Type of the Paper (Article, Review, Communication, etc.)

Near chromosome-level genome assembly and annotation of *Rhodotorula babjevae* strains reveals high intraspecific divergence

Giselle C. Martín-Hernández[1], Bettina Müller[1], Christian Brandt[2,3], Martin Hölzer[2], Adrian Viehweger[2,4] and Volkmar Passoth[1,*]

[1] Department of Molecular Sciences, Swedish University of Agricultural Sciences, 75007 Uppsala, Sweden; Giselle.Martin@slu.se (G.C.M.-H); Bettina.Muller@slu.se (B.M.)
[2] nanozoo GmbH, Leipzig, Germany; christian@nanozoo.com (C.B.); martin@nanozoo.com (M.H.); adrian@nanozoo.com (A.V.)
[3] Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany; christian.brandt@med.uni-jena.de
[4] Institute of Medical Microbiology and Virology, University Hospital Leipzig, Germany; adrian.viehweger@medizin.uni-leipzig.de
* Correspondence: Volkmar.Passoth@slu.se; Tel.: +4618673380
G.C.M.-H. and B.M. equally contributed to this paper and shall both be regarded as first authors.

Abstract: The genus *Rhodotorula* includes basidiomycetous oleaginous yeast species. *R. babjevae* can produce compounds of biotechnological interest such as lipids, carotenoids and biosurfactants from low value substrates such as lignocellulose hydrolysate. High-quality genome assemblies are needed to develop genetic tools and to understand fungal evolution and genetics. Here, we combined short- and long-read sequencing to resolve the genomes of two *R. babjevae* strains, CBS 7808 (type strain) and DBVPG 8058 at chromosomal level. Both genomes have a size of 21 Mbp and a GC content of 68.2%. Allele frequency analysis indicated tetraploidy in both strains. They harbor 21 putative chromosomes with sizes ranging from 0.4 to 2.4 Mb. In both assemblies, the mitochondrial genome was recovered in a single contig, which shared 97% pairwise identity. The pairwise identity between the majority of chromosomes ranges from 82% to 87%. We found indications for strain-specific extrachromosomal endogenous DNA. 7,591 protein-coding genes and 7,607 associated transcripts were annotated in CBS 7808 and 7,481 protein-coding genes and 7,516 associated transcripts in DBVPG 8058. CBS 7808 has accumulated a higher number of tandem duplications than DBVPG 8058. We identified large translocation events between putative chromosomes and a high genetic divergence between the two strains.

Keywords: *Rhodotorula babjevae; de-novo* hybrid assembly; Nanopore sequencing; genome divergence

1. Introduction

Oleaginous yeasts have received considerable attention in recent years due to a high number of potential biotechnological applications of microbial lipids. *Rhodotorula* species are basidiomycetous oleaginous yeasts whose lipid production has been accounted for higher than 70 % of dry cell weight. They showed high tolerance to inhibitors, enabling them to convert lignocellulosic hydrolysates into lipids [1–4]. Microbial lipids from *R. babjevae* and other oleaginous yeasts have a similar fatty acid composition as vegetable oils, representing an environmental and ethically suitable alternative raw material for the production of biofuels, oleochemicals, feed, and food additives [2,5,6]. Under nitrogen-limited conditions *R. babjevae* can simultaneously accumulate biotechnologically important enzymes, glycolipids, and carotenoids [5]. Glycolipids from *R. babjevae* have promising environmental applications in the biodegradation of hydrocarbon pollutants and in replacing synthetic compounds and chemical surfactants [7–9]. They are also attractive for further applications in different industrial sectors due to antifungal, antibacterial, antiviral and anti-carcinogenic activities [7–10]. However, obtaining more robust *R. babjevae* strains is desirable to overcome the high production costs of microbial lipids and biosurfactants.

There are currently no described methods for the molecular manipulation of *R. babjevae* strains. A high-quality genome assembly is needed for the development of genetic tools for *R. babjevae* and to deepen our understanding of the biology and evolution of the species. Combinations of short- and long-reads have been previously found to accomplish high-quality genome hybrid assemblies in terms of completeness, contiguity, and chromosome reconstruction [3,11–13]. Here we present for the first time *de novo* genome assemblies and annotations of two strains from *R. babjevae* species, CBS 7808 (type strain) and DBVPG 8058, by combining short- and long-read sequencing technologies. We also provide a genome divergence analysis between both *R. babjevae* strains.

2. Materials and Methods

2.1. Yeast strains

R. babjevae type strain CBS 7808 was obtained from the CBS-KNAW collection (Utrecht, the Netherlands). *R. babjevae* strain DBVPG 8058 was isolated and identified at SLU Uppsala (strain number in the strain collection of the Department of Molecular Sciences J195) [2] and deposited in the Industrial Yeasts Collection (Perugia, Italy).

2.2. DNA purification

The yeasts were cultivated on 50 mL YPD until reaching exponential growth phase [14]. Cell wall degradation was performed according to [15] with some modifications. Briefly, the cells were suspended in SCEM pH 5.8 (1M Sorbitol, 0.01 M EDTA, 0.03 M β-mercaptoethanol, 0.1 M sodium citrate) after harvesting. Lyticase solution was added to the cell suspensions (100 U/ml) of CBS 7808 and DBVPG 8058, which were then incubated for 9 h or overnight, respectively. After Lyticase digestion, cells were harvested at 3000 rpm, suspended in SCEM buffer, and incubated overnight with Zymolyase (200 U/mL). Genomic DNA extraction from protoplasts was performed using the NucleoBond® CB 20 Kit (Macherey-Nagel, Germany). DNA concentration, purity, and quality were confirmed through Qubit™ 4 Fluorometer (Thermo Fisher Scientific, Singapore), NanoDrop® ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA), and agarose gel electrophoresis, respectively.

2.3. Library preparation and sequencing

The extracted DNA samples were sequenced using MinION (Oxford Nanopore Technologies) and Illumina sequencing platforms. Nanopore DNA libraries were prepared according to [16]. Briefly, 31.5 μL of AMPure magnetic beads were added to 5 μg of DNA for a "pre-cleaning" step. Library preparation was then performed according to a modified protocol [16] using a Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK). Each DNA library was loaded onto a FLO-MIN106 flow cell mounted on a MINION device (Oxford Nanopore Technologies). MinKNOW software (version 19.06.8) was used for sequencing as described by [16]. The basecalling was run using Guppy version 3.2.4-1--195590e and model HAC-mod (modified base sensitive high accuracy model).

From the 6,665,174 long reads recovered from the CBS 7808 DNA library, the mean read length was 2,789.7 bases and the read length N50 5,553 bases yielding a total of 18,593 Mbp sequenced. For DBVPG 8058, 2,953,255 long reads were retrieved containing a total of 15,702 Mbp sequenced. The mean read length was 5,317 bases and the read length N50 7,411 bases. Aliquots of the extracted DNA from both *R. babjevae* strains were also subjected to short-read paired-end sequencing using the Illumina Novaseq platform (S prime, 2x 150 bp) and the TruSeq PCR free DNA library preparation kit (Illumina Inc.). 179,163,622 short reads were recovered from CBS 7808 DNA library, corresponding to a total of 27,053 Mbp sequenced. For DBVPG 8058, 203,873,550 short reads were retrieved containing a total of 30,784 Mbp sequenced.

2.4. Genome assembly and annotation

R. babjevae genome assembly and annotation was performed using a custom pipeline described elsewhere [3], applying the program versions listed in Table S1. To further improve the annotation of transcripts and exon-intron boundaries, we additionally mapped RNA-Seq data from the closely related *R. toruloides* CBS 14 (PRJEB40807) to the *R. babjevae* genomes like also shown before [3]. nQuire (v0.0) was used for estimating the ploidy level of the *R. babjevae* strains [17]. Genomics data were visualized using Circa (http://omgenomics.com/circa).

The reconstruction of lipid metabolic pathway maps was performed using KEGG Mapper version 4.3. The KEGG Orthology (KO) identifiers were affiliated to the annotated transcripts of *R. babjevae* CBS 7808 and *R. babjevae* DBVPG 8058 using KofamKOALA [18] with an e-value cut-off of 0.01.

2.5. Genome divergence analysis

Synteny relationship analysis between *R. babjevae* CBS 7808 and *R. babjevae* DBVPG 8058 was performed using NUCmer (MUMmer, version 3.23). The maximum gap between adjacent matches in a cluster was set to 500 and the minimum cluster length to 100. Visualization of NUCmer alignments was done through Circa.

The level of sequence divergence between both *R. babjevae* strains as well as with other closely related *Rhodotorula* species, including *R. glutinis* ZHK (JAAGPT010000000.1), *R. graminis* WP1 (JTAO00000000.1) and *R. toruloides* strains CBS 14 (PRJEB40807), CGMCC 2.1609 (LKER00000000.1), VN1 (SJTE00000000.1) and NBRC 0880 (LCTV00000000.2), was evaluated using the alignment-free distance measure *kr* [19]. We calculated Average Nucleotide Identity (ANI) values using the web-based calculator available at Kostas Lab [20]. DNA–DNA homology (DDH) was estimated using the Genome-to-Genome Distance Calculator (GGDC) 2.1 (http://ggdc.dsmz.de/distcalc2.php) with GBDP2_MUMMER program [21].

Whole genome alignments of *R. babjevae* strains were performed using LASTZ (version 7.0.2) implemented in Geneious prime, version 2021.0.1 (Biomatters Ltd.) [22]. Nucleotide alignment and phylogenetic tree construction using MAFFT v7.450 [23] and PhyML 3.3.20180621 [24] with 100 bootstraps, respectively, were performed on Geneious prime platform too.

We applied OrthoVenn2 web-based platform (https://orthovenn2.bioinfotoolkits.net) for whole genome comparison and identification of orthologous gene clusters and paralogous genes in the strains CBS 7808 and DBVPG 8058, respectively, using 1e-15 threshold e-value and 1.5 inflation [25]. In order to identify duplicated genes (paralogs) with high sequence identity an all-against-all sequence identity search was performed on NCBI Genome Workbench version 3.7.0 [26] using BLASTp (BLOSUM62 matrix) and a cut-off e-value of 1e-15. The output was filtered for a minimum of 70% coverage and 70% sequence identity.

