BMC Genomics

RESEARCH ARTICLE Open Access

Whole genome sequencing and comparative genomic analysis of oleaginous red yeast *Sporobolomyces pararoseus* NGR identifies candidate genes for biotechnological potential and ballistospores-shooting

Chun-Ji Li[1,2], Die Zhao[3], Bing-Xue Li[1*], Ning Zhang[4], Jian-Yu Yan[1] and Hong-Tao Zou[1]

Abstract

Background: *Sporobolomyces pararoseus* is regarded as an oleaginous red yeast, which synthesizes numerous valuable compounds with wide industrial usages. This species hold biotechnological interests in biodiesel, food and cosmetics industries. Moreover, the ballistospores-shooting promotes the colonizing of *S. pararoseus* in most terrestrial and marine ecosystems. However, very little is known about the basic genomic features of *S. pararoseus*. To assess the biotechnological potential and ballistospores-shooting mechanism of *S. pararoseus* on genome-scale, the whole genome sequencing was performed by next-generation sequencing technology.

Results: Here, we used Illumina Hiseq platform to firstly assemble *S. pararoseus* genome into 20.9 Mb containing 54 scaffolds and 5963 predicted genes with a N50 length of 2,038,020 bp and GC content of 47.59%. Genome completeness (BUSCO alignment: 95.4%) and RNA-seq analysis (expressed genes: 98.68%) indicated the high-quality features of the current genome. Through the annotation information of the genome, we screened many key genes involved in carotenoids, lipids, carbohydrate metabolism and signal transduction pathways. A phylogenetic assessment suggested that the evolutionary trajectory of the order Sporidiobolales species was evolved from genus *Sporobolomyces* to *Rhodotorula* through the mediator *Rhodosporidiobolus*. Compared to the lacking ballistospores *Rhodotorula toruloides* and *Saccharomyces cerevisiae*, we found genes enriched for spore germination and sugar metabolism. These genes might be responsible for the ballistospores-shooting in *S. pararoseus* NGR.

Conclusion: These results greatly advance our understanding of *S. pararoseus* NGR in biotechnological potential and ballistospores-shooting, which help further research of genetic manipulation, metabolic engineering as well as its evolutionary direction.

Keywords: *Sporobolomyces pararoseus*, Genome sequencing, Comparative genomic, Biotechnological potential, Ballistospores-shooting, Evolutionary direction

* Correspondence: libingxue1027@163.com
[1]College of Land and Environment, Shenyang Agricultural University, Shenyang 110866, People's Republic of China
Full list of author information is available at the end of the article

Background

Genomic studies of the oleaginous red yeasts have gained increased attention due to their great biotechnological potential for biomass-based biofuel production [1–4]. The red yeast *Sporobolomyces pararoseus* (previously known as *Sporidiobolus pararoseus*) belongs to the order Sporidiobolales [5], which is classified in the subphylum Pucciniomycotina, an earliest branching lineage of Basidiomycota. This species has been documented from a broad spectrum of environments, ranging from freshwater and marine ecosystem, soil, and to plant tissue [6]. Biomass of this yeast constitutes sources of carotenoid, lipid, exopolysaccharide, and enzyme [7, 8]. Colony color of *S. pararoseus* includes shades of pink and red due to the presence of lipid droplets full of carotenoid pigments, containing β-carotene, torulene and torularhodin [9–11].

However, there is little information on bioactivity and nutritional value of torulene and torularhodin, perhaps because they are rare in food, but its structure and sparse evidence provide some hints. For example, tests performed on human and mice showed that torulene and torularhodin have anti-prostate tumor activity [12]. Furthermore, torularhodin represents antimicrobial properties, and it may become a new natural antibiotic [13]. Previous studies have reported their safety to be used as a food additive [14]. In consideration of their valuable properties, torulene and torularhodin might be successfully used as food and pharmaceutical industries in the future. Members of the order Sporidiobolales comprise of genera *Sporobolomyces*, *Rhodosporidiobolus*, and *Rhodotorula*, are known as competent producers of torulene and torularhodin [15]. Consequently, genetic manipulation of *S. pararoseus* for large-scale torulene and torularhodin production will be one of the major aims of future research efforts.

Additionally, *S. pararoseus* is regarded as one of the most efficient microorganisms for bioconversion of crude glycerol into lipids [16]. Lipids content comprises from 20% up to 60% of the dry biomass [16]. These lipids are not only important sources of polyunsaturated fatty acids, such as arachidonic acid and docosahexaenoic acid, but also for the production of biodiesel [8]. Microbial lipids' components are similar to that of vegetable oils, while have several advantages over vegetable oils [17, 18]. Such as a short life cycle, low space demands and independent of location and climates [19, 20]. Thus, the *S. pararoseus* also has been considered as potential feed stock for biodiesel industry [8].

