Short Communication

Chromosome-level genome assembly and transcriptome-based annotation of the oleaginous yeast *Rhodotorula toruloides* CBS 14

Giselle C. Martín-Hernández [a,1], Bettina Müller [a,1], Mikołaj Chmielarz [a], Christian Brandt [b,c], Martin Hölzer [b,d], Adrian Viehweger [b,e], Volkmar Passoth [a,*]

[a] *Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden*
[b] *nanozoo GmbH, Leipzig, Germany*
[c] *Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany*
[d] *RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Jena, Germany*
[e] *Department of Medical Microbiology, University Hospital Leipzig, Germany*

ARTICLE INFO

ABSTRACT

Rhodotorula toruloides is an oleaginous yeast with high biotechnological potential. In order to understand the molecular physiology of lipid synthesis in *R. toruloides* and to advance metabolic engineering, a high-resolution genome is required. We constructed a genome draft of *R. toruloides* CBS 14, using a hybrid assembly approach, consisting of short and long reads generated by Illumina and Nanopore sequencing, respectively. The genome draft consists of 23 contigs and 3 scaffolds, with a N50 length of 1,529,952 bp, thus largely representing chromosomal organization. The total size of the genome is 20,534,857 bp and the overall GC content is 61.83%. Transcriptomic data from different growth conditions was used to aid species-specific gene annotation. We annotated 9464 genes and identified 11,691 transcripts. Furthermore, we demonstrated the presence of a potential plasmid, an extrachromosomal circular structure of about 11 kb with a copy number about three times as high as the other chromosomes.

1. Introduction

The basidiomycete yeast *Rhodotorula toruloides* is an oleaginous microorganism with high biotechnological potential for lipid and carotenoid production. This yeast can naturally accumulate lipids up to 70% of dry cell weight, and a number of carotenoids [20,42]. *R. toruloides* can be cultivated to high cell densities on a wide range of substrates, including lignocellulose hydrolysate and other residual products [11,26,48,51,61]. This makes *R. toruloides* a promising host for the production of single-cell oil, as a sustainable and less controversial alternative to plant-derived oils for the production of biofuels, food and feed additives [42].

The molecular physiology behind lipid synthesis in *R. toruloides* has been relatively little explored, which hinders effective metabolic engineering for improved lipid production. Some draft genome sequences

from *R. toruloides* strains have been determined using short-read sequencing technologies, however a complete picture of the *R. toruloides* genome on chromosomal level is still lacking [23,33,38,43,47,55,62,64]. The combination of short- and long-read sequencing strategies has been shown to improve the accuracy of genome sequences in yeast [41,54].

Thus, we combined Nanopore long-read sequencing and Illumina short-read sequencing to obtain a chromosome-level genome assembly. We further generated comprehensive transcriptomic data from different growth conditions to aid species-specific annotation. The results provide a valuable resource for pathway analysis and manipulation of *R. toruloides* and enable better understanding of genome biology and evolution of basidiomycetous yeasts.

---

* Corresponding author at: Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden.
  E-mail address: Volkmar.Passoth@slu.se (V. Passoth).
[1] Giselle C. Martín-Hernández and Bettina Müller equally contributed to this paper and shall both be regarded as first authors.

2. Materials and methods

2.1. DNA extraction

R. toruloides CBS 14 (Westerdijk Fungal Biodiversity Institute, Utrecht, the Netherlands) was grown in 50 mL YPD medium [10]. Cells were harvested during exponential growth and the cell wall was digested according to Pi et al. [45]. DNA was extracted from the protoplasts using NucleoBond® CB 20 Kit (Macherey-Nagel, Germany). Concentration, purity and integrity were assessed with Qubit™ 4 Fluorometer, NanoDrop® ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA) and agarose gel electrophoresis, respectively.

2.2. RNA extraction

R. toruloides was cultivated in 500 mL bioreactors (Multifors, Infors HT, Bottmingen, Switzerland) with either a mixture of lignocellulose hydrolysate and crude glycerol or crude glycerol only as carbon source [5,11]. Additionally, the media contained 0.75 g/L yeast extract (Bacto™ Yeast Extract, BD, France), 1 g/L $MgCl_2$ $6xH_2O$ (Merck KGaA, Germany), 2 g/L $(NH_4)_2HPO_4$ (≥98%, Sigma-Aldrich, USA), and 1.7 g/L YNB without amino acids and ammonium sulfate (Difco™, Becton Dickinson, France). Cultivations were performed in triplicates, at 25 °C, pH 6 and oxygen tension of 21%.

