	31	24	4.2 Mbp	1.8 Mbp	55.2	35.1
Final		Racon (2×), Pilon (2×)	24.4	12	11	3.8 Mbp	2.8 Mbp	45.6	11.0

[a] Statistics were produced with Quast v. 4.5 (15). To estimate mismatches and indels, SPAdes assembly based on Illumina short reads was used as a reference. With SPAdes, the result was filtered for length >100 and coverage >10. Canu assembly used only reads overlapping SPAdes by >200 bp, and we filtered out contigs supported by fewer than 5 reads. All assemblies were polished with Pilon v. 1.21 (16) and Racon v. 1.3.1 (17). Most of the size differences between candidate assemblies can be accounted for by mtDNA and rRNA gene fragments as well as other repetitive sequences.

We obtained 204,824 long reads (mean, 9,011 nucleotides [nt]; longest read, 211,620 nt) totaling 1.8 Gbp (~74× coverage) with a MinION Mk-1B device on a R9.4.1 flow cell with a SQK-LSK109 kit and base called with ONT Albacore (v. 2.3.1). A paired-end (2 × 101 nt) TruSeq PCR-free DNA library was sequenced on a HiSeq 2000 platform in Macrogen, Korea, which yielded 64,378,402 reads (6.4 Gbp, ~262× coverage). RNA-Seq was performed with NovaSeq 6000 system in Macrogen, Korea, which yielded 42,932,052 reads from a TruSeq mRNA V2 nonstranded paired-end (2 × 101 nt) library.

Table 1 presents candidate genome assemblies. The final assembly is based on miniasm, which had the smallest number of contigs and did not show apparent assembly artifacts. To further improve this assembly, we removed contigs containing fragments of mitochondrial DNA (mtDNA) and rRNA genes, individually polished rRNA gene repeats, and replaced regions upstream and downstream of rRNA gene repeats with 505 bp from DBG2OLC and 309 bp from Canu assemblies, respectively. The nuclear genome has a GC content of 39.5% and likely consists of at least 7 chromosomes, because both ends of 4 contigs and one end of 6 contigs are terminated by telomeric repeats with a predominant motif CA_3G_{5-7}. About 2% of the genome (508 kbp) is covered by simple and low-complexity repeats identified with RepeatMasker v. 4.0.7 (8).

RNA-Seq reads processed with Trimmomatic v. 0.36 (9) were assembled into transcripts with Trinity v. 2.8.3 (10). We trained Augustus v. 3.2.3 (11) on the *Magnusiomyces capitatus* data set (12) and, using RNA-Seq transcripts aligned to the reference with blat v. 34 × 1 (13), we predicted 8,119 protein-coding genes.

The genome sequence of *S. suaveolens* will provide a basis for understanding metabolic pathways involved in the production of volatile organic compounds, suitable as flavors and aromas in the food industry, and genetic traits associated with the ability to colonize humans.

Data availability. This whole-genome shotgun assembly has been deposited in EMBL ENA under the accession no. CAAAMA010000000. Illumina, MinION, and RNA-Seq reads have been deposited under accession no. ERR3039972, ERR3040055, and ERR3039974, respectively. Genome annotations are available through a genome browser at http://genome.compbio.fmph.uniba.sk/ and are also archived through Zenodo (14).

ACKNOWLEDGMENTS

We thank Cletus P. Kurtzman and James Swezey (Agricultural Research Service, Peoria, IL, USA) for providing us with the yeast strain.

Nanopore sequencing, genome assembly, and genome annotation were performed during the hackathon at the #NGSchool2018: Nanopore sequencing & personalised medicine (September 16 to 23, 2018) bioinformatics school organized in Lublin, Poland (https://ngschool.eu/2018) supported by International Visegrad Fund project 21810033.