Despite its long history of use for carotenoids fermentation, biodiesel production and ballistospores-shooting, very little is known about the basic genomic features of *S. pararoseus*. Advances in sequencing technology have drastically changed the strategies for studying genetic systems of microorganisms. Here, we present the first de novo genome assembly of *S. pararoseus*, as well as genes prediction and annotation. Subsequently, we performed a comparative analysis to investigate candidate orthologous and specific genes between *S. pararoseus*, *R. toruloides* and *S. cerevisiae*. The gene inventories provide vital insights into the genetic basis of *S. pararoseus* and facilitate the discovery of new genes applicable to the metabolic engineering of natural chemicals.

Results

Genome assembly and assessment

Here, the genome of oleaginous red yeast *S. pararoseus* NGR was sequenced using the Illumina Hiseq 2500 platform. A total of 8347 Mb raw data was generated from two DNA libraries: a pair-end library with an insert size of 500 bp (2631 Mb) and a mate-pair library with an insert size of 5 kb (5716 Mb). After, removing adapters, low-quality reads and ambiguous reads, we obtained 6073 Mb clean data (Q20 > 95%, Q30 > 90%) for genome assembly. For the genome size estimation of *S. pararoseus* NGR, we calculated the total 15 *k*-mer number is 705,505,006 and the *k*-mer depth is 28.41. According to the 15-mer depth frequency distribution formula, the estimated genome size of *S. pararoseus* NGR was calculated to be 24.44 Mb. Our final assembly consists of 54 scaffolds, a N50 length of 2,038,020 bp, the longest length scaffold of 4,025,647 bp, the shortest length scaffold of 513 bp, a GC content of 47.59% and a size of 20.9 Mb (85.52% of the estimated genome size). We identified 5963 genes in the genome with an average length of 1620 bp and a mean GC-content of 47.26% that occupied 55.07% of the genome. The results of BUSCO alignment showed that our final assembly contains 1273 complete BUSCOs (95.4%), of which 1268 were single-copy, while 5 were duplicated (Additional file 1). For the RNA-seq results, a total of 2662 Mb raw reads were generated. Using assessment of RNA-seq data, we found 98.68% (5884) of genes predicted in the NGR genome regions and 767 novel genes were expressed (Additional file 2). In addition, the RNA-seq data showed that 74.07% of reads matched to exon regions, 4.03% to intron regions, and 21.9% to intergenic regions. These reads are aligned to the intron region, mostly due to intron retention or alternative splicing events. In total 488 SNPs/InDel (Additional file 3) were identified when comparing RNA-seq data with the NGR genome sequences. From the RNA-seq data, we also identified the boundaries of 5'UTR and 3'UTR of 2772 genes (Additional file 4). Both BUSCO alignment and RNA-seq mapping suggested that our current genome assembly is characterized as high-quality, completeness and accuracy [21].

Functional annotation

Among the 5963 predicted genes, 4595 (77.05%) genes could be annotated by BLASTN (E-value <1e^{-5}) using NCBI Nr databases based on sequence homology. In addition, 1940 (32.53%), 3002 (50.34%), 4237 (71.05%), 1806 (30.3%) and 4659 (78.13%) genes could be annotated according to KEGG, KOG, NOG, SwissProt, and TrEMBL databases, respectively. It should be noted that among these genes assigned to Nr database, the top 3 species of matched genes number are *R. toruloides* (3484, 75.82%), *Rhodotorula glutinis* (555, 12.08%) and *Microbotryum violaceum* (340, 7.4%). Furthermore, 4057 genes could be classified into three Gene Ontology (GO) categories (Additional file 5): cellular component (1883 genes), biological process (2802 genes), and molecular function (3388 genes). In addition, 194 tRNA, 1753 dispersed repetitive sequences, 2092 tandem repeats, 1178 minisatellite DNA (Additional file 6) and 659 microsatellites DNA (Additional file 7) were identified in the genome. A total of 132,885 full-length TEs were predicted in the NGR whole genome. These TEs include 838 LTR-REs, 59 SINE-REs, 31 RC-REs, 598 DNA transposons, 208 LINE-REs and 7 Unknowns, of which 47.17% are Class LTR element, mainly assigned to Gypsy (346) and Copia (190). The full-length TEs totally comprised 132,885 bp, accounting for 0.61% of the NGR whole genome.