3. Results and Discussion

3.1. Genome assembly, ploidy estimation and gene annotation of R. babjevae strains

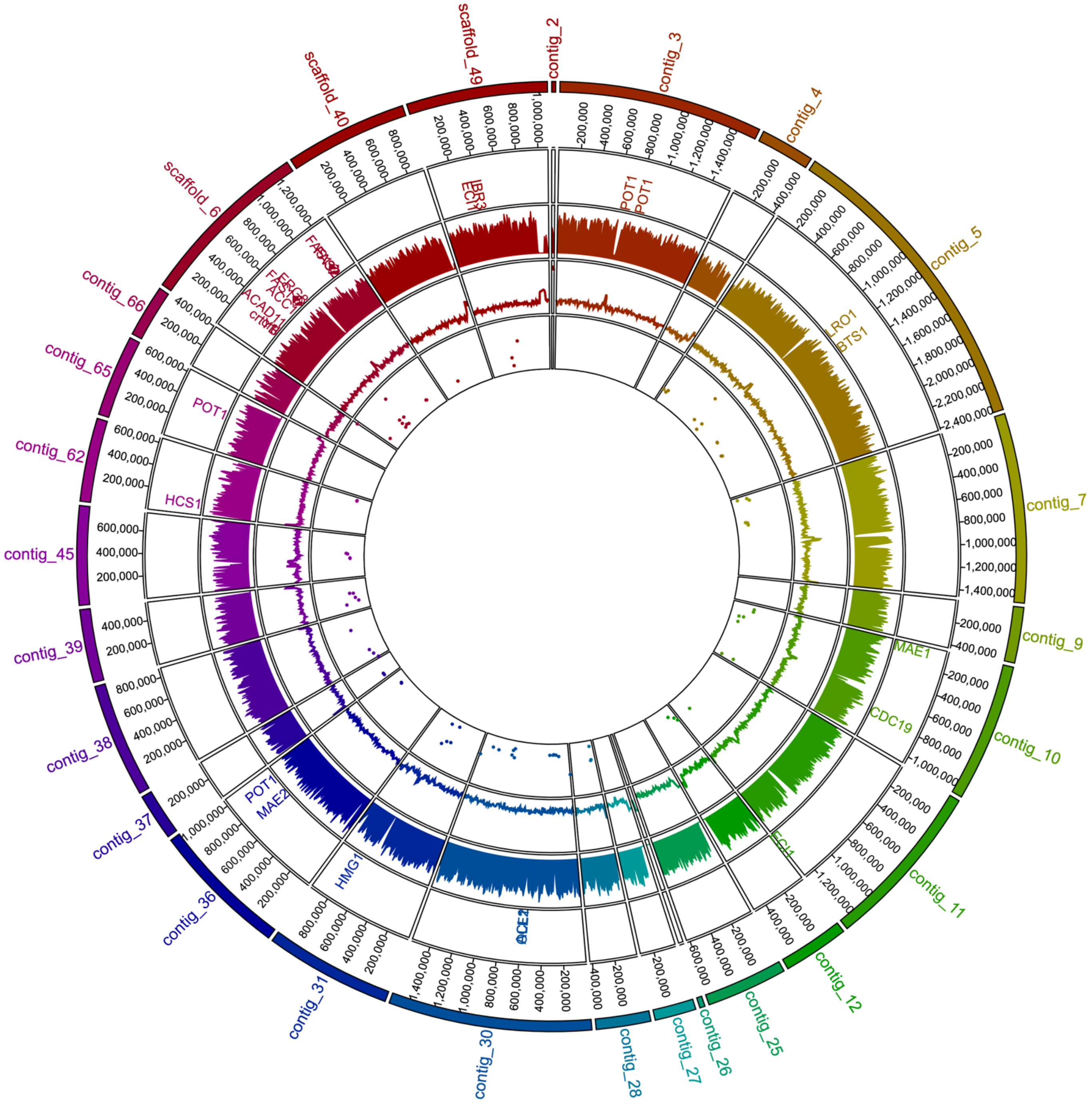
The genome of both *R. babjevae* strains was assembled by a combined approach of long- and short-read sequencing with a coverage depth of about 2000 X. A summary of the genomic data is presented in Table 1. *R. babjevae* CBS 7808 draft genome has a total size of 21,862,387 bp and a GC content of 68.23%. Repetitive sequences represent 5.93% of the total length of the genome, from which 4.98% are single repeats and 0.96% low complexity regions. The draft genome of DBVPG 8058 has a total size of 21,522,072 bp and a GC content of 68.24%. The approach identified 6.73% as repetitive sequences, including 5.65% as single repeats and 1.09% as low complexity regions. Genome features such as genome size, GC content and percentage of repetitive regions are highly similar between both strains validating that they are closely related. The genome size is comparable to that of other *Rhodotorula* species, but the GC content is slightly higher [3,12,27–29] (Table 1).

Table 1. Genomic data from *Rhodotorula* species.

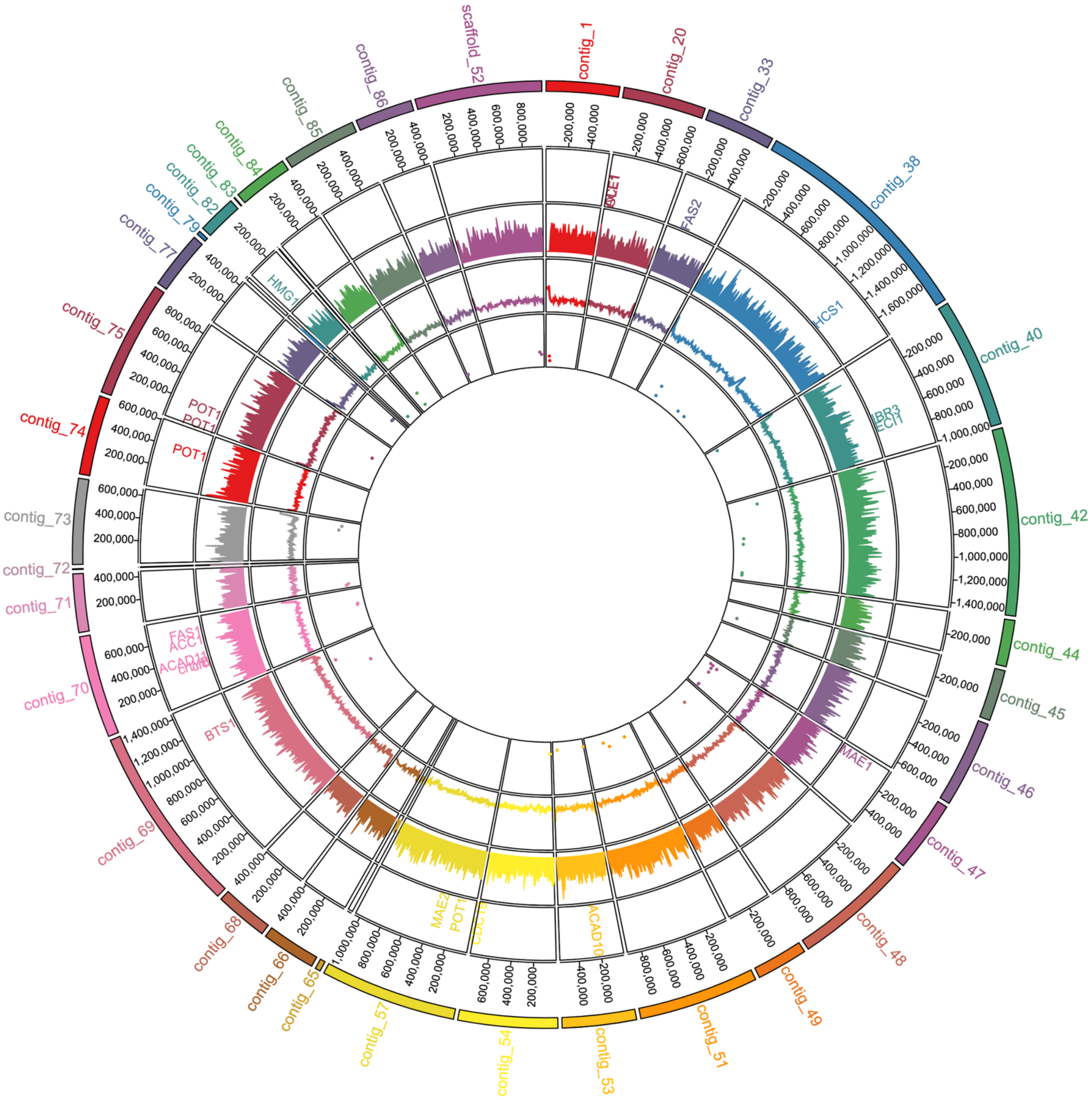
Reference	This study	This study	[28]	[30]	[3]	[29]
Strain number	*R. babjevae* CBS 7808	*R. babjevae* DBVPG 8058	*R. graminis* WP1	*R. glutinis* ZHK	*R. toruloides* CBS 14	*R. toruloides* NP11
Genome size (Mb)	21.9	21.5	21.0	21.8	20.5	20.2
Coverage	2,058	2,122	8.6	470	1,514	96
GC content (%)	68.23	68.24	67.76	67.8	61.83	62.05
bases masked (%)	5.93	6.73	6.5	NA	2.01	2.53
No. Scaffolds	3	1	26	30	3	34
No. Contigs	24	33	325	NA	23	NA
Protein-coding genes	7,591	7,481	7,283 a	6,774 a	9,464	8,171
Avg. no. exons per gene	4.0	3.9	6.2	NA	5.9	NA
Sequencing platform	Nanopore and Illumina	Nanopore and Illumina	Sanger	PacBio & Illumina	Nanopore and Illumina	Illumina and Sanger

NA - not available; a - refer to predicted genes.

Sequence assembly resulted for *R. babjevae* CBS 7808 in 24 contigs and 3 scaffolds with a length N50 of 1,067,634 bp (Figure 1A, Table S2). A telomeric region was predicted at one of the termini for 13 contigs and scaffolds larger than 250,000 bp. The draft genome of strain DBVPG 8058 consists of 33 contigs and one scaffold with a length N50 of 789,767 bp (Figure 1B, Table S3). From the contigs and scaffolds with sizes larger than 250,000 bp in DBVPG 8058 genome assembly, two have telomere sequences at both termini and 15 at one terminus each. The low numbers of contigs and scaffolds in the genome assemblies from both *R. babjevae* strains indicate high accuracy, contiguity and completeness. Two putative circular sequences were identified in each strain. Among them, contig_2 in CBS 7808 and contig_79 in DBVPG 8058 contained the mitochondrial genes. Both mitochondrial genomes are similar in size with 30.876 bp and 28.432 bp, respectively, and have a GC content of 38.9%. (Table S2, S3).


scaffold_40
scaffold_49
contig_2
contig_3
contig_4
contig_5
contig_7
contig_9
contig_10
contig_11
contig_12
contig_25
contig_26
contig_27
contig_28
contig_30
contig_31
contig_36
contig_37
contig_38
contig_39
contig_45
contig_62
contig_65
contig_66
scaffold_6
IBR3
ECH
POT1
POT1
LRO1
BTS1
MAE1
CDC19
ECH
ACE2
HMG1
POT1
MAE2
HCS1
POT1
FAS1
FAS2
ERG8
ACC1
ACAD11
crtYB

(a)