Total RNA was extracted in triplicates from 5 mL samples withdrawn after 10, 40, 72 and 96 h, respectively from bioreactors with mixed carbon source, and 10, 30 and 60 h, respectively from bioreactors with glycerol as sole carbon source, using the Quick-RNA™ Fecal/Soil Microbe MicroPrep Kit (Zymo Research, USA) according to the manufacturer's instructions with some modifications [37]. Briefly, the cells were harvested and resuspended in 1 mL Trizol and disrupted in a FastPrep −24 bead beater (M.P. Biomedicals, USA) at speed 6.0 m/s for 40 s. The homogenate was separated into layers by adding 0.2 mL chloroform and centrifugation. 400 μL of the upper layer containing the RNA was further processed as described in the manufacturer's manual. DNAse I treatment was performed as described in the manual applying 26 U/mL at 37° for 15 min. The technical replicates of the retrieved RNA were pooled prior rRNA depletion, which was performed using the human riboPOOL™ (siTOOLsBiotech, Germany) and Streptavidin-coated magnetic beads (Thermo Fisher Scientific, Norway) following the two-step protocol of the manufacturer. rRNA-depleted samples were purified by ethanol precipitation. RNAse inhibitor (1 U/μL, Thermo Fisher) was added before storage. Total RNA and rRNA-depleted samples were tested for integrity and quantity on RNA Nano Chips (Agilent 2100 Bioanalyzer System, Agilent Technologies, Germany).

2.3. Library preparation and sequencing

The extracted DNA was divided into two samples and sequenced using either MinION (Oxford Nanopore Technologies) or Illumina sequencing platform. Before Nanopore library preparation, 5 μg of the retrieved DNA were "pre-cleaned" using 31.5 μL of AMPure magnetic beads, for removing short DNA fragments [6]. The DNA library was prepared using the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK) and a modified protocol [6]. The library was loaded onto a FLO-MIN106 flow cell attached to the MinION device (Oxford Nanopore Technologies). Sequencing was performed using the MinKNOW software (Oxford Nanopore Technologies) according to the protocol from Brandt et al. [6]. The retrieved long reads had a mean length of 7889 bp and a read length N50 of 11,739 bp. The total number of reads was 124,417 and the total number of base pairs sequenced was 981.6 Mb.

Short-read paired-end sequencing of DNA and rRNA-depleted samples was performed on the Illumina Novaseq platform (S prime, 2 × 150 bp) using TruSeq PCR free DNA library preparation kit (Illumina Inc.).

In order to verify the circular structure of contig_63, Sanger sequencing (Macrogen Europe B·V, the Netherlands) was performed from PCR amplicons using the primers shown in supplementary table s1. The PCR mixture consisted of 0.3 μL DNA, 1.25 μL primer, 12.5 μL Dream Taq Green PCR Mix (Thermo Scientific, Lithuania) and 0.8 μL DMSO in a total volume of 25 μL. Amplification was conducted as follows: initial denaturation at 95 °C for 5 min, 35 cycles of denaturation (95 °C for 30 s), annealing (1 min) and elongation (72 °C) followed by a final extension at 72 °C for 10 min. Annealing temperatures and elongation times were adapted to the respective primer combination (table S2). Amplification was assessed on agarose gel electrophoresis at 8.2 V/cm electric field strength. The gel was prepared using GelRed® Nucleic Acid Gel Stain (Biotium), 1.0% agarose and TAE 1× buffer (VWR Life Science). The PCR products from the corresponding sizes were purified using GeneJet Gel Extraction kit (Thermo Scientific, Lithuania).

Geneious prime version 2021.0.1 (Biomatters Ltd.) was used for assembly of the Sanger sequences obtained and to align *FAS2* and *FAS21* from *R. toruloides* CBS 14 and *FAS2* from *R. toruloides* NP11. Whole genome alignment was performed using the LASTZ sequence alignment tool (version 7.0.2) implemented in Geneious [22].

2.4. Quality control and genome assembly

Short and long reads generated by Illumina and Nanopore sequencing, respectively, were combined for hybrid *de-novo* assembly. Before, the quality of the long reads was ensured using NanoPlot v1.25.0 [15] and all short reads were removed until reaching a target base coverage with Filtlong v0.2.0 (https://github.com/rrwick/Filtlong) [59] applying the parameters --target_bases 5,000,000,000 and --length_weight 8. Short reads were quality-trimmed and adapter-clipped with fastp v0.20.0 [9] using parameters -5 -3 -W 4 -M 20 -l 25 -x -n 5 -z 6. To achieve high contiguity in the initial assembly step, a draft assembly was produced from the preprocessed long reads using Flye v2.8 [30], setting the suggested genome size to 20 Mbp and keeping the plasmid and meta options activated. The Flye draft assembly was further subjected to long-read-based polishing using two rounds of Racon v1.4.7 [56] followed by one round of Medaka v0.10.0 (https://github.com/nanoporetech/medaka) [53] using the model r941_min_high. We used minimap2 (−x map-ont) v2.17 [34] and samtools v.1.10 [36] to prepare alignment files for the Racon polishing. The assembly graph was visualized and investigated via Bandage v0.8.1 [60]. The long-read assembly was finally subjected to two Pilon v1.23 [57] rounds, in which the quality-filtered short reads were used for final polishing. Alignment files for Pilon were prepared using BWA v0.7.17 [35] and samtools. We mapped the Illumina and Nanopore reads with BWA and minimap2, respectively, to the final assembly and used the pileup.sh script from BBMap v38.86 [8] to calculate coverage histograms for each contig with a bin size of 1000 nt and plotted them with ggplot2. Besides the visual inspection of coverage patterns, we used the samtools coverage function and the faidx command on the long-read-mapped data to filter out contigs violating at least one of the following cutoffs: minimum read number = 100, minimum bases covered by reads = 5000, minimum read coverage = 15, and minimum read coverage depth (amount of bases covered by reads in comparison to contig length) = 10. The resulting quality-controlled and cleaned assembly file was used for annotation and chromosome analyses.