Based on KEGG pathways mapping, we annotated the coding genes of candidate for biotechnological potential in the NGR genome. A summary of the candidates (Additional files 8, 9, 10 and 11 for details) is presented as following: 1) carotenoids biosynthesis, including *crtI* (phytoene desaturase, GenBank: KR108014) [22], *crtYB* (lycopene cyclase/phytoene synthase, GenBank: KR108013) [23], *crtE* (GGPP synthase, GenBank: KY652916), and other genes encoding hydroxylase, monooxygenase, or ketolase/carboxylase which might be responsible for the transformation from torulene to torularhodin; 2) lipid metabolism, including genes encoding acetyl-CoA carboxylase, acyl-CoA oxidase, phospholipid: diacylglycerol acyltransferase, glycerol 3-phosphate dehydrogenase; 3) carbohydrate metabolism, including genes encoding pyruvate dehydrogenase, pyruvate carboxylase and acyl-CoA: diacylglycerol acyltransferase; 4) stress responses, including genes involved in MAPK signaling pathway and calcium signal transduction.

Phylogenetic relationships between red yeasts of the order Sporidiobolales

Among phylum Basidiomycetes yeasts, there are a number of species that grow as pigmented colonies, and are for this reason known as red yeast [24]. Among them, 42 red yeasts belong to the order Sporidiobolales. Recently, the order Sporidiobolales has been reconstructed, including three genera *Sporobolomyces* (17 species) *Rhodosporidiobolus* (9 species) and *Rhodotorula* (16 species) [5, 25]. In order to determine the possible evolutionary trajectories between these red yeasts, we constructed the phylogenetic tree with available 26S rDNA sequences. As shown in Fig. 1, as for genus *Sporobolomyces*, the NGR showed a closer evolutionary relationship with *S. ruberrimus* and *S. koalae* than the other species, particularly for *S. johnsonii* and *S. salmonicolor*. The genus *Rhodosporidiobolus* situates a closer evolutionary relationship with *Rhodotorula* than with *Sporobolomyces*. The ballistospores are not uniform in the species of order Sporidiobolales, however, being a specialized mode of genus *Sporobolomyces* but absent in *Rhodotorula* and two characterized species of *Rhodosporidiobolus (R. lusitaniae and R. colostri)* [26–28]. It suggests that the same ancestor of *Sporobolomyces* and *Rhodosporidiobolus* species shoot ballistospores. However, the ballistospores-shooting ability was gradually lost in *R. lusitaniae/R. colostri* or other undescribed *Rhodosporidiobolus* species. Subsequently, some *Rhodosporidiobolus* species of lacking ballistospores-shooting ability has undergone a series of evolutionary processes to form *Rhodotorula* species. While these basic hypotheses are non-controversial, further verification basing on discovering more new Sporidiobolales species and obtaining their genome data is required.

Comparative analysis of protein families and genes

The NGR genome has predicted 5963 protein-coding genes, and the most of genes were annotated into the specie *R. toruloides* NP11. This motivates us to perform a comparative genomic analysis between *S. pararoseus* NGR and *R. toruloides* NP11. In order to exclude the inherent quality of yeast, we added the model yeast *S. cerevisiae* S288C as a control. As shown in Fig. 2a, we compared the distribution of genes among the three yeasts. In order to identify species-specific gene/protein families, we performed pairwise comparisons using a series of BLASTX searches within the three species. As shown in Fig. 2b, a total of 14,408 protein families were identified based on sequence similarities (5751 families for the NGR, 7935 families for NP11, and 5485 families for S288C). 1975 (2077 genes), 4102 (4159 genes), and 4485 (4736 genes) protein families were species-specific in *S. pararoseus* NGR, *R. toruloides* NP11 and *S. cerevisiae* S288C, respectively. As shown in Fig. 2c, we conducted the GO analysis using respective species-specific genes of the three species. As for the genes of *S. pararoseus* NGR, 106 (16.4%), 280 (43.3%) and 261 (40.3%) terms were enriched in the CC, MF and BP, respectively. We found that the significantly enriched GO terms of the *S. pararoseus* NGR species-specific genes containing, CC: nucleus, membrane, and integral to membrane; MF:

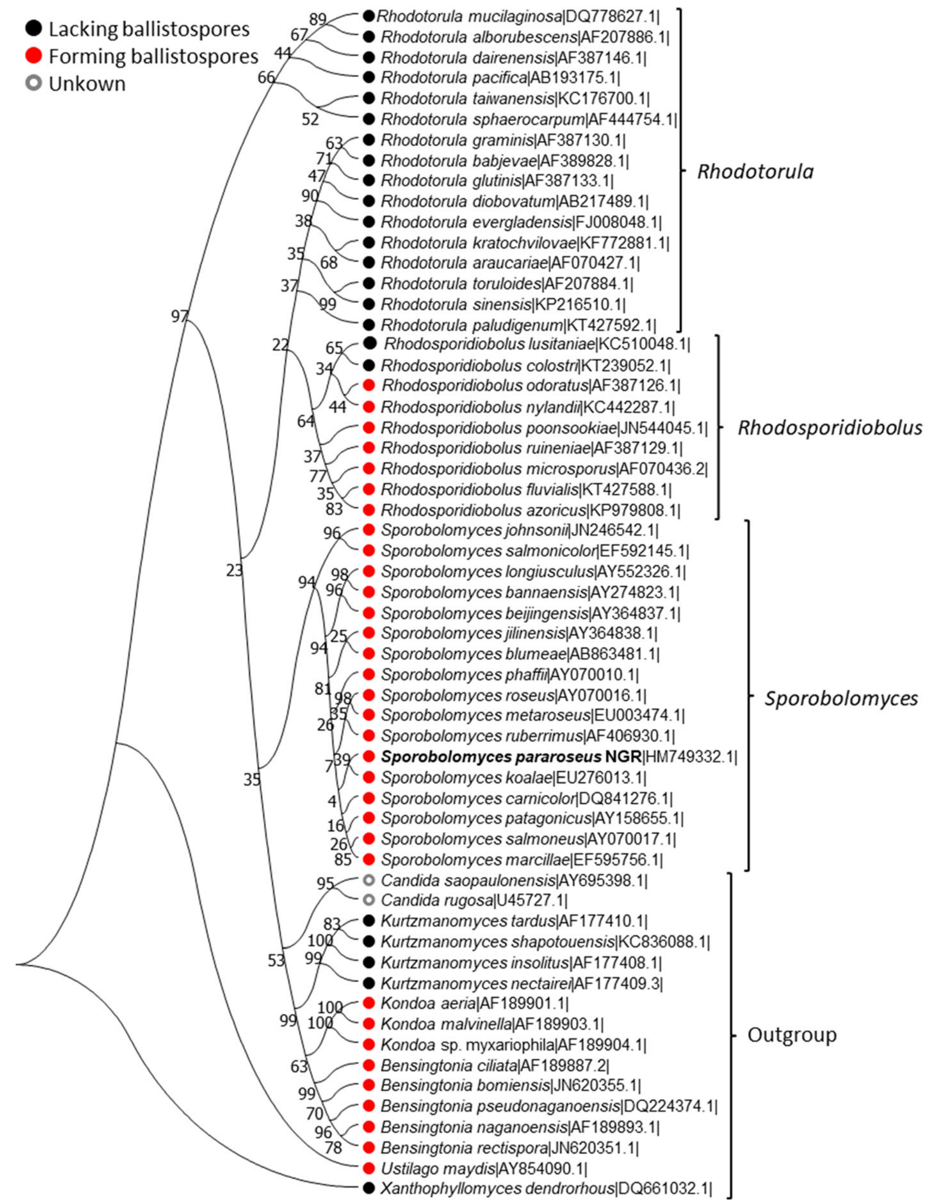
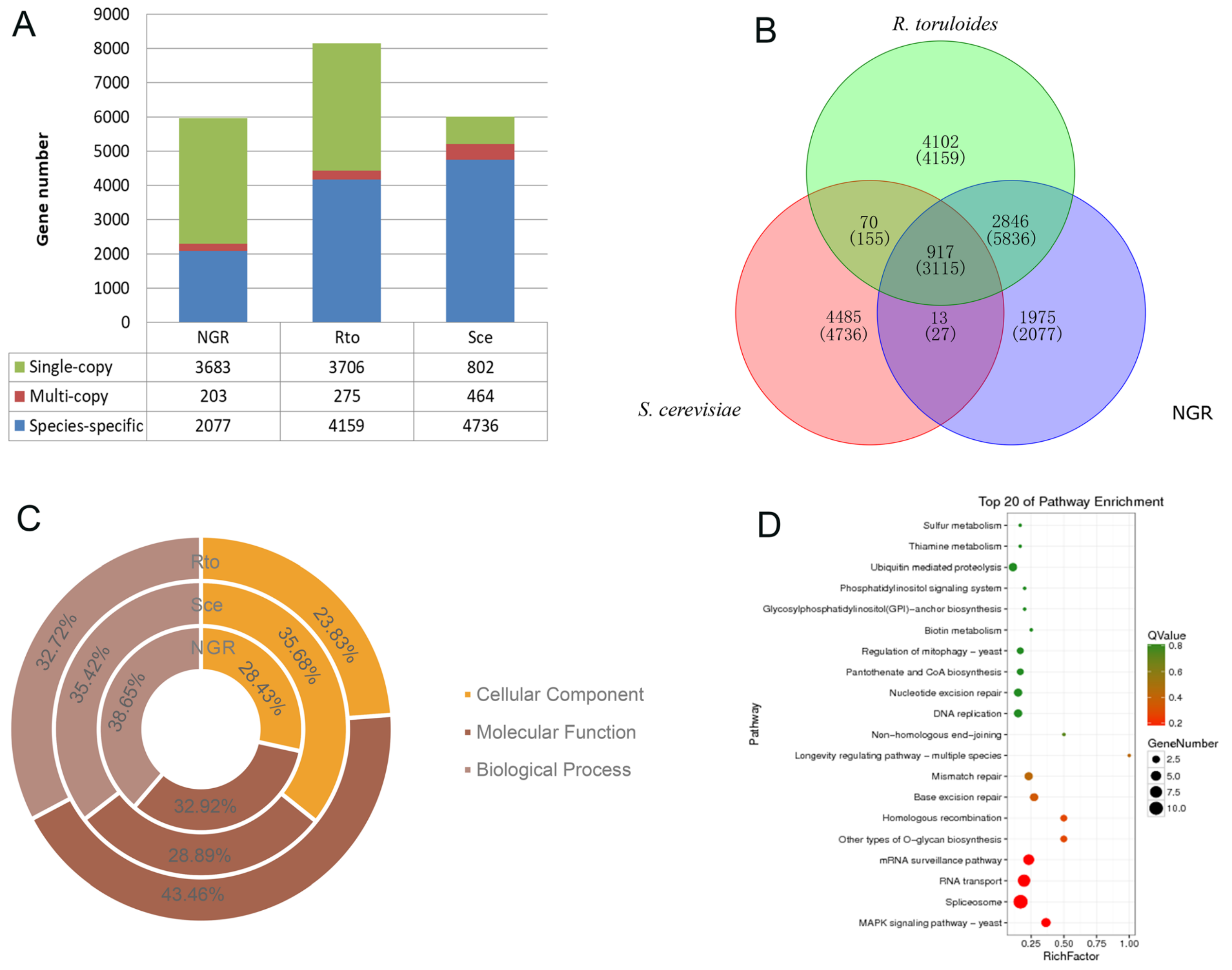