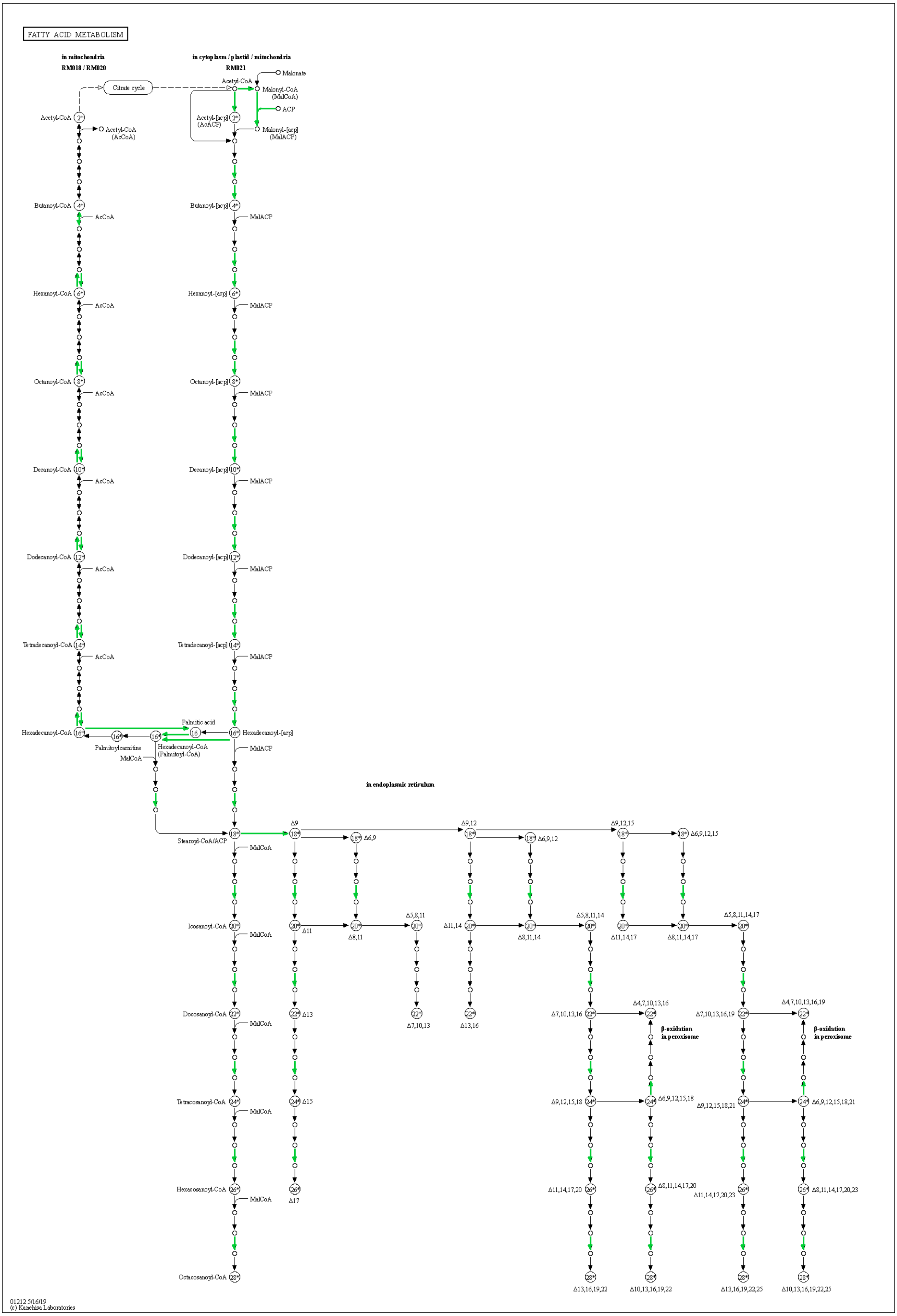


(**b**)

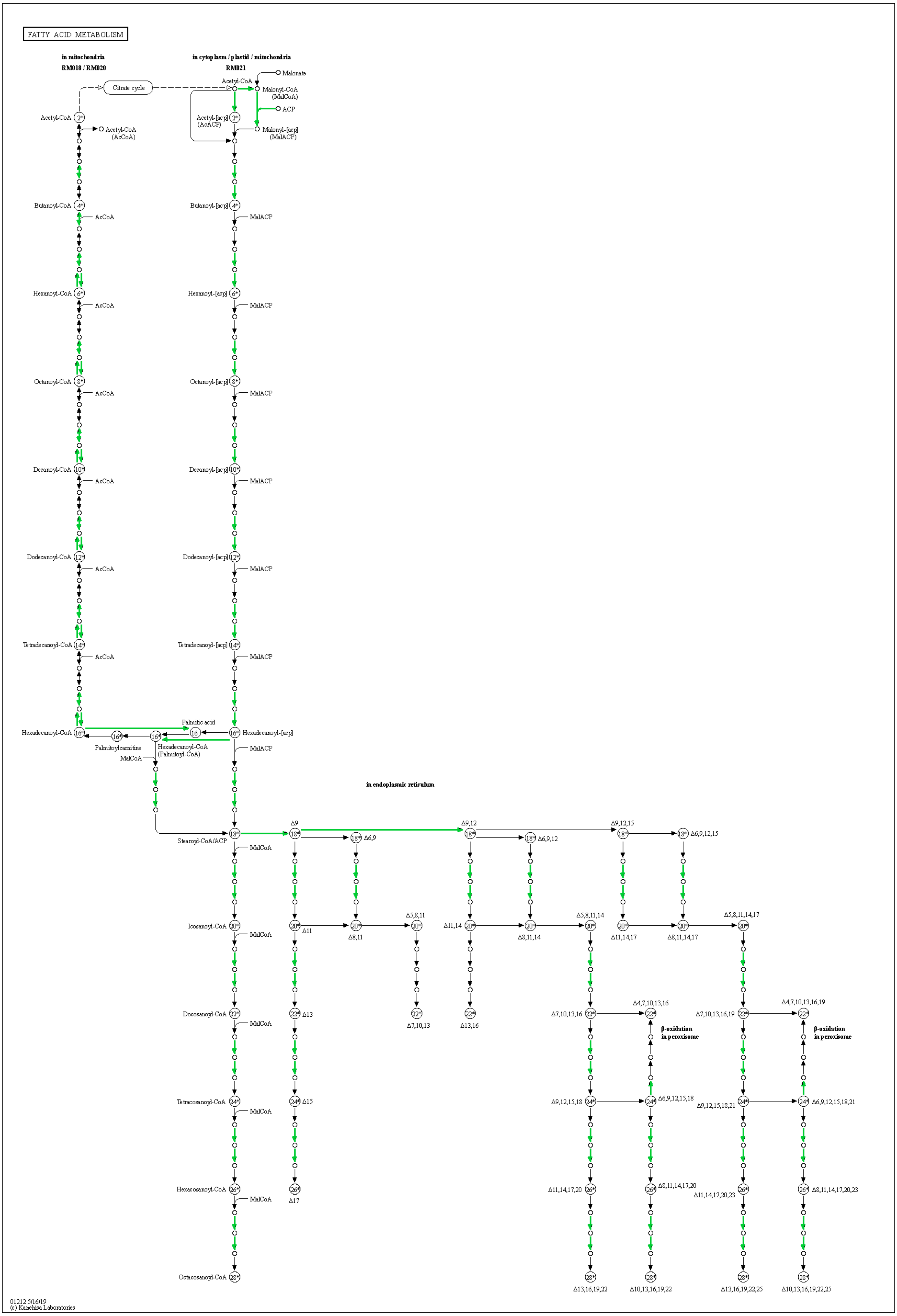
Figure 1. Overview of the genome assemblies of *Rhodotorula babjevae* strains: (a) CBS 7808; (b) DBVPG 8058. The concentric circles show from outside to inside the contig name and sizes, distribution of lipid and carotenoid metabolism related genes, and in non-overlapping 10 kb windows, the gene density, the deviation from the average GC content and the density of duplicated genes with 70% sequence coverage.

To estimate the ploidy in *R. babjevae* strains we used nQuire, which quantitatively distinguishes between different ploidy levels based on the distribution of base frequencies at variable sites [17]. In both strains, most of the alleles occur with frequencies of around 25% and 75%, indicating that both *R. babjevae* strains are tetraploid (Figure S1). The ploidy level of *R. babjevae* strains has not been studied before. While published whole genome sequences from strains of closely related *Rhodotorula* species such as *R. toruloides* NP11 and *R. mucilaginosa* JGTA-S1 have been reported to present haploid genomes [12,29], tetraploidy has been widely acknowledged in yeasts previously [31–33]. The identification of ploidy can be a valuable resource for developing genetic manipulation tools and establishing protocols.

A final number of 7,591 protein-coding genes and 7,607 associated transcripts were annotated for CBS 7808 using MetaEuk (Table 1). The average number of estimated exons per gene is 3.97 (Table 1). DBVPG 8058 has 7,481 protein-coding genes, 7,516 associated transcripts and 3.93 estimated exons per gene (Table 1). Hence, for both strains we corroborated the presence of split genes in high proportion within the genome, with final numbers of 6,390 and 6,305 for CBS 7808 and DBVPG 8058, respectively. This is in line with previous findings for *Rhodotorula* spp [3,12,29]. The distribution of exon counts in the genomes of *R. babjevae* strains CBS 7808 and DBVPG 8058 is shown in Table S4. 315 and 309 open reading frames (ORF) complementary to annotated genes were predicted in CBS 7808 and DBVPG 8058, respectively. The presence of antisense transcripts has been reported previously in the related species *R. toruloides* [3]. Figures S2-S4 show clustered annotations of coding sequences through Gene Ontology (GO) terms into the categories of biological processes, cellular components and molecular functions. Some examples of annotated genes that encode crucial enzymes for lipid and carotenoid metabolism are *CDC19*, *MAE1*, *MAE2*, *ACL1*, *ACL2*, *ACC1*, *FAS1*, *FAS2*, *OLE1*, *ACAD10*, *ACAD11*, *IBR3*, *D6C81_05617*, *POT1*, *LRO1*, *HMG1*, *HCS1*, *ERG8*, *crtYB*, *crtI* and *BTS1* (Tables S5-S6). Some differences among them are the absence of *ACL2*, *LRO1* and *ERG8* and the presence of *ACAD10* in DBVPG 8058. A total of 2,691 and 2,660 CDS from CBS 7808 and DBVPG 8058, respectively, could be assigned KO numbers with which we reconstructed metabolic pathways involved in biosynthesis of saturated and unsaturated fatty acid, glycerolipid metabolism, terpenoid backbone biosynthesis, carbon metabolism and fatty acid metabolism (Figure 2, Figure S5-S6).