2.5. Annotation

The final assembly was first screened for repetitive regions using RepeatMasker v4.0.9 [40], checked for completeness using BUSCO v3.0.2 [50,58] with the fungi_odb9 database as reference and visualized using chromoMap v0.2 [2]. Contigs, larger than 10 kb but smaller than 250 kb were checked for circularization using a python script (https://github.com/Kzra/Simple-Circularise, v1.0) [29].

Taxonomic classification was performed using sourmash v2.0.1 [46] and its "LCA" method. Both indexing the tree and querying genomes

used a k-mer size of 31 and a sampling fraction of 10^{-4}. The LCA index was derived from publicly available genomes (GenBank, https://osf.io/4f8n3/).

The repeat-masked assembly was further annotated using a combination of homology-based gene comparison and RNA-Seq-derived transcript information. First, we used MetaEuk v1.ea903e5 [27] with the "easy-predict" subcommand for draft gene annotation providing all proteins (filtered models; best) obtained from the JGI fungal genome portal MycoCosm (http://jgi.doe.gov/fungi; downloaded April 2020) as a database. Next, we combined and improved the MetaEuk annotation with the obtained RNA-Seq data. Prior to this, RNA-Seq datasets were quality-checked using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [3] and trimmed and adapter-clipped using fastp with the aforementioned parameters. Potential residual rRNA was removed by SortMeRNA v2.1b [31]. The processed RNA-Seq reads were per-sample-mapped to the assembly using HISAT v2.2.0 [28] and subsequently provided to StringTie v2.1.1 [32,44] for transcript annotation, guided by the initial MetaEuk gene annotation. Finally, the MetaEuk-guided StringTie information from each individual RNA-Seq sample were merged into a final single annotation using the StringTie merge functionality. In this step, we again provided the full MycoCosm-based MetaEuk annotation next to the individual RNA-Seq annotations to also include genes that were not recovered in the RNA-Seq data. Lastly, we extracted all annotated gene sequences and used the dammit pipeline v1.2 (http://www.camillescott.org/dammit) [49] for functional annotation. Dammit unites different databases for annotation: Pfam-A, Rfam, OrthoDB, and again BUSCO (fungi_odb9) that we have decorated with additional information derived from UniProt/Swiss-Prot. We extracted the GO terms annotated by dammit and counted their appearance to apply a weight per GO term. GO terms that occured at least 10 times were subjected to REVIGO (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800) to summarize them by removing redundancy and to visualize the results in semantic similarity-based scatterplots and interactive graphs.

We used a custom python script to identify potential telomeres by using known motifs at contig ends. This information was provided along with the Nanopore reads to Tapestry v1.0.0 [14].

3. Results and discussion

The hybrid *de-novo* assembly for *R. toruloides* CBS 14 (Fig. 1) resulted in a 20,534,857 bp genome, which is in line with the 20.4 Mb median reported for this and other *R. toruloides* strains (Table 1). The overall GC content of the obtained genome is 61.83%, which corresponds to previously reported values (61.9% on average). The identified repetitive

sequences represent 2.01% of the total genome length, of which 1.56% are single repeats and 0.45% low complexity regions. The assembled genome resulted in 23 contigs and 3 scaffolds ranging from 5778 to 1,965,970 bp, and an N50 length of 1,529,952 bp (Table 2). This number of contigs and scaffolds was significantly lower than that obtained using short-read sequencing only: the lowest number achieved in previous studies is 186 contigs (Table 1). This clearly shows the higher accuracy, contiguity and completeness of the genome presented here, achieved through the improved coverage using a hybrid assembly approach. 56.37% of the genome is represented by 4 contigs and 3 scaffolds, each larger than 1 Mb. The rest is allocated in 10 medium-size contigs (0.5–1 Mb) and in 9 small contigs (<0.5 Mb), four of which are circularized. Telomere sequences could be detected at both termini of two contigs and at one terminus each of 15 contigs (Table 2). De Jonge et al. [16] identified at least 11 chromosomes in *R. toruloides* CBS 14 by pulsed field gel electrophoresis (PFGE). Zhu et al. [64] predicted 16 chromosomes in *R. toruloides* NP11 applying Illumina sequencing and gap closing using genome walking. When comparing the size of the obtained contigs and scaffolds with the corresponding bands found by PFGE, we identified at least 18 chromosomes in *R. toruloides* CBS 14.