● Lacking ballistospores
● Forming ballistospores
○ Unkown
Rhodotorula mucilaginosa|DQ778627.1|
Rhodotorula alborubescens|AF207886.1|
Rhodotorula dairenensis|AF387146.1|
Rhodotorula pacifica|AB193175.1|
Rhodotorula taiwanensis|KC176700.1|
Rhodotorula sphaerocarpum|AF444754.1|
Rhodotorula graminis|AF387130.1|
Rhodotorula babjevae|AF389828.1|
Rhodotorula glutinis|AF387133.1|
Rhodotorula diobovatum|AB217489.1|
Rhodotorula evergladensis|FJ008048.1|
Rhodotorula kratochvilovae|KF772881.1|
Rhodotorula araucariae|AF070427.1|
Rhodotorula toruloides|AF207884.1|
Rhodotorula sinensis|KP216510.1|
Rhodotorula paludigenum|KT427592.1|
Rhodotorula
Rhodosporidiobolus lusitaniae|KC510048.1|
Rhodosporidiobolus colostri|KT239052.1|
Rhodosporidiobolus odoratus|AF387126.1|
Rhodosporidiobolus nylandii|KC442287.1|
Rhodosporidiobolus poonsookiae|JN544045.1|
Rhodosporidiobolus ruineniae|AF387129.1|
Rhodosporidiobolus microsporus|AF070436.2|
Rhodosporidiobolus fluvialis|KT427588.1|
Rhodosporidiobolus azoricus|KP979808.1|
Rhodosporidiobolus
Sporobolomyces johnsonii|JN246542.1|
Sporobolomyces salmonicolor|EF592145.1|
Sporobolomyces longiusculus|AY552326.1|
Sporobolomyces bannaensis|AY274823.1|
Sporobolomyces beijingensis|AY364837.1|
Sporobolomyces jilinensis|AY364838.1|
Sporobolomyces blumeae|AB863481.1|
Sporobolomyces phaffii|AY070010.1|
Sporobolomyces roseus|AY070016.1|
Sporobolomyces metaroseus|EU003474.1|
Sporobolomyces ruberrimus|AF406930.1|
Sporobolomyces pararoseus NGR|HM749332.1|
Sporobolomyces koalae|EU276013.1|
Sporobolomyces carnicolor|DQ841276.1|
Sporobolomyces patagonicus|AY158655.1|
Sporobolomyces salmoneus|AY070017.1|
Sporobolomyces marcillae|EF595756.1|
Sporobolomyces
Candida saopaulonensis|AY695398.1|
Candida rugosa|U45727.1|
Kurtzmanomyces tardus|AF177410.1|
Kurtzmanomyces shapotouensis|KC836088.1|
Kurtzmanomyces insolitus|AF177408.1|
Kurtzmanomyces nectairei|AF177409.3|
Kondoa aeria|AF189901.1|
Kondoa malvinella|AF189903.1|
Kondoa sp. myxariophila|AF189904.1|
Bensingtonia ciliata|AF189887.2|
Bensingtonia bomiensis|JN620355.1|
Bensingtonia pseudonaganoensis|DQ224374.1|
Bensingtonia naganoensis|AF189893.1|
Bensingtonia rectispora|JN620351.1|
Ustilago maydis|AY854090.1|
Xanthophyllomyces dendrorhous|DQ661032.1|
Outgroup
89
67
44
66
52
63
71
47
90
38
35
68
37
99
97
22
65
34
44
64
37
77
35
83
23
96
94
98
96
25
94
81
98
35
26
7
39
35
4
16
26
85
95
83
100
53
99
100
99
100
63
99
70
96
78