FATTY ACID METABOLISM
in mitochondria
RM018 / RM020
in cytoplasm / plastid / mitochondria
RM021
Citrate cycle
Acetyl-CoA
Malonate
Malonyl-CoA
(MalCoA)
ACP
Acetyl-[acp]
(AcACP)
Malonyl-[acp]
(MalACP)
Acetyl-CoA
(AcCoA)
Butanoyl-CoA
Butanoyl-[acp]
AcCoA
MalACP
Hexanoyl-CoA
Hexanoyl-[acp]
Octanoyl-CoA
Octanoyl-[acp]
Decanoyl-CoA
Decanoyl-[acp]
Dodecanoyl-CoA
Dodecanoyl-[acp]
Tetradecanoyl-CoA
Tetradecanoyl-[acp]
Palmitic acid
Hexadecanoyl-CoA
Hexadecanoyl-[acp]
Palmitoylcarnitine
Hexadecanoyl-CoA
(Palmitoyl-CoA)
MalCoA
in endoplasmic reticulum
Stearoyl-CoA/ACP
Δ9
Δ6,9
Δ9,12
Δ6,9,12
Δ9,12,15
Δ6,9,12,15
Icosanoyl-CoA
Δ11
Δ8,11
Δ5,8,11
Δ11,14
Δ8,11,14
Δ5,8,11,14
Δ11,14,17
Δ8,11,14,17
Δ5,8,11,14,17
Docosanoyl-CoA
Δ13
Δ7,10,13
Δ13,16
Δ7,10,13,16
Δ4,7,10,13,16
Δ7,10,13,16,19
Δ4,7,10,13,16,19
β-oxidation
in peroxisome
Tetracosanoyl-CoA
Δ15
Δ9,12,15,18
Δ6,9,12,15,18
Δ9,12,15,18,21
Δ6,9,12,15,18,21
Hexacosanoyl-CoA
Δ17
Δ11,14,17,20
Δ8,11,14,17,20
Δ11,14,17,20,23
Δ8,11,14,17,20,23
Octacosanoyl-CoA
Δ13,16,19,22
Δ10,13,16,19,22
Δ13,16,19,22,25
Δ10,13,16,19,22,25
01212 5/16/19
(c) Kanehisa Laboratories

(a)


FATTY ACID METABOLISM
in mitochondria
RM018 / RM020
in cytoplasm / plastid / mitochondria
RM021
Citrate cycle
Acetyl-CoA
Malonate
Malonyl-CoA (MalCoA)
ACP
Acetyl-[acp] (AcACP)
Malonyl-[acp] (MalACP)
Acetyl-CoA (AcCoA)
Butanoyl-CoA
Butanoyl-[acp]
AcCoA
MalACP
Hexanoyl-CoA
Hexanoyl-[acp]
Octanoyl-CoA
Octanoyl-[acp]
Decanoyl-CoA
Decanoyl-[acp]
Dodecanoyl-CoA
Dodecanoyl-[acp]
Tetradecanoyl-CoA
Tetradecanoyl-[acp]
Hexadecanoyl-CoA
Palmitic acid
Hexadecanoyl-[acp]
Palmitoylcarnitine
Hexadecanoyl-CoA (Palmitoyl-CoA)
MalCoA
in endoplasmic reticulum
Stearoyl-CoA/ACP
Icosanoyl-CoA
Docosanoyl-CoA
Tetracosanoyl-CoA
Hexacosanoyl-CoA
Octacosanoyl-CoA
Δ9
Δ6,9
Δ9,12
Δ6,9,12
Δ9,12,15
Δ6,9,12,15
Δ11
Δ8,11
Δ5,8,11
Δ11,14
Δ8,11,14
Δ5,8,11,14
Δ11,14,17
Δ8,11,14,17
Δ5,8,11,14,17
Δ13
Δ7,10,13
Δ13,16
Δ7,10,13,16
Δ4,7,10,13,16
Δ7,10,13,16,19
Δ4,7,10,13,16,19
β-oxidation in peroxisome
Δ15
Δ9,12,15,18
Δ6,9,12,15,18
Δ9,12,15,18,21
Δ6,9,12,15,18,21
Δ17
Δ11,14,17,20
Δ8,11,14,17,20
Δ11,14,17,20,23
Δ8,11,14,17,20,23
Δ13,16,19,22
Δ10,13,16,19,22
Δ13,16,19,22,25
Δ10,13,16,19,22,25
01212 5/16/19
(c) Kanehisa Laboratories

(b)

Figure 2. Fatty acid metabolism pathways reconstructed by KEGG Mapper: (a) *Rhodotorula babjevae* CBS 7808; (b) *R. babjevae* DBVPG 8058. The CDSs with affiliated KEGG Orthology (KO) identifiers involved in each metabolic pathway are colored in green.

Benchmarking of universal single-copy orthologs (BUSCOs, using fungi_odb9) identified that 95.5% and 96.9% of the assessed genes in CBS 7808 and DBVPG 8058, respectively, were complete and single-copy (Figure S7). This supports the high quality of the draft genome assemblies reported herein. Furthermore, 0.7 % and 0.3% of the assessed genes were fragmented in CBS 7808 and DBVPG 8058, respectively, and the rest were missing (Figure S7).

3.2. Chromosome organization

The *R. babjevae* genome assemblies were aligned for comparison using NUCmer. From a total of 27 assembled contigs and scaffolds in CBS 7808, 24 had matches with 30 of the 34 assembled sequences in DBVPG 8058 (Figure 3). Even when there is a high number of undisturbed segments of conservation, a high proportion of chromosomal rearrangements can be spotted (Figure 3). LASTZ alignments of each contig from one *R. babjevae* strain with the whole genome of the other strain confirmed the results given by the synteny analysis (Table S7, Figure S8-S9). From these results, we deduce that *R. babjevae* has 21 putative chromosomes with sizes ranging from 0.4 to 2.4 Mbp (Table 2). The pairwise identity between chromosomes ranges mainly from 82% to 87%. The mitochondrial genomes have 97% pairwise identity (Table S7, Figure S8). Four of the chromosomes are affected by large translocation events between putative chromosomes 3 and 6, and between putative chromosomes 9 and 14 (Table 2). Smaller inversions are noticed in other chromosomes (Table S7, Figure S9). Each *R. babjevae* strain contains two contigs that are strain-specific (Table S7, Table S8). They comprise small-size linear contigs with higher read depths than the chromosomes, except for circular contig_26 in CBS 7808, which has a lower read depth than the chromosomes. These variations in read depth may be indicative of relaxed replication regulation. The linear DNA sequence from CBS 7808 contig_46 has two annotated genes, one of which encodes for Retrovirus-related Pol polyprotein from transposon 17.6. DNA plasmids have been previously found in filamentous fungi, including the close relative *R. toruloides*, with sizes ranging from 2.5 to 11 kb and typically encoding enzymes involved in plasmid replication [3,34,35]. This might indicate the presence of extrachromosomal endogenous DNA that is not shared between *R. babjevae* strains.

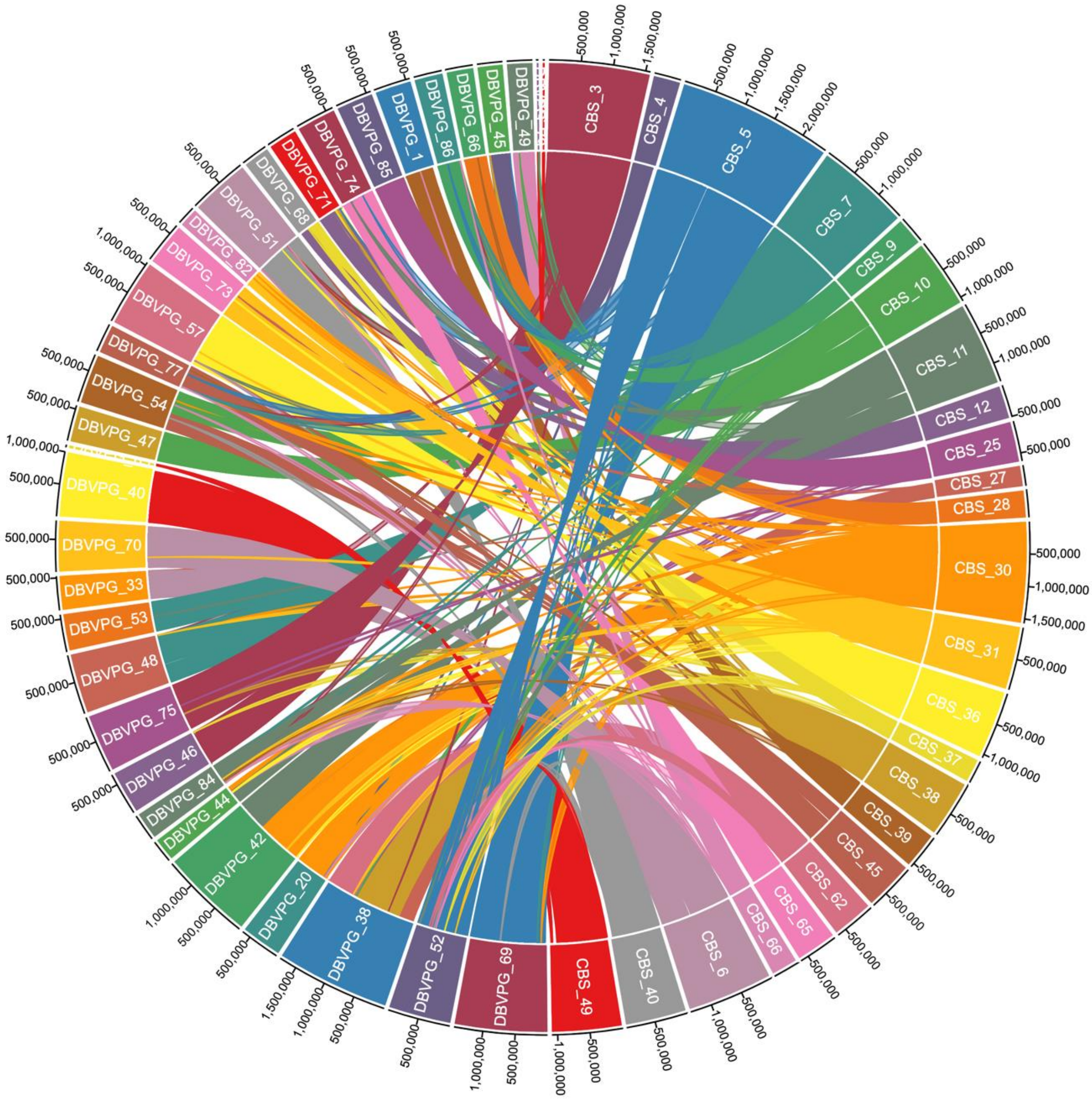


Figure 3. Genome alignment of *Rhodotorula babjevae* strains CBS 7808 and DBVPG 8058. Maximal unique matches between CBS 7808 and DBVPG 8058 were obtained using NUCmer 3.0 and visualized with Circa. The circular plot shows unique and repetitive alignments as ribbons using CBS 7808 contigs and scaffolds as the reference. Contig and scaffolds of CBS 7808 and DBVPG 8058 are labeled "CBS" and "DBVPG", respectively.