The mitochondrial genome was recovered in one contig with a length of 157 kb. However, the final step of circularizing sequences in an assembly remains challenging and thus can result in shorter or even longer contigs, since assembly programs cannot clearly define where a circular DNA ends, even despite longer reads [24]. Thus, tandem repeats at the end of a contig can increase the length of a linear representation of a circular sequence. Such repeats can be identified by overlapping sequences at the ends of a contig. To account for this, in an additional step, we aligned the contig against itself to identify and remove circular repeats, ultimately resulting in an actual mitochondrial genome size of 69 kb. The mitochondrial genome has a GC content of 40.9%, which agrees with the GC values previously reported for *R. toruloides* strains NP11 and NBRC 0880 [13,63].

In addition to the mitochondrial genome, three further circular contigs (contig_64, contig_49, contig_63) were predicted. They have a comparable GC content of about 62% (Table 2). Among these, contig_63 was particularly noticeable because it showed a read depth approximately three times higher than the other chromosomes, which may indicate relaxed replication regulation (Table 2). We confirmed the circular structure of contig_63 by overlapping amplification using sequence-specific primers and subsequent Sanger sequencing of the amplicons obtained (Supplementary Fig. S1, supplementary file 2). This showed that the circular structure is about 1143 bp larger than the one predicted (Supplementary Fig. S1). As explained above, circular sequence structures possess a certain challenge for assembly tools and
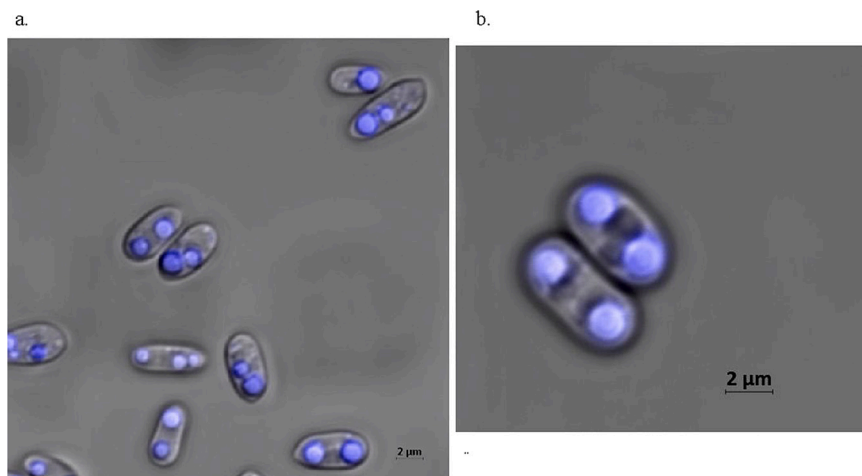
Fig. 1. Confocal microscope picture of Nile Red-stained lipid droplets of *Rhodotorula toruloides* CBS 14. (a) Group of cells at stationary phase (b) Magnified view of cells.

Table 1
Genome assembly statistics of *Rhodotorula toruloides* strains and sequencing platforms used.

Reference	Strain	Sequencing platform	Coverage	Genome size (Mb)	GC content (%)	Contigs	Scaffolds	coding sequences
[33]	MTCC 457	Illumina	121 X	20.09	62	689	644	NA
[64]	NP11	Illumina	96 X	20.2	NA	17,814	34	8171 a
[38]	CECT1137	Illumina, Roche 454 FLX	NA	20.45	61.9	NA	62	8206 a
[43]	ATCC 204091	Illumina, Roche 454 FLX, Sanger	NA	20.48	61.9	186	29	3359 b
[62]	IFO0559	Illumina	170 X	20.28	NA	NA	246	8100 c
[62]	IFO0880	Illumina	139 X	20.36	NA	NA	219	7920 c
[23]	ATCC 10788	Illumina	241 X	20.75	62.01	NA	61	7730 a
[23]	ATCC 10657	Illumina	233 X	21.49	61.81	NA	137	7800 a
[47]	CGMCC 2.1609	Illumina, Roche 454 FLX	13 X	33.4	61.9	868	365	9820
[55]	VN1	Illumina	125 X	20.02	61.8	424	NA	8021 a
This study	CBS 14	Illumina, Oxford Nanopore	1000 X	20.53	61.83	23	3	9464 a

NA, not available; a, b and c refer to protein-coding sequences, putative genes and predicted proteins, respectively.

Table 2
Characteristics of the assembled contigs and scaffolds.