Fig. 1 Phylogenetic tree of the order Sporidiobolales yeasts and outgroup species were constructed by Neighbor-Joining method and bootstrap analysis (1000 replicates) based on the alignment of the 26S rDNA sequence. The strain NGR font has been bolded. The numbers at the nodes indicate the bootstrap probabilities of the particular branch. Organisms belonging to the same genus have been represented on the right-side, representing as *Rhodotorula*, *Rhodosporidiobolus*, and *Sporobolomyces*. The scale (value: 0.01) representing nucleotide substitution per side is displayed. The accession numbers of the corresponding database entries are listed in behind the Latin name of each species. The ballistospores-forming ability for each entry of the phylogenetic tree is represented in front of the Latin name of each species. A red dot for those forming ballistospores, black dot for those not forming them and gray for those for which no information is available

protein binding, DNA binding, and zinc ion binding; BP: regulation of transcription-DNA-dependent, transport, transmembrane transport, intracellular protein transport, carbohydrate metabolic process and oxidation-reduction process. Subsequently, we carried out the KEGG pathway mapping of *S. pararoseus* NGR species-specific genes. As shown in Fig. 2d, the significantly enriched pathways (Top 20) of the *S. pararoseus* NGR species-specific genes including MAPK signaling pathway-yeast, spliceosome, RNA transport, and mRNA surveillance pathways (Additional file 12).

Among the species-specific genes, NGR-1A3721 that assigned to the GO term of spore germination (GO: 0009847) was considered to be one of the candidates for the formation of ballistospores. Moreover, the species-specific genes of the NGR involved in the KEGG



	NGR	Rto	Sce
🟩 Single-copy	3683	3706	802
🟥 Multi-copy	203	275	464
🟦 Species-specific	2077	4159	4736

Fig. 2 Comparative genomic analysis of *S. pararoseus* NGR, *R. toruloides* NP11, and *S. cerevisiae* S228C. **a** Distribution of single-copy, multi-copy and species-specific genes among three yeasts. **b** Venn diagram representation of shared/unique genes in *S. pararoseus* NGR and comparison with those in *R. toruloides* and *S. cerevisiae*. **c** Percentage of the gene numbers of species-specific protein families matched to different GO categories, in three yeast genomes, respectively. **d** Top 20 enriched KEGG pathways of species-specific genes in *S. pararoseus* NGR genomes. A rich factor is the ratio of the enriched genes numbers to total gene number in this pathway. The greater the rich factor, the higher the degree of enrichment. The Q-value ranges from 0 to 1 and the closer it is to zero, the more significant the enrichment

pathways of sugar metabolism, including amino sugar and nucleotide sugar metabolism (ko00520), pentose and glucuronate interconversions (ko00040), starch and sucrose metabolism (ko00500), galactose metabolism (ko00052), fructose and mannose metabolism (ko00051), and butanoate metabolism (ko00650) might be related to the ballistospores dissemination as reported in previous studies [29, 30]. Recently, Ianiri et al. reported that 3-hydroxyacyl-CoA dehydratase gene *Phs1* is not only responsible for the very long chain fatty acid biosynthesis, but also for the ballistospores-shooting in *Sporobolomyces* sp. IAM 13481 [31]. However, we found this *Phs1* gene in the both *S. pararoseus* NGR and *R. toruloides* genomes. Moreover, the *Phs1* gene was not strong positive or negative selected in substitution rates (Ka/Ks) analysis. Therefore, the *Phs1* should be an indirect determinant of the ballistospores-shooting in genus *Sporobolomyces*.