Table 2. Putative chromosomes in *Rhodotorula babjevae* deduced from whole genome LASTZ alignments (TableS7, Figure S9)

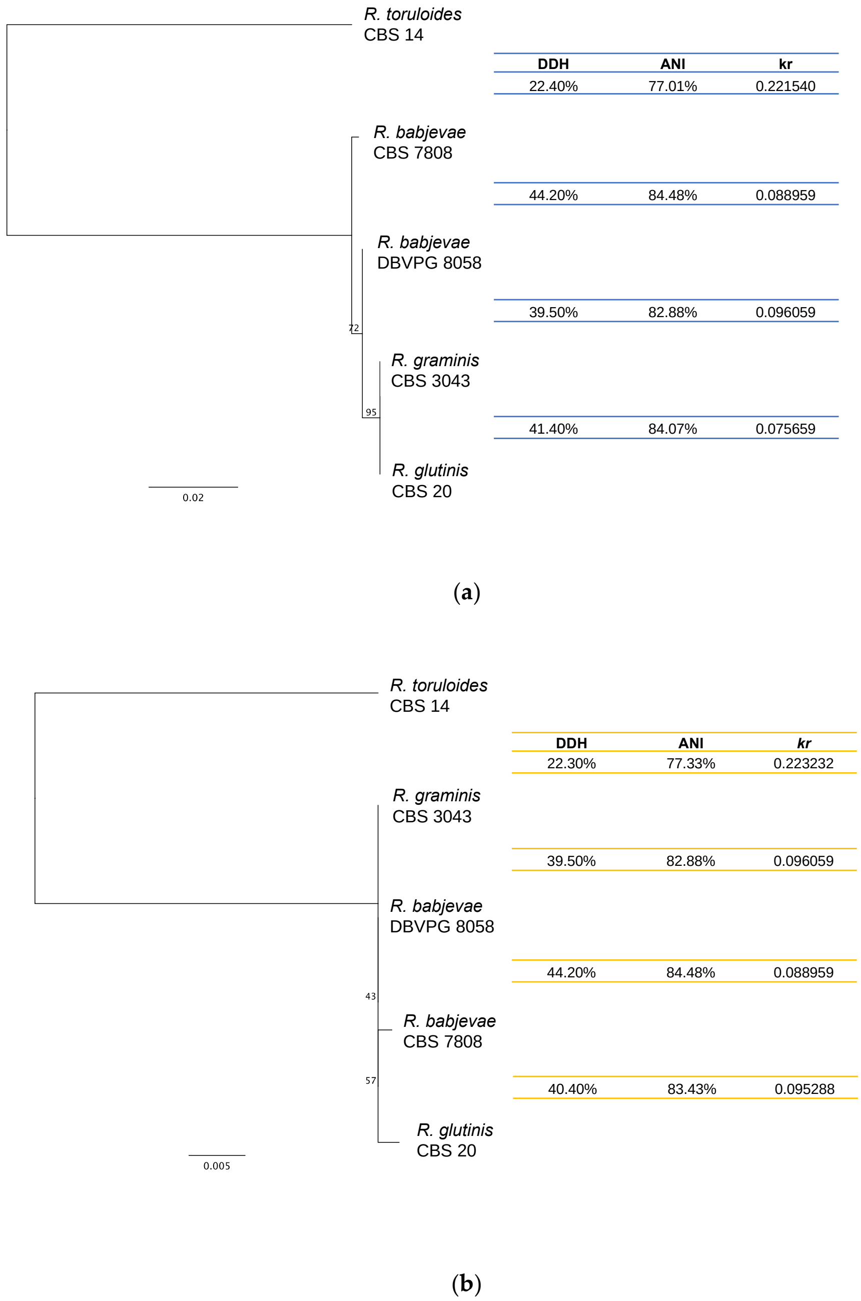
R. bajevae **CBS 7808**	***R. bajevae*** **DBVPG 8058**	**Genetic structure**	**GC content**	**Comments**	**Size (Mb)**
Contig_5 (2,415,752 bp)	Contig_69 (1,447,990 bp) Scaffold_52 (977,625 bp)	Putative chromosome 1	67-69%	Figure S9A	2.4
Contig_27 (320,063 bp) Contig_38 (881,966 bp) Contig_62 (644,441 bp)	Contig_38 (1,780,658 bp)	Putative chromosome 2	67-69%	Figure S9B	1.8
Contig_30 (1,569,459 bp)	Contig_20 (637,402 bp) Contig_42 (1,446,680 bp) Contig_44 (357,974 bp)	Putative chromosome 3	67-69%	Figure S9C Large translocation event between Chr. 3 and Chr.6	1.6

Contig_3 (1,574,520 bp)	Contig_46 (670,828 bp) Contig_75 (900,917 bp)	Putative chromosome 4	67-69%	Figure S9D	1.6
Contig_7 (1,460,653 bp)	Contig_48 (931,129 bp) Contig_53 (571,073 bp)	Putative chromosome 5	67-69%	Figure S9E	1.5
Contig_11 (1,300,441 bp)	Contig_42 (1,446,680 bp) Contig_44 (357,974 bp) Contig_84 (425,340 bp)	Putative chromosome 6	67-69%	Figure S9F Large translocation event between Chr. 3 and Chr.6	1.3
Scaffold_6 (1,337,997 bp)	Contig_33 (529,001 bp) Contig_70 (789,767 bp)	Putative chromosome 7	67-69%	Figure S9G	1.3
Scaffold_49 (1,089,446 bp)	Contig_40 (1,004,683 bp) Contig_65 (41,334 bp)	Putative chromosome 8	67-69%	Figure S9H	1.1
Contig_10 (1,067,634 bp)	Contig_47 (557,103 bp) Contig_54 (766,724 bp)	Putative chromosome 9	67-69%	Large translocation event between Chr. 9 and Chr.14 Figure S9I	1.1
Contig_36 (1,056,323 bp)	Contig_57 (1,049,892 bp)	Putative chromosome 10	67-69%	Figure S9J	1.1
Contig_31 (979,228 bp)	Contig_73 (659,761 bp) Contig_82 (299,180 bp)	Putative chromosome 11	67-69%	Figure S9K	1.0
Scaffold_40 (948,604 bp)	Contig_51 (924,743 bp)	Putative chromosome 12	67-69%	Figure S9L	0.9
Contig 37 (362,520 bp) Contig_12 (511,897 bp)	Contig_68 (408,627 bp) Contig_71 (449,691 bp)	Putative chromosome 13	67-69%	Figure S9M	0.9
Contig_45 (762,860 bp)	Contig_54 (766,724 bp) Contig_77 (446,828 bp)	Putative chromosome 14	67-69%	Figure S9N Large translocation event between Chr. 9 and Chr.14	0.8
Contig_65 (630,535 bp)	Contig_74 (614,034 bp	Putative chromosome 15	67-69%	Figure S9O	0.6
Contig_25 (627,118 bp)	Contig_85 (573,802 bp)	Putative chromosome 16	67-69%	Figure S9P	0.6
Contig_39 (564,129 bp)	Contig_1 (565,532 bp)	Putative chromosome 17	67-69%	Figure S9Q	0.6
Contig_9 (429,397 bp)	Contig_86 (443,617 bp)	Putative chromosome 18	67-69%	Figure S9R	0.4
Contig_28 (422,133 bp)	Contig_66 (419,035 bp)	Putative chromosome 19	67-69%	Figure S9S	0.4
Contig_4 (418,972 bp)	Contig_45 (394,205 bp)	Putative chromosome 20	67-69%	Figure S9T	0.4
Contig_66 (406,102 bp)	Contig_49 (396,114 bp)	Putative chromosome 21	67-69%	Figure S9U	0.4

Chr., chromosome

3.2. Genome divergence analysis

The genomes from the described *R. babjevae* strains were further compared to each other and to genomes from other closely related *Rhodotorula* species in terms of DDH, ANI and *kr* for tracing genome divergence (Figure 4, Table S9). *R. babjevae* strains share 44.20% DDH estimates, 84.48% ANI and *kr* values of 0.09. In general, the genetic divergence between *R. babjevae* strains was comparable to the divergence with *R. graminis* and *R. glutinis* and higher than expected for strains of the same species. For instance, divergence between strains of *R. toruloides* was much lower than for the two investigated *R. babjevae* strains. (Table S9).



(**a**)

(**b**)

Figure 4. Phylogenetic relationship of *Rhodotorula babjevae* strains and their placement within the *Rhodotorula* genus. The phylogenetic tree was built based on: (a) ITS; and (b) D1/D2 LSU of rRNA gene sequences. It was inferred using PhyML with 100 bootstraps on Geneious prime version 2021.0.1. *R. toruloides* was selected as outgroup. Similarities between whole genome sequences of the corresponding strains are presented in terms of the alignment-free distance measure *kr*, Average Nucleotide Identity (ANI) and DNA–DNA homology (DDH). *R. graminis* WP1 and *R. glutinis* ZHK genome sequences were used for the calculations instead of *R. graminis* CBS 3043 and *R. glutinis* CBS 20, respectively.

Moreover, the protein-coding sequences from the described *R. babjevae* strains and their closest relatives *R. graminis* and *R. glutinis* were analyzed using OrthoVenn2 web platform to identify and compare orthologous gene clusters. *R. babjevae* species share 6,598 from a total of 7,223 orthologous clusters produced by OrthoVenn2, including both single-copy gene clusters and overlapping gene clusters such as paralogs (Figure 5). 5,933 of the shared clusters are common within the three assessed *Rhodotorula* species, representing putative shared orthologous proteins that have evolved from common ancestral genes. In addition, CBS 7808 has 389 single genes and one cluster that didn´t have orthologs in the other genomes, while the strain DBVPG 8058 has 355 single genes. These unique genes could account for the specific functional capabilities of the described *R. babjevae* strains as a result of gene loss or gain events. From the 79 orthologous clusters shared only between *R. babjevae* strains, some of the assigned GO terms are positive regulation of the unsaturated fatty acid biosynthetic process by positive regulation of transcription from RNA polymerase II promoter (GO:0036083), protein O-linked glycosylation (GO:0006493), glucan catabolic process (GO:0009251), cellular calcium ion homeostasis (GO:0006874), sulfate assimilation (GO:0000103) and carbohydrate transport (GO:0008643). Likewise, in the previous results, the two *R. babjevae* strains have a high genome pairwise similarity and number of shared orthologous clusters, though not as high as for *R. graminis* and *R. glutinis* (Figure 5). In general, *R. babjevae*, *R. glutinis* and *R. graminis* are very closely related species with a short evolutionary distance between them compared to other species in the genus (i.e., *R. toruloides*). The described strains CBS 7808 and DBVPG 8058 have high inter-strain variability and a greater evolutionary distance to *R. graminis* than *R. glutinis*.