Contig/ scaffold	Size (bp)	GC content (%)	Read depth	No. telomeric regions
contig_1	666,847	62.12	21	1
contig_19	763,076	61.61	20	0
contig_20	906,560	61.88	21	2
contig_28	659,395	61.82	21	1
contig_29	785,970	61.85	19	0
contig_30	1,712,111	62.15	21	1
contig_31	5778	39.10	0	1
contig_33	1,538,587	62.03	25	1
contig_34	1,529,952	62.08	21	1
contig_35	930,820	61.92	22	1
contig_36	1,965,970	62.07	21	0
contig_40	309,449	61.55	17	1
contig_43	618,889	62.24	20	0
contig_49	63,359	62.14	14	1
contig_5	577,483	61.83	20	1
contig_53	670,355	61.86	20	2
contig_58	477,283	62.04	22	0
contig_60	668,592	62.28	22	1
contig_63	11,496	61.73	63	1
contig_64	152,705	62.29	20	1
contig_65	265,355	60.97	20	0
contig_67	268,355	62.24	20	0
contig_8	157,597	40.90	438	0
scaffold_27	1,925,638	61.87	22	1
scaffold_3	1,219,652	62.22	21	0
scaffold_32	1,683,583	62.10	22	1

can particularly result in longer but also shorter contigs that miss parts of the actual sequence at the end of a contig [24]. The genes annotated in contig 63 are *UTP22*, *H2A* and *H2B* which are encoding for RNA-associated protein 22, Histone H2A and Histone H2B, respectively (Supplementary Table S3). A truncated copy of *UTP22* from *R. toruloides* NP11 was also annotated in contig 63 and identified through Blast search (Supplementary Table S3). The CDS of *UTP22*, *H2A* and *H2B* are also found in scaffold 32 sharing a pairwise identity of 99.8%, 96.7%, and 98.5%, respectively (Supplementary Figs. S2, S3). *H2B* and *UTP22* are localized next to each other, *H2A* is localized elsewhere on scaffold 32 (Supplementary Fig. S3). Both of the *UTP22* genes encoded in contig 63 are shortened compared to the *UTP22* gene that we found encoded on scaffold 32. The *UTP22* gene on scaffold 32 shares 100% sequence identity with the *UTP22* gene identified in strain NP11. In addition to scaffold 32, the CDS of *H2B* aligned partially to contig_29 and to scaffold 30 sharing a pairwise identity of 88% (Supplementary Fig. S2). Five transcripts were annotated within the circular contig 63 (Supplementary Table S3). All of them were expressed under the experimental conditions. Extrachromosomal endogenous DNA have been previously found in *Saccharomyces cerevisiae* [39,52]. The presence of DNA mitochondrial plasmids in filamentous fungi, including some Basidiomycota species, have also been widely acknowledged before ([21]; [65]; [12]). Within them, mitochondrial circular DNA plasmids are less frequently found

than linear. They are typically within a length range of 2.5–5 kb, and encoding enzymes involved in their replication such as a DNA polymerase or a reverse transcriptase ([12]; [65]). In *S. cerevisiae* DNA replication initiate at certain autonomously replicating sequences (ARS) [17]. An ARS consensus sequence (ACS), or ACS like pattern, was not found in contig 63. AT-rich ACS elements of ARS have been found to be degenerate and in *Schizosaccharymces pombe*, which is closer related to *S. cerevisiae* than *R. toruloides*, the absence of such consensus sequence has been reported [17,25]. This is the first time that such extrachromosomal circular DNA has been detected in Basidiomycetes. The replication origin of such structures may be useful for developing efficient episomal vectors for the manipulation of *Rhodotorula* yeasts.

Analysis of the RNA-Seq data of *R. toruloides* CBS 14 using StringTie software and guided by the MetaEuk annotation (based on all protein sequences from MycoCosm) resulted in the annotation of 9464 genes with 11,691 transcripts. This number is significantly higher than previously reported for *R. toruloides* genomes (Table 1). The higher number of transcripts than protein-coding genes can be explained by alternative splicing and non-coding RNAs. Zhu et al. [64] found 1371 genes encoding two or more transcript isoforms in *R. toruloides* NP11. The average number of exons per gene is 5.95 in *R. toruloides* CBS 14 (supplementary table 4). Confirming the findings of Zhu et al. [64], we observed a predominance of split genes in the genome, with a total of 8550 for *R. toruloides* CBS 14. After functional annotation of coding sequences through Gene Ontology (GO) terms we clustered annotations into the categories: biological processes, cellular components and molecular functions (Supplementary Figs. S4–S6). Genes encoding enzymes crucial for lipid and carotenoid metabolisms such as *CDC19*, *MAE1*, *MAE2*, *ACL1*, *ACC1*, *FAS1*, *FAS2*, *OLE1*, *ACAD10*, *ACAD11*, *crtYB*, *crtI* and *BTS1* are present in the genome (supplementary table s5). Similar to the situation in *S. cerevisiae* [7], the genes encoding the α- and β-subunits of FAS are located in different chromosomes (Supplementary Table S5).

The MetaEuk annotation identified a gene (*FAS21*) on the opposite strand of the *FAS2*- gene (Supplementary Fig. S7), which encodes the α-subunit of the fatty acid synthase complex (FAS) in *R. toruloides* NP11 [64]. *FAS21* contains 2 exons and its mRNA would be complementary to parts of the *FAS2*-sequence. There is a growing number of natural antisense transcripts identified in fungal transcriptome analyses [18,19]. *FAS21* transcript might be involved in controlling *FAS2* expression. However, we did not identify a *FAS21* transcript under the experimental conditions carried out, thus the involvement of *FAS21* in the regulation of *FAS2*- expression and thus fatty acid synthesis still needs to be verified. Three different transcripts of *FAS2* were identified containing 3, 12 and 16 exons, respectively (Supplementary Table S6).