Discussion

S. pararoseus is recognized as a kind of biotechnologically important oleaginous red yeast, which potentially can be used for biodiesel production as well as other important bio-products, such as carotenoids, enzymes and exopolysaccharide [32]. However, little is currently known about its genomic sequence and features. In the present study, the genome of *S. pararoseus* NGR will enable direct access to the genes responsible for its biology and biotechnological potential. To date, the only yeast belonging to the *Sporobolomyces* genus for which genome sequence is available is *S. salmonicolor* CBS 6832 [33]. As shown in Table 1, we compared the general

Table 1 Genome features of *S. pararoseus* NGR and *S. salmonicolor* CBS 6832

Features	NGR	CBS 6832
Genome assembly size (Mb)	20.9	20.5
Number of contings	135	744
Number of scaffolds	54	395
Scaffolds N50 length (bp)	2,038,020	538,656
GC contents (%)	47.59	61.3%
Predicted genes (Nr)	5963	5147
Sequence platform	Illumina	Illumina + PacBio

genome features of *S. pararoseus* NGR and *S. salmonicolor* CBS 6832. The genome assembly quality of NGR is better than CBS 6832. The genome GC-content of CBS 6832 (61.3%) is higher than NGR (47.59%), but the predicted genes amount of CBS 6832 (5147) is less than NGR (5963). The *S. pararoseus* NGR genome will also serve as a useful basis of comparative genomics studies to investigate functional peculiarities specific to this yeast and its relative lineage within the *Sporobolomyces* clade.

Moreover, one of the most notable characteristics of *S. pararoseus* is the process of ballistospores discharge. Ballistospores discharge is a unique type of spore produced by phylum Basidiomycetes fungi, however, does not occur in other fungal phyla [34]. As shown in Fig. 3, the *S. pararoseus* NGR was patched on agar medium to form colonies, and the ballistospores are vertically shot into the lid of the plate to form a "mirror" with their colonies. Ballistospores-booting is the main reason for this eukaryotic lineage colonizing in the most ecosystems. The *Sporobolomyces* species are endowed with many similar phenotypes with *Rhodotorula* species, such as carotenoids and lipid production, and morphological characteristics. However, an obvious difference between them is that *Rhodotorula* species are usually considered as marine microorganisms, and does not produce ballistospores. The ancestor of the order Sporidiobolales might be certain *Sporobolomyces* species and lived on land without the convenience of an aqueous environment. Dissemination of ballistospores is for finding new nutrient sources. As they entered and adapted to the marine environment, they gradually reduce the efficacy of ballistospores-shooting to form the *Rhodosporidiobolus* species and further lost ballistospores to evolve into the *Rhodotorula* species. Because of the ballistospores-shooting is widely considered as a biological process of energy consumption. When they exposed to excessive sea water, its energy should be preserved as much as possible to resist cold and high salt stresses, instead of discharging ballistospores. Both cold and salt stresses might play critical roles in positive selection and rapid evolution of genera *Sporobolomyces* to *Rhodotorula* species.

The Ka/ Ks ratio is widely considered to be an indicator of selective pressure during evolution [35]. To assess the overall difference in the selective restriction of gene levels within genera *Sporobolomyces* and *Rhodotorula*, the free ratio model was used to calculate the substitution rate for each orthologous gene [36]. Among the 700 pair's single-copy homologous genes, we found that 80 pairs with a Ka/Ks value 0.1 < Ka/Ks < 0.5, 165 pairs with a Ka/Ks value Ka/Ks < 0.1, and 455 pairs with a Ks value = 0 (Additional file 13). The top four functional