The computations were done with the help of cloud services and resources from national e-infrastructure providers through the Training Infrastructure of the EGI Fed-

eration. The project was supported by grants from the Slovak Research and Development Agency (no. APVV-14-0253 to J.N.) and VEGA (no. 1/0684/16 to B.B. and no. 1/0458/18 to T.V.). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 665778 (to L.P.P.). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. de Hoog GS, Smith MT. 2011. *Saprochaete* Coker & Shanor ex D.T.S. Wagner & Dawes (1970), p 1317–1327. *In* Kurtzman CP, Fell JW, Boekhout T (ed), The yeasts: a taxonomic study, 5th ed. Elsevier, London, United Kingdom.
2. Kolecka A, Khayhan K, Groenewald M, Theelen B, Arabatzis M, Velegraki A, Kostrzewa M, Mares M, Taj-Aldeen SJ, Boekhout T. 2013. Identification of medically relevant species of arthroconidial yeasts by use of matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 51:2491–2500. https://doi.org/10.1128/JCM.00470-13.
3. Grondin E, Shum Cheong Sing A, Caro Y, Raherimandimby M, Randrianierenana AL, James S, Nueno-Palop C, François JM, Petit T. 2015. A comparative study on the potential of epiphytic yeasts isolated from tropical fruits to produce flavoring compounds. Int J Food Microbiol 203:101–108. https://doi.org/10.1016/j.ijfoodmicro.2015.02.032.
4. Grondin E, Shum Cheong Sing A, Caro Y, de Billerbeck GM, François JM, Petit T. 2015. Physiological and biochemical characteristics of the ethyl tiglate production pathway in the yeast *Saprochaete suaveolens*. Yeast 32:57–66.
5. Grondin E, Shum Cheong Sing A, James S, Nueno-Palop C, François JM, Petit T. 2017. Flavour production by *Saprochaete* and *Geotrichum* yeasts and their close relatives. Food Chem 237:677–684. https://doi.org/10.1016/j.foodchem.2017.06.009.
6. Hodorova V, Lichancova H, Bujna D, Nebohacova M, Tomaska L, Brejova B, Vinar T, Nosek J. 2018. De novo sequencing and high-quality assembly of yeast genomes using a MinION device. London Calling, 24 and 25 May 2018, London, UK. https://nanoporetech.com/resource-centre/de-novo-sequencing-and-high-quality-assembly-yeast-genomes-using-minion-device.
7. Collart MA, Oliviero S. 1993. Preparation of yeast RNA. Curr Protoc Mol Biol 23:13.12.1–13.12.5. https://doi.org/10.1002/0471142727.mb1312s23.
8. Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.
9. 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.
10. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883.
11. Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62. https://doi.org/10.1186/1471-2105-7-62.
12. Brejová B, Lichancová H, Brázdovič F, Hegedűsová E, Forgáčová Jakúbková M, Hodorová V, Džugasová V, Baláž A, Zeiselová L, Cillingová A, Neboháčová M, Raclavský V, Tomáška Ľ, Lang BF, Vinař T, Nosek J. 2018. Genome sequence of the opportunistic human pathogen *Magnusiomyces capitatus*. Curr Genet. https://doi.org/10.1007/s00294-018-0904-y.
13. Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res 12:656–664. https://doi.org/10.1101/gr.229202.
14. Lichancová H, Hodorová V, Sienkiewicz K, Penir SMU, Afanasyev P, Boceck D, Bonnin S, Hakobyan S, Krawczyk PS, Smyczynska U, Zhivkoplias E, Zlatohurska M, Odrzywolski A, Tralle E, Frolova A, Pryszcz LP, Brejová B, Vinař T, Nosek J. 2019. Annotations of sapSuaA1 genome assembly (version 1) [data set]. Zenodo. https://doi.org/10.5281/zenodo.2555651.
15. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
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. 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.
18. 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.
19. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
20. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669–2677. https://doi.org/10.1093/bioinformatics/btt476.
21. 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.
22. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
23. Ye C, Hill CM, Wu S, Ruan J, Ma Z. 2016. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep 6:31900. https://doi.org/10.1038/srep31900.
24. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 24:1384–1395. https://doi.org/10.1101/gr.170720.113.