BUSCO orthologs analysis showed that 96.9% of the assessed genes were identified and complete (96.6% single-copy and 0.3% duplicated), 1.7% fragmented and only 1.4% were lacking, demonstrating the high quality of the hybrid genome assembly (Supplementary Fig. S8).

In the case of the transcriptomic data, 84.1% of the genes were indicated as complete (69.3% single copy,14.8% duplicated), 14.1% as

fragmented and 1.8% genes as missing.

The genome annotation and assembly correspond to the following taxonomic classification: Eukaryota superkingdom, Basidiomycota phylum, Microbotryomycetes class, Sporidiobolales order, Sporidiobolaceae family, *Rhodotorula* genus, *Rhodotorula toruloides* species, CBS 14 strain.

Acknowledgements

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.10.006.

References

[2] L. Anand, ChromoMap: an R package for interactive visualization and annotation of chromosomes, BioRxiv 605600 (2019), https://doi.org/10.1101/605600.

[3] S. Andrews, et al., FastQC: a quality control tool for high throughput sequence data, in: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[5] J. Blomqvist, et al., Oleaginous yeast as a component in fish feed, Sci. Rep. 8 (1) (2018) 1–8, https://doi.org/10.1038/s41598-018-34232-x.

[6] C. Brandt, E. Bongcam-Rudloff, B. Müller, Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants, Appl. Sci. 10 (21) (2020) 1–14, https://doi.org/10.3390/app10217518.

[7] G. Burkl, H. Castorph, E. Schweizer, Mapping of a complex gene locus coding for part of the *Saccharomyces cerevisiae* fatty acid synthetase multienzyme complex, MGG Mol. Gen. Genet. 119 (4) (1972) 315–322, https://doi.org/10.1007/BF00272089.

[8] B. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner. United States. https://www.osti.gov/servlets/purl/1241166, 2014.

[9] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) i884–i890, https://doi.org/10.1093/bioinformatics/bty560.

[10] M. Chmielarz, et al., FT-NIR: a tool for rapid intracellular lipid quantification in oleaginous yeasts, Biotechnol. Biofuels 12 (1) (2019) 1–9, https://doi.org/10.1186/s13068-019-1513-9.

[11] M. Chmielarz, J. Blomqvist, S. Sampels, M. Sandgren, V. Passoth, Microbial lipid production from crude glycerol and hemicellulosic hydrolysate with oleaginous yeasts, Biotechnol. Biofuels 14 (1) (2021) 1–12, https://doi.org/10.1186/s13068-021-01916-y.

[12] P. Cahan, J.C. Kennell, Identification and distribution of sequences having similarity to mitochondrial genomes of filamentous fungi, Mol. Gen. Genomics. 273 (6) (2005) 462–473, https://doi.org/10.1007/s00438-005-1133-x.

[13] S.T. Coradetti, et al., Functional genomics of lipid metabolism in the oleaginous yeast *Rhodosporidium toruloides*, ELife 7 (2018) 1–55, https://doi.org/10.7554/eLife.32110.

[14] J.W. Davey, S.J. Davis, J.C. Mottram, P.D. Ashton, Tapestry: validate and edit small eukaryotic genome assemblies with long reads, BioRxiv (2020), https://doi.org/10.1101/2020.04.24.059402.

[15] W. De Coster, S. D'Hert, D.T. Schultz, M. Cruts, C. Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, Bioinformatics 34 (15) (2018) 2666–2669, https://doi.org/10.1093/bioinformatics/bty149.

[16] P. De Jonge, F.C.M. De Jongh, R. Meijers, H.Y. Steensma, W.A. Scheffers, Orthogonal-field-alternation gel electrophoresis banding patterns of DNA from yeasts, Yeast 2 (3) (1986) 193–204, https://doi.org/10.1002/yea.320020307.

[17] M.K. Dhar, S. Sehgal, S. Kaul, Structure, replication efficiency and fragility of yeast *ARS* elements, Res. Microbiol. 163 (4) (2012) 243–253, https://doi.org/10.1016/j.resmic.2012.03.003.

[18] M.E. Donaldson, B.J. Saville, Natural antisense transcripts in fungi, Mol. Microbiol. 85 (3) (2012) 405–417, https://doi.org/10.1111/j.1365-2958.2012.08125.x.

[19] N. Chacko, X. Lin, Non-coding RNAs in the development and pathogenesis of eukaryotic microbes, Appl. Microbiol. Biotechnol. 97 (18) (2013) 7989–7997, https://doi.org/10.1007/s00253-013-5160-y.

[20] Y. González-García, et al., High lipids accumulation in *Rhodosporidium toruloides* by applying single and multiple nutrients limitation in a simple chemically defined medium, Ann. Microbiol. 67 (8) (2017) 519–527, https://doi.org/10.1007/s13213-017-1282-2.