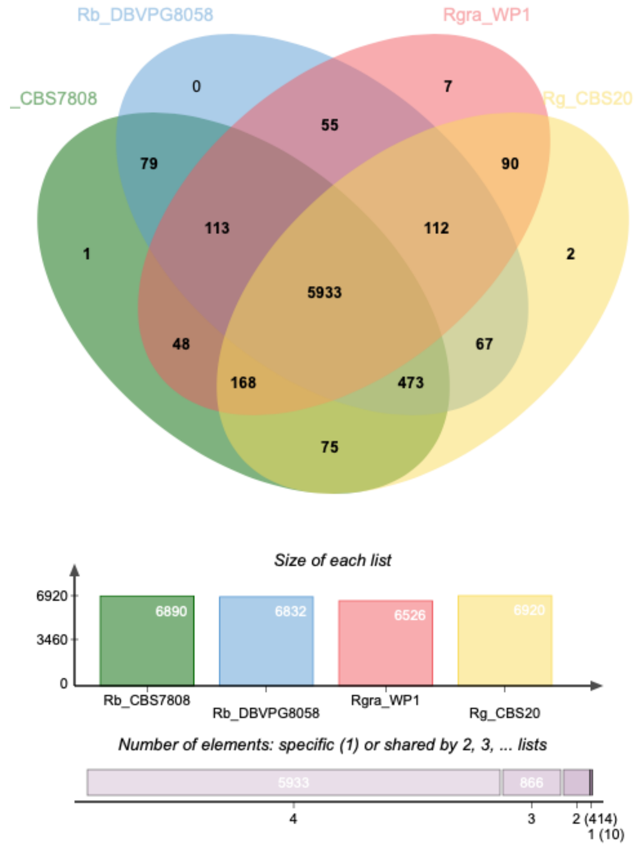


Figure 5. Distribution of shared orthologous clusters between *Rhodotorula babjevae* strains CBS 7808 and DBVPG 8058, *R. graminis* WP1 and *R. glutinis* CBS 20. The Venn diagram was generated using OrthoVenn2.

In total 59 and 30 paralogous gene clusters were identified in CBS 7808 and DBVPG 8058, respectively, using OrthoVenn2 (Table S10-S11). When applying a cut-off value of

70% sequence coverage to them, we identified 29 and 19 duplicated genes, respectively, that potentially haven´t diverged in function. On the other hand, an all-against-all protein sequence similarity search was performed in each of the two strains using BLASTp with e-value 1e-15, 70% coverage and 70% sequence identity. It resulted in a total of 34 and 21 duplicated sequences in CBS 7808 and DBVPG 8058, respectively, and for a total of 41 and 29 duplicated sequences with 70% sequence coverage, respectively, that were identified by any of the tools (Figure 1, Table 3, Table S12-13). The higher accumulation of duplicated genes in CBS 7808 could be related to a higher number of gene duplication events due to faster evolution of the strain. The majority of these duplications lies adjacent to each other or in close proximity. Tandem duplications have been suggested as a mechanism of adaptative evolution to changing environments [36]. They can have arisen through homologous recombination between sequences on sister chromatids or homologous chromosomes [36]. A substantial redundancy of duplicate gene pairs has been reported to maintain even after 100 million years of evolution in *Saccharomyces cerevisiae* [37]. Some of the predicted functions from genes that are duplicated only in CBS 7808 are Uncharacterized protein C17G8.02 (NAD biosynthesis), Mannose-6-phosphate isomerase and Phosphoenolpyruvate carboxykinase (ATP) (carbon metabolism), Acetyl-CoA carboxylase (fatty acid metabolism), Alpha-ketoglutarate-dependent sulfonate dioxygenase, Sulfite reductase [NADPH] hemoprotein beta-component and Sulfite reductase [NADPH] subunit beta (Sulfur metabolism), and Probable quinate permease (import of quinic acid as a carbon source). On the other hand, some of the duplicated genes involved in metabolic processes identified only in DBVPG 8058 are mitochondrial Aspartate aminotransferase (intracellular NAD(H) redox balance) and Leucine-rich repeat extensin-like protein 3 (AtLRX3, cell morphogenesis). In both strains, the most common duplicated gene is *SRRM2*, which codes for Ser/Arg repetitive matrix protein 2 and is involved in mRNA splicing. Cwc21p is encoded by *CWC21*, an ortholog of human *SRRM2* in *S. cerevisiae*. It has been proposed to be in the catalytic center of the spliceosome and possibly perform its role in response to changing conditions of the cell environment [38]. The predicted function Ser/Arg repetitive matrix protein 2 was annotated in 1055 genes in CBS 7808 and 1068 in DBVPG 8058. Alternative splicing is an essential driver of proteomic diversity and can potentially provide a high level of evolutionary plasticity.

Table 3. Duplicated genes in *Rhodotorula babjevae* identified by BLASTp and OrthoVenn2 with a minimum coverage of 70%.

Genetic structure	CBS 7808	Functional prediction	DBVPG 8058	Functional prediction
Putative chromosome 1	4319/4337	Protein_bcp1	4979/4980	Carbamoyl-phosphate_synthase_arginine-specific_large_chain
	4341/4342	Ser/Arg_repetitive_matrix_protein_2	5060/5062	Ser/Arg_repetitive_matrix_protein_2 & Pantothenate_transporter_liz1
	4606/4608	Phenylalanine--tRNA_ligase_alpha_subunit & Probable_feruloyl_esterase_B-2	7488/7489	EF-1-alpha
	4665	Uncharacterized_protein_C17G8.02	7497/7498	Glycoprotein_gp2
	4799/4806	Immediate-early_protein_2 & Putative_uncharacterized_protein_ENSP00000383309		
	4904	A-agglutinin_anchorage_subunit		
Putative chromosome 2	2814/2815	Ser/Arg_repetitive_matrix_protein_1	699/700	Quinone-oxidoreductase_homolog,_chloroplastic & MUC-5AC

	2854/2877	Putative_uncharacterized_protein_ENSP00000383309 & Ser/Arg_repetitive_matrix_protein_3	881/918	Histone_H2A
	3041/3043	Heat_shock_70_kDa_protein & Heat_shock_protein_SSC1,_mitochondrial		
	5361/5364	Protein_arginine_N-methyltransferase_1		
Putative chromosome 3	1409	Ser/Arg_repetitive_matrix_protein_1	1575/1776*	Cytochrome_P450_monooxygenase_ALT8
	1471/1488	Ser/Arg_repetitive_matrix_protein_2 & Protein_YIP5	1785/1823*	MUC-5AC & Pre-mRNA-splicing_factor_CWC22
	1500/1501	Putative_protein_TPRXL	2011/2024*	Ser/Arg_repetitive_matrix_protein_2 & AtPERK9
	1515/1516	Ser/Arg_repetitive_matrix_protein_2		
	1749/1763	Protein_SON & Pre-mRNA-splicing_factor_CWC21		
	1851/1892	RNA-binding_protein_with_serine-rich_domain_1 & ERF3		
	1960	Pneumococcal_serine-rich_repeat_protein		
Putative chromosome 4			6248/6249	IE2
Putative chromosome 5	5765/5781	Ser/Arg_repetitive_matrix_protein_2 & Ser/Arg_repetitive_matrix_protein_1	2853/2854	Heat_shock_protein_60,_mitochondrial
	5915	Uncharacterized_protein_C17G8.02	3517	Uncharacterized_serine-rich_protein_C215.13
	5929/5930	Heat_shock_protein_60,_mitochondrial	3695/3696	40S_ribosomal_protein_S1
Putative chromosome 6	701	Alpha-ketoglutarate-dependent_sulfonate_dioxygenase	1575/1776*	Cytochrome_P450_monooxygenase_ALT8
			1785/1823*	MUC-5AC & Pre-mRNA-splicing_factor_CWC22
			2011/2024*	Ser/Arg_repetitive_matrix_protein_2 & AtPERK9
			6681/6687	Endochitinase_2 & Glycoprotein_gp2
Putative chromosome 7	7144/7145	Tryptophan_synthase	5308/5318	Histone_H3.2
	7266/7271	Mannose-6-phosphate_isomerase		
	7286/7301	Histone_H3.2		
	7312/7313	Acetyl-CoA_carboxylase		
	7498/7497	Sulfite_reductase_[NADPH]_hemoprotein_beta-component & Sulfite_reductase_[NADPH]_subunit_beta		
Putative chromosome 8	6916	Chitin_deacetylase_1	1190/1191	AtLRX3
	6968/6969	60S_ribosomal_protein_L3		
Putative chromosome 9	24/38	Ser/Arg_repetitive_matrix_protein_2 & Vitellogenin-1	2581/2590	Ser/Arg_repetitive_matrix_protein_2

	84	Alpha-ketoglutarate-dependent_dioxygenase_cnsM	2601/2627	MUC-5AC
	122/123	MUC-5AC & Ser/Arg-rich_splicing_factor_SR45	2722/2726	Ser/Arg_repetitive_matrix_protein_1
	326/327	Ser/Arg_repetitive_matrix_protein_2 & Ser/Arg_repetitive_matrix_protein_1	3739/3743*	Trimethylguanosine_synthase
Putative chromosome 10	2615/2616	Ser/Arg_repetitive_matrix_protein_1		
Putative chromosome 11	2140/2163	MUC-5AC & AtOPT4	5763/5786	AtOPT4
	2182	Ser/Arg_repetitive_matrix_protein_1	6580	Histone_H3
Putative chromosome 12	6541/6542	Fumarate_hydratase,_mitochondrial	3245/3246	Splicing_factor_YJU2 & Protamine
			3354/3355	Actin
			3393/3394	Fumarate_hydratase,_mitochondrial
Putative chromosome 13	801/812	Ser/Arg_repetitive_matrix_protein_2	5441/5451	Ser/Arg_repetitive_matrix_protein_2
	830/831	Phosphoenolpyruvate_carboxykinase_(ATP)		
	2717/2727	Ser/Arg_repetitive_matrix_protein_2 & Putative_protein_TPRXL		
Putative chromosome 14	4016/4020	Trimethylguanosine_synthase	6501/6510	Putative_GPI-anchored_protein_pfl2
	4044/4049	Branched_chain_2-oxo-acid_dehydrogenase_complex_component_E2	3739/3743*	Trimethylguanosine_synthase
Putative chromosome 17	3191/3192	Ser/Arg_repetitive_matrix_protein_1	07/08	Aspartate_aminotransferase,_mitochondrial
	3207/3228	MUC-5AC & Uncharacterized_serine-rich_protein		
	3262/3263	Probable_aldo-keto_reductase_2 & Aldo-keto_reductase_yakc		
Putative chromosome 18			7028	Bromodomain_and_WD_repeat-containing_protein_3
Putative chromosome 19	1276/1280	Probable_quinate_permease		
Putative chromosome 20			2269/2270	TCP-1-zeta
Putative chromosome 21	5744	Pneumococcal_serine-rich_repeat_protein		

* Paralogous sequences from DBVPG 8058 that are located in chromosomes containing large translocations compared to CBS 7808. The paralogous sequences with orthologs in the same putative chromosome from the other *R. babjevae* strain are indicated in black.