[21] A.J.F. Griffiths, Natural plasmids of filamentous fungi, Microbiol. Rev. 59 (4) (1995) 673–685, https://doi.org/10.1128/mmbr.59.4.673-685.1995.

[22] R.S. Harris, Improved Pairwise Alignment of Genomic DNA, Ph.D. Thesis, The Pennsylvania State University, 2007.

[23] J. Hu, L. Ji, Draft genome sequences of *Rhodosporidium toruloides* strains ATCC 10788 and ATCC 10657 with compatible mating types, Genome Announc. 4 (2) (2016) 1083–1084, https://doi.org/10.1128/genomeA.00098-16.

[24] M. Hunt, et al., Circlator: automated circularization of genome assemblies using long sequencing reads, Genome Biol. 16 (1) (2015) 1–10, https://doi.org/10.1186/s13059-015-0849-0.

[25] R. Iwakiri, S. Eguchi, Y. Noda, H. Adachi, K. Yoda, Isolation and structural analysis of efficient autonomously replicating sequences (*ARSs*) of the yeast *Candida utilis*, Yeast 22 (13) (2005) 1049–1060, https://doi.org/10.1002/yea.1296.

[26] H.S.J. Lopes, N. Bonturi, E.J. Kerkhoven, E.A. Miranda, P. Lahtvee, C/N ratio and carbon source-dependent lipid production profiling in *Rhodotorula toruloides*, Appl. Microbiol. Biotechnol. 104 (6) (2020) 2639–2649, https://doi.org/10.1007/s00253-020-10386-5.

[27] E.L. Karin, M. Mirdita, J. Söding, MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics, Microbiome 8 (48) (2020), https://doi.org/10.1186/s40168-020-00808-x.

[28] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (4) (2015) 357–360, https://doi.org/10.1038/nmeth.3317.

[29] Kitson, E. (n.d.). SimpleCircularise. https://github.com/Kzra/Simple-Circularise.

[30] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, Nat. Biotechnol. 37 (5) (2019) 540–546, https://doi.org/10.1038/s41587-019-0072-8.

[31] E. Kopylova, L. Noé, H. Touzet, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data, Bioinformatics 28 (24) (2012) 3211–3217, https://doi.org/10.1093/bioinformatics/bts611.

[32] S. Kovaka, et al., Transcriptome assembly from long-read RNA-seq alignments with StringTie2, Genome Biol. 20 (1) (2019) 1–13, https://doi.org/10.1186/s13059-019-1910-1.

[33] S. Kumar, H. Kushwaha, A.K. Bachhawat, G.P.S. Raghava, K. Ganesan, Genome sequence of the oleaginous red yeast *Rhodosporidium toruloides* MTCC 457, Eukaryot. Cell 11 (8) (2012) 1083–1084, https://doi.org/10.1128/EC.00156-12.

[34] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics 34 (18) (2018) 3094–3100, https://doi.org/10.1093/bioinformatics/bty191.

[35] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows—Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760, https://doi.org/10.1093/bioinformatics/btp324.

[36] H. Li, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

[37] S. Manzoor, E. Bongcam-Rudloff, A. Schnürer, B. Müller, Genome-guided analysis and whole transcriptome profiling of the mesophilic syntrophic acetate oxidising bacterium *Syntrophaceticus schinkii*, PLoS One 11 (11) (2016) 1–24, https://doi.org/10.1371/journal.pone.0166520.

[38] N. Morin, et al., Draft genome sequence of *Rhodosporidium toruloides* CECT1137, an oleaginous yeast of biotechnological interest, Genome Announc. 2 (4) (2014) 578–579, https://doi.org/10.1128/genomeA.00641-14.

[39] H.D. Møller, L. Parsons, T.S. Jørgensen, D. Botstein, B. Regenberg, Extrachromosomal circular DNA is common in yeast, Proc. Natl. Acad. Sci. U. S. A. 112 (24) (2015) E3114–E3122, https://doi.org/10.1073/pnas.1508825112.

[40] D. Nishimura, RepeatMasker. Biotech Software & Internet Report 1(1–2), 2000, pp. 36–39, https://doi.org/10.1089/152791600319259.

[41] R.A. Olsen, et al., De novo assembly of *Dekkera bruxellensis*: a multi technology approach using short and long-read sequencing and optical mapping, GigaScience 4 (1) (2015), https://doi.org/10.1186/s13742-015-0094-1.

[42] Y.K. Park, J.M. Nicaud, R. Ledesma-Amaro, The engineering potential of *Rhodosporidium toruloides* as a workhorse for biotechnological applications, Trends Biotechnol. 36 (3) (2018) 304–317, https://doi.org/10.1016/j.tibtech.2017.10.013.

[43] D. Paul, et al., Genome sequence of the oleaginous yeast *Rhodotorula glutinis* ATCC, Genome Announc. 2 (1) (2014) 2009–2010, https://doi.org/10.1128/genomeA.00046-14.

[44] M. Pertea, et al., StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol. 33 (3) (2015) 290, https://doi.org/10.1038/nbt.3122.