The here investigated type strain of *R. babjevae*, CBS 7808, was first isolated from herbaceous plants in Moscow, Russia [39]. *R. babjevae* DBVPG 8058 was isolated from wild apples in Uppland locality, Sweden. Their phylogenetic placement was done through aligning 5.8S-ITS rDNA and D1/D2 26S rDNA regions as shown in Figure 4. The estimated genome divergence values through DDH, ANI and *kr* proved to be more sensitive for delineating *Rhodotorula* species. Phylogenetic placement based on the standard rDNA regions may not be enough to understand yeast diversity and species delineation as shown before [40,41]. These *R. babjevae* strains showed different behavior during enzymatic cell wall degradation for DNA purification within this study and when they were grown on xylose medium in another study [42]. Highly dynamic genome structures have been previously found within closely related yeast species [13,43–46]. A dynamic genome structure of *R. babjevae* could enhance the adaptability of the species to each environment and the subsequent physiological differences [47–50]. However, their genetic divergence suggests that they could belong to different species. A genome comparison study using whole genome sequences from different strains from closely related *Rhodotorula* species will allow to gain a deeper knowledge about their genome diversity, evolution and to identify novel yeast species.

A taxonomic classification using Sourmash [51] and a GenBank reference (https://osf.io/4f8n3) assigns to both genome assemblies: Eukaryota superkingdom, Basidiomycota phylum, Microbotryomycetes class, Sporidiobolales order, Sporidiobolaceae family, *Rhodotorula* genus, *Rhodotorula graminis* species. The retrieved taxonomic classification might indicate that *R. graminis* was the closest relative of *R. babjevae* with available genomic data. Previous studies have shown a close evolutionary relationship between *R. babjevae* and *R. graminis*, which was also demonstrated here [30,52].

4. Conclusions

We present the *de novo* genome assembly and annotation of the tetraploid strains from *R. babjevae* DBVPG 8058 and CBS 7808^T. We predict the number of putative chromosomes in the species and identify large-scale translocation events. Moreover, we demonstrate a high genome divergence between the *R. babjevae* strains, comparable to the divergence to other closely related *Rhodotorula* species.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Allele frequency values of single nucleotide polymorphisms (SNP) in *Rhodotorula babjevae*.; Figure S2: Gene Ontology (GO) term summary related to the GO topic: molecular functions; Figure S3: Gene Ontology (GO) term summaries belonging to the GO topic: biological processes; Figure S4: Gene Ontology (GO) term summaries belonging to the GO topic: cellular components; Figure S5: Examples of lipid metabolism pathways in *Rhodotorula babjevae* CBS 7808 reconstructed by KEGG Mapper; Figure S6: Examples of lipid metabolism pathways in *Rhodotorula babjevae* DBVPG 8058 reconstructed by KEGG Mapper; Figure S7: Quantitative assessment of the hybrid genome assemblies and annotation completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO); Figure S8: LASTZ alignment of the mitochondrial genome sequences from *Rhodotorula babjevae* CBS 7808 and DBVPG 8058; Figure S9: LASTZ alignment of contigs with assigned homology from *Rhodotorula babjevae* CBS 7808 and DBVPG 8058 representing putative chromosomes; Table S1: Program versions used for the genome assembly and annotation pipeline; Table S2: Characteristics from the contigs and scaffolds of *Rhodotorula babjevae* CBS 7808 genome assembly; Table S3: Characteristics from the contigs and scaffolds of *Rhodotorula babjevae* DBVPG 8058 genome assembly; Table S4: Distribution of exon counts in the transcriptome of two strains of *Rhodotorula babjevae*; Table S5: Examples of lipid and carotenoid metabolism related genes in *Rhodotorula babjevae* CBS 7808 genome assembly; Table S6: Examples of lipid and carotenoid metabolism related genes in *Rhodotorula babjevae* DBVPG 8058 genome assembly; Table S7: Contigs with assigned homology between *Rhodotorula babjevae* strains; Table S8: Summary of features from strain-unique contigs in *Rhodotorula babjevae*; Table S9: Genetic divergence between *Rhodotorula babjevae* strains and closely related *Rhodotorula* species; Table S10: Alignment statistics of the duplicated genes from the genome of *Rhodotorula babjevae* CBS 7808 identified by OrthoVenn2; Table S11: Alignment statistics of the duplicated genes from the genome of *Rhodotorula babjevae* DBVPG 8058 identified by OrthoVenn2; Table S12: Alignment statistics of the duplicated genes from the genome of *Rhodotorula babjevae* CBS 7808 identified by BLASTp; Table S13: Alignment statistics of the duplicated genes from the genome of *Rhodotorula babjevae* DBVPG 8058 identified by BLASTp.

Author Contributions: Conceptualization, V.P.; methodology, B.M.; validation, C.B., M.H. and A.V.; formal analysis, G.C.M.-H. and B.M.; investigation, G.C.M.-H. and BM; resources, V.P., C.B., M.H. and A.V.; data curation, C.B., M.H. and A.V.; writing—original draft preparation, G.C.M.-H.; writing—review and editing, B.M., V.P., C.B., M.H. and A.V.; visualization, G.C.M.-H.; supervision,

B.M.; project administration, V.P.; funding acquisition, V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas), grant number 2018-01877.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
Data Availability Statement: This project has been deposited at ENA under the accession PRJEB48745.

Acknowledgments: Illumina sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, which is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. A.V., C.B. and M.H. are shareholders of nanozoo GmbH.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Poontawee, R.; Yongmanitchai, W.; Limtong, S. Efficient oleaginous yeasts for lipid production from lignocellulosic sugars and effects of lignocellulose degradation compounds on growth and lipid production. *Process Biochem.* **2017**, *53*, 44–60, doi:10.1016/j.procbio.2016.11.013.
2. Brandenburg, J.; Poppele, I.; Blomqvist, J.; Puke, M.; Pickova, J.; Sandgren, M.; Rapoport, A.; Vedernikovs, N.; Passoth, V. Bioethanol and lipid production from the enzymatic hydrolysate of wheat straw after furfural extraction. *Appl. Microbiol. Biotechnol.* **2018**, *102*, 6269–6277, doi:10.1007/s00253-018-9081-7.
3. Martín-Hernández, G.C.; Müller, B.; Chmielarz, M.; Brandt, C.; Hölzer, M.; Viehweger, A.; Passoth, V. Chromosome-level genome assembly and transcriptome- based annotation of the oleaginous yeast *Rhodotorula toruloides* CBS 14. *Genomics* **2021**, *113*, 4022–4027, doi: 10.1016/j.ygeno.2021.10.006.
4. Chmielarz, M.; Blomqvist, J.; Sampels, S.; Sandgren, M.; Passoth, V. Microbial lipid production from crude glycerol and hemicellulosic hydrolysate with oleaginous yeasts. *Biotechnol. Biofuels* **2021**, *14*, 1–12, doi:10.1186/s13068-021-01916-y.
5. Ayadi, I.; Belghith, H.; Gargouri, A.; Guerfali, M. Screening of new oleaginous yeasts for single cell oil production, hydrolytic potential exploitation and agro-industrial by-products valorization. *Process Saf. Environ. Prot.* **2018**, *119*, 104–114, doi:10.1016/j.psep.2018.07.012.
6. Blomqvist, J.; Pickova, J.; Tilami, S.K.; Sampels, S.; Mikkelsen, N.; Brandenburg, J.; Sandgren, M.; Passoth, V. Oleaginous Yeast as a Component in Fish Feed. *Sci. Rep.*, **2018**, *8*, 1–8, doi:10.1038/s41598-018-34232-x
7. Guerfali, M.; Ayadi, I.; Mohamed, N.; Ayadi, W.; Belghith, H.; Bronze, M.R.; Ribeiro, M.H.L.; Gargouri, A. Triacylglycerols accumulation and glycolipids secretion by the oleaginous yeast *Rhodotorula babjevae* Y-SL7: Structural identification and biotechnological applications. *Bioresour. Technol.* **2019**, *273*, 326–334, doi:10.1016/j.biortech.2018.11.036.
8. Sen, S.; Borah, S.N.; Bora, A.; Deka, S. Production, characterization, and antifungal activity of a biosurfactant produced by *Rhodotorula babjevae* YS3. *Microb. Cell Fact.* **2017**, *16*, 1–14, doi:10.1186/s12934-017-0711-z.
9. Seveiri, R. Characterization and prospective applications of the exopolysaccharides poduced by *Rhodosporidium babjevae*. *Adv Pharm Bull* **2020**, *10*, 254–263, doi:10.34172/apb.2020.030.
10. Sen, S.; Borah, S.N.; Kandimalla, R.; Bora, A.; Deka, S. Sophorolipid biosurfactant can control cutaneous dermatophytosis caused by *Trichophyton mentagrophytes*. *Front. Microbiol.* **2020**, *11*, 1–15, doi:10.3389/fmicb.2020.00329.
11. Olsen, R.A.; Bunikis, I.; Tiukova, I.; Holmberg, K.; Lötstedt, B.; Pettersson, O.V.; Passoth, V.; Käller, M.; Vezzi, F. De novo assembly of *Dekkera bruxellensis*: A multi technology approach using short and long-read sequencing and optical mapping. *Gigascience* **2015**, *4*, doi:10.1186/s13742-015-0094-1.
12. Sen, D.; Paul, K.; Saha, C.; Mukherjee, G.; Nag, M.; Ghosh, S.; Das, A.; Seal, A.; Tripathy, S. A unique life-strategy of an endophytic yeast *Rhodotorula mucilaginosa* JGTA-S1—a comparative genomics viewpoint. *DNA Res.* **2019**, *26*, 131–146, doi:10.1093/dnares/dsy044.
13. Tiukova, I.A.; Pettersson, M.E.; Hoeppner, M.P.; Olsen, R.A.; Käller, M.; Nielsen, J.; Dainat, J.; Lantz, H.; Söderberg, J.; Passoth, V. Chromosomal genome assembly of the ethanol production strain CBS 11270 indicates a highly dynamic genome structure in the yeast species *Brettanomyces bruxellensis*. *PLoS One* **2019**, *14*, 1–20, doi:10.1371/journal.pone.0215077.
14. Chmielarz, M.; Sampels, S.; Blomqvist, J.; Brandenburg, J.; Wende, F.; Sandgren, M.; Passoth, V. FT-NIR: A tool for rapid intracellular lipid quantification in oleaginous yeasts. *Biotechnol. Biofuels* **2019**, *12*, 1–9, doi:10.1186/s13068-019-1513-9.
15. Pi, H.W.; Anandharaj, M.; Kao, Y.Y.; Lin, Y.J.; Chang, J.J.; Li, W.H. Engineering the oleaginous red yeast *Rhodotorula glutinis* for simultaneous β-carotene and cellulase production. *Sci. Rep.* **2018**, *8*, 2–11, doi:10.1038/s41598-018-29194-z.