[45] H.W. Pi, et al., Engineering the oleaginous red yeast *Rhodotorula glutinis* for simultaneous β-carotene and cellulase production, Sci. Rep. 8 (1) (2018) 2–11, https://doi.org/10.1038/s41598-018-29194-z.

[46] N.T. Pierce, L. Irber, T. Reiter, P. Brooks, C.T. Brown, Large-scale sequence comparisons with sourmash, F1000 Res. 8 (2019) 1006, https://doi.org/10.12688/f1000research.19675.1.

[47] C. Sambles, et al., Genome sequence of the oleaginous yeast *Rhodotorula toruloides* strain CGMCC 2.1609, Genomics Data 13 (2017) 1–2, https://doi.org/10.1016/j.gdata.2017.05.009.

[48] V. Sànchez Nogué, et al., Integrated diesel production from lignocellulosic sugars: via oleaginous yeast, Green Chem. 20 (18) (2018) 4349–4365, https://doi.org/10.1039/c8gc01905c.

[49] Scott, C. (n.d.). dammit: An Open and Accessible De Novo Transcriptome Annotator. http://www.camillescott.org/dammit.

[50] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy

orthologs, Bioinformatics 31 (19) (2015) 3210–3212, https://doi.org/10.1093/bioinformatics/btv351.

[51] G. Singh, S. Sinha, K.K. Bandyopadhyay, M. Lawrence, D. Paul, Triauxic growth of an oleaginous red yeast *Rhodosporidium toruloides* on waste "extract" for enhanced and concomitant lipid and β-carotene production, Microb. Cell Factories 17 (1) (2018) 1–10, https://doi.org/10.1186/s12934-018-1026-4.

[52] P.K. Strope, et al., 2μ plasmid in *Saccharomyces* species and in *Saccharomyces cerevisiae*, FEMS Yeast Res. 15 (8) (2015) 1–8, https://doi.org/10.1093/femsyr/fov090.

[53] Technologies, O. N. (n.d.). Medaka: Long-read Polishing Pipeline. https://github.com/nanoporetech/medaka.

[54] I.A. Tiukova, et al., Chromosomal genome assembly of the ethanol production strain CBS 11270 indicates a highly dynamic genome structure in the yeast species *Brettanomyces bruxellensis*, PLoS One 14 (5) (2019) 1–20, https://doi.org/10.1371/journal.pone.0215077.

[55] T.N. Tran, D.H. Ngo, N.T. Nguyen, D.N. Ngo, Draft genome sequence data of *Rhodosporidium toruloides* VN1, a strain capable of producing natural astaxanthin, Data Brief 26 (2019) 104443, https://doi.org/10.1016/j.dib.2019.104443.

[56] R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate *de novo* genome assembly from long uncorrected reads, Genome Res. 27 (5) (2017) 737–746, https://doi.org/10.1101/gr.214270.116.

[57] B.J. Walker, et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS One 9 (11) (2014), https://doi.org/10.1371/journal.pone.0112963.

[58] R.M. Waterhouse, et al., BUSCO applications from quality assessments to gene prediction and phylogenomics, Mol. Biol. Evol. 35 (3) (2018) 543–548, https://doi.org/10.1093/molbev/msx319.

[59] Wick, R. (n.d.). Filtlong: Quality Filtering Tool for Long Reads. https://github.com/rrwick/Filtlong.

[60] R.R. Wick, M.B. Schultz, J. Zobel, K.E. Holt, Bandage: interactive visualization of de novo genome assemblies, Bioinformatics 31 (20) (2015) 3350–3352, https://doi.org/10.1093/bioinformatics/btv383.

[61] M.G. Wiebe, K. Koivuranta, M. Penttilä, L. Ruohonen, Lipid production in batch and fed-batch cultures of *Rhodosporidium toruloides* from 5 and 6 carbon carbohydrates, BMC Biotechnol. 12 (2012), https://doi.org/10.1186/1472-6750-12-26.

[62] S. Zhang, et al., Engineering *Rhodosporidium toruloides* for increased lipid production, Biotechnol. Bioeng. 113 (5) (2016) 1056–1066, https://doi.org/10.1002/bit.25864.

[63] R. Zhou, Z. Zhu, S. Zhang, Z.K. Zhao, The complete mitochondrial genome of the lipid-producing yeast *Rhodotorula toruloides*, FEMS Yeast Res. 20 (6) (2020), https://doi.org/10.1093/femsyr/foaa048.

[64] Z. Zhu, et al., A multi-omic map of the lipid-producing yeast *Rhodosporidium toruloides*, Nat. Commun. 3 (2012) 1111–1112, https://doi.org/10.1038/ncomms2112.

[65] Y. Wang, F. Zeng, C.C. Hon, Y. Zhang, F.C.C. Leung, The mitochondrial genome of the Basidiomycete fungus Pleurotus ostreatus (oyster mushroom), FEMS Microbiol. Lett. 280 (1) (2008) 34–41, https://doi.org/10.1111/j.1574-6968.2007.01048.x.