16. Brandt, C.; Bongcam-Rudloff, E.; Müller, B. Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants. *Appl. Sci.* **2020**, *10*, 1–14, doi:10.3390/app10217518.
17. Weiß, C.L.; Pais, M.; Cano, L.M.; Kamoun, S.; Burbano, H.A. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* **2018**, *19*, 122, doi:10.1186/s12859-018-2128-z.
18. Aramaki, T.; Blanc-Mathieu, R.; Endo, H.; Ohkubo, K.; Kanehisa, M.; Goto, S.; Ogata, H. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **2020**, *36*, 2251–2252, doi:10.1093/bioinformatics/btz859.
19. Gremme, G.; Steinbiss, S.; Kurtz, S. Genome tools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2013**, *10*, 645–656, doi:10.1109/TCBB.2013.68.
20. Rodriguez-R, L.; Konstantinidis, K. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr.* **2016**, 4:e1900v1, doi:10.7287/peerj.preprints.1900.
21. Meier-Kolthoff, J.P.; Auch, A.F.; Klenk, H.P.; Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **2013**, *14*, doi:10.1186/1471-2105-14-60.
22. Harris, B.; Riemer, C.; Miller, W. Lastz. **2007**.
23. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.
24. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321, doi:10.1093/sysbio/syq010.
25. Xu, L.; Dong, Z.; Fang, L.; Luo, Y.; Wei, Z.; Guo, H.; Zhang, G.; Gu, Y.Q.; Coleman-Derr, D.; Xia, Q.; et al. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **2019**, *47*, W52–W58, doi:10.1093/nar/gkz333.
26. Kuznetsov, A.; Bollin, C.J. NCBI Genome Workbench: desktop software for comparative genomics, visualization, and genbank data submission. *Methods Mol. Biol.* **2021**, *2231*, 261–295, doi:10.1007/978-1-0716-1036-7_16.
27. Goordial, J.; Raymond-Bouchard, I.; Riley, R.; Ronholm, J.; Shapiro, N.; Woyke, T.; LaButti, K.M.; Tice, H.; Amirebrahimi, M.; Grigoriev, I. V.; et al. Improved high-quality draft genome sequence of the eurypsychrophile *Rhodotorula* sp. JG1b, isolated from permafrost in the hyperarid upper-elevation McMurdo Dry Valleys, Antarctica. *Genome Announc.* **2016**, *4*, 15–17, doi:10.1128/genomeA.00069-16.
28. Firrincieli, A.; Otillar, R.; Salamov, A.; Schmutz, J.; Khan, Z.; Redman, R.S.; Fleck, N.D.; Lindquist, E.; Grigoriev, I. V.; Doty, S.L. Genome sequence of the plant growth promoting endophytic yeast *Rhodotorula graminis* WP1. *Front. Microbiol.* **2015**, *6*, doi:10.3389/fmicb.2015.00978.
29. Zhu, Z.; Zhang, S.; Liu, H.; Shen, H.; Lin, X.; Yang, F.; Zhou, Y.J.; Jin, G.; Ye, M.; Zou, H.; et al. A multi-omic map of the lipid-producing yeast *Rhodosporidium toruloide*s. *Nat. Commun.* **2012**, *3*, 1111–1112, doi:10.1038/ncomms2112.
30. Li, C.J.; Zhao, D.; Cheng, P.; Zheng, L.; Yu, G.H. Genomics and lipidomics analysis of the biotechnologically important oleaginous red yeast R*hodotorula glutinis* ZHK Provides New Insights into Its Lipid and Carotenoid Metabolism. BMC Genomics, **2020**, 21, 1–16, doi:10.1186/s12864-020-07244-z.
31. Krahulec, J.; Lišková, V.; Boňková, H.; Lichvariková, A.; Šafranek, M.; Turňa, J. The ploidy determination of the biotechnologically important yeast *Candida utilis*. *J. Appl. Genet.* **2020**, *61*, 275–286, doi:10.1007/s13353-020-00544-w.
32. Fijarczyk, A.; Hénault, M.; Marsit, S.; Charron, G.; Fischborn, T.; Nicole-Labrie, L.; Landry, C.R. The genome sequence of the Jean-Talon strain, an archeological beer yeast from Québec, reveals traces of adaptation to specific brewing conditions. *G3 Genes|Genomes|Genetics* **2020**, *10*, 3087–3097, doi:10.1534/g3.120.401149.
33. Gallone, B.; Steensels, J.; Prahl, T.; Soriaga, L.; Saels, V.; Herrera-Malaver, B.; Merlevede, A.; Roncoroni, M.; Voordeckers, K.; Miraglia, L.; et al. Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **2016**, *166*, 1397-1410.e16, doi:10.1016/j.cell.2016.08.020.
34. Cahan, P. and Kennell, J.C. Identification and distribution of sequences having similarity to mitochondrial plasmids in mitochondrial genomes of filamentous fungi. *Mol. Genet. Genomics* **2005**, *273*, 462–473, doi:10.1007/s00438-005-1133-x.
35. Wang, Y.; Zeng, F.; Hon, C.C.; Zhang, Y.; Leung, F.C.C. The mitochondrial genome of the Basidiomycete fungus *Pleurotus ostreatus* (oyster mushroom). *FEMS Microbiol. Lett.* **2008**, *280*, 34–41, doi:10.1111/j.1574-6968.2007.01048.x.
36. Lallemand, T.; Leduc, M.; Landès, C.; Rizzon, C.; Lerat, E. An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes* **2020**, *11*, 1046, doi:10.3390/genes11091046
37. Dean, E.J.; Davis, J.C.; Davis, R.W.; Petrov, D.A. Pervasive and persistent redundancy among duplicated genes in yeast. *PLOS Genet.* **2008**, *4*, e1000113, doi: 10.1371/journal.pgen.1000113
38. Grainger, R.J.; Barrass, J.D.; Jacquier, A.; Rain, J.-C.; Beggs, J.D. Physical and genetic interactions of yeast Cwc21p, an ortholog of human SRm300/SRRM2, suggest a role at the catalytic center of the spliceosome. *RNA* **2009**, *15*, 2161–2173, doi:10.1261/rna.1908309.
39. Golubev, W. *Rhodosporidium babjevae*, a new heterothallic yeast species (Ustilaginales). *Syst. Appl. Microbiol.* **1993**, *16*, 445–449, doi:10.1016/S0723-2020(11)80278-X.
40. Chand Dakal, T.; Giudici, P.; Solieri, L. Contrasting patterns of rDNA homogenization within the *Zygosaccharomyces rouxii*

species complex. *PLoS One* **2016**, *11*, e0160744, doi: 0.1371/journal.pone.0160744

41. Conti, A.; Corte, L.; Pierantoni, D.C.; Robert, V.; Cardinali, G. What is the best lens? Comparing the resolution power of genome-derived markers and standard barcodes. *Microorganisms* **2021**, *9*, 1–19, doi:10.3390/microorganisms9020299.
42. Brandenburg, J.; Blomqvist, J.; Shapaval, V.; Kohler, A.; Sampels, S. Oleaginous yeasts respond differently to carbon sources present in lignocellulose hydrolysate. *Biotechnol. Biofuels* **2021**, 1–13, doi:10.1186/s13068-021-01974-2.
43. Wang, Q.; Sun, M.; Zhang, Y.; Song, Z.; Zhang, S.; Zhang, Q.; Xu, J.R.; Liu, H. Extensive chromosomal rearrangements and rapid evolution of novel effector superfamilies contribute to host adaptation and speciation in the basal ascomycetous fungi. *Mol. Plant Pathol.* **2020**, *21*, 330–348, doi:10.1111/mpp.12899.
44. Passoth, V.; Hansen, M.; Klinner, U.; Emeis, C.C. The electrophoretic banding pattern of the chromosomes of *Pichia stipiti*s and *Candida shehatae*. *Curr. Genet.* **1992**, *22*, 429–431, doi:10.1007/BF00352445.
45. Legrand, M.; Jaitly, P.; Feri, A.; d'Enfert, C.; Sanyal, K. *Candida albicans*: an emerging yeast model to study eukaryotic genome plasticity. *Trends Genet.* **2019**, *35*, 292–307, doi:10.1016/j.tig.2019.01.005.
46. Narayanan, A.; Vadnala, R.N.; Ganguly, P.; Selvakumar, P.; Rudramurthy, S.M.; Prasad, R.; Chakrabarti, A.; Siddharthan, R.; Sanyal, K. Functional and comparative analysis of centromeres reveals clade-specific genome rearrangements in *Candida auris* and a chromosome number change in related species. *MBio* **2021**, *12*, doi:10.1128/mBio.00905-21.
47. Gordon, J.L.; Byrne, K.P.; Wolfe, K.H. Mechanisms of chromosome number evolution in yeast. *PLoS Genet.* **2011**, *7*, 0–3, doi:10.1371/journal.pgen.1002190.
48. Hellborg, L.; Piškur, J. Complex nature of the genome in a wine spoilage yeast, *Dekkera bruxellensis*. *Eukaryot. Cell* **2009**, *8*, 1739–1749, doi:10.1128/EC.00115-09.
49. Chang, S.L.; Lai, H.Y.; Tung, S.Y.; Leu, J.Y. Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. *PLoS Genet.* **2013**, *9*, doi:10.1371/journal.pgen.1003232.
50. Vassiliadis, D.; Wong, K.H.; Blinco, J.; Dumsday, G.; Andrianopoulos, A.; Monahan, B. Adaptation to industrial stressors through genomic and transcriptional plasticity in a bioethanol producing fission yeast isolate. *G3 Genes, Genomes, Genet.* **2020**, *10*, 1375–1391, doi:10.1534/g3.119.400986.
51. Pierce, N.T.; Irber, L.; Reiter, T.; Brooks, P.; Brown, C.T. Large-scale sequence comparisons with sourmash. *F1000 Res.* **2019**, *8*, 1006, doi:10.12688/f1000research.19675.1
52. Civiero, E.; Pintus, M.; Ruggeri, C.; Tamburini, E.; Sollai, F.; Sanjust, E.; Zucca, P. Physiological and phylogenetic characterization of *Rhodotorula diobovata* DSBCA06, a nitrophilous yeast. *Biology (Basel).* **2018**, *7*, doi:10.3390/biology7030039.