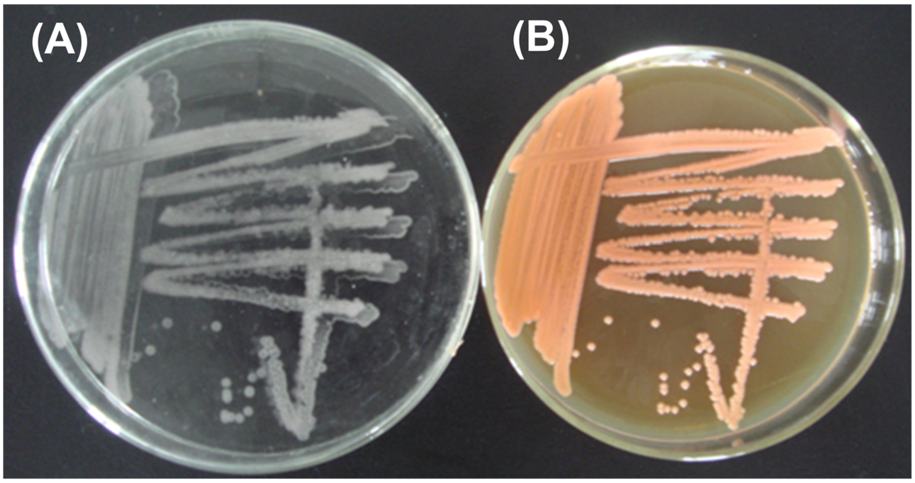


Fig. 3 Ballistospores shoot in the *S. pararoseus* NGR. **a** The ballistospores have shot to the lid of the YPD plate to form mirror symmetry; **b** Colony morphology of the NGR patched on YPD plate. The trajectories of the ballistospores-shooting are perpendicular to the surface of the YPD plate. At the base of the ballistospores is a liquid droplet resulting from the drop coalescence that powers the explosive launch. The process of ballistospores-shooting is termed as "Buller's drop" [29]

KEGG terms enriched among the negatively selected genes were "Carbohydrate metabolism", "Translation", "Lipid metabolism", and "Amino acid metabolism", which are associated with energy metabolism and the progress of protein synthesis or hydrolysis. This result indicates that the *Rhodotorula* species might have evolved a better energy metabolism and osmoregulation system to adapt to the marine environments and delay or prevent potential injury. But, the ballistospores-shooting is not necessary for its spreading in the marine environments. Given that the genes responsible for ballistospores-shooting still remain unknown. Our results provided valuable genetic data for the further characterization of the molecular mechanisms for ballistospores-shooting.

Conclusions

Here, the high-quality *S. pararoseus* genome was reported. It established a genomic basis for further studying on its carotenoids, lipid, carbohydrate metabolism and stress responses. Furthermore, we proposed the evolutionary trajectories that *Rhodotorula* species were evolved from *Sporobolomyces* through the mediator *Rhodosporidiobolus.* Comparative genomic analysis revealed that the species-specific genes of *S. pararoseus* NGR related to spore germination and sugar metabolism, which might be involved in ballistospores-shooting. In conclusion, our work provides an important foundation for genes with potential biotechnological applications and foster comparative genomics studies to elucidate fundamental biological processes and evolutionary consequences of the order Sporidiobolales.

Methods

Strain material and DNA extraction

S. pararoseus NGR was isolated from strawberry fruit in the greenhouse of Shenyang Agricultural University (41°49′N, 123°34′E) in Shenyang City, Liaoning Province, China. Species identification was performed through morphological and molecular methods. The available GenBank accession number of *S. pararoseus* NGR 26S rDNA is HM749332. The strain number is recorded in the China General Microbiological Culture Collection Center as CGMCC 2.5280. *S. pararoseus* NGR cultures were grown for 72 h in 250 mL Erlenmeyer baffle flasks containing 50 mL of the YPD medium (10 g/L yeast extract, 20 g/L peptone and 20 g/L glucose, pH 6.5 ± 0.5) at 28 °C on a rotary shaker at 180 rpm. Genomic DNA of *S. pararoseus* NGR was extracted using the DNAiso Reagent kit (Code No.: 9770A) (Takara Bio, Dalian, China) according to the manufacturer's protocols. The extracted genomic DNA was subjected to quality control by agarose gel electrophoresis and quantified by Qubit 2.0 fluorometer (Life Technologies, USA). The obtained genomic DNA (≥500 ng/μL) was used for whole genome sequencing and PCR verification.

Genome sequencing

Genome sequencing of the strain NGR was performed utilizing the Illumina HiSeq 2500 platform (Illumina, USA). In order to obtain a high-quality de novo assembly, the strategy used was to combine data generated from standard short insert paired-end libraries with those from mate-pair libraries. Two DNA libraries were constructed: a paired-end library with an insert size of approximately 500 bp using TruSeq Nano DNA Kit (Illumina, USA) and a mate-pair library with an insert size of approximately 5 kb using Nextera DNA Library Preparation Kit (Illumina, USA). The 500 bp library and the 5 kb library were sequenced using the PE125 strategy at the Novogene Bioinformatics Technology Co., Ltd. (Beijing, China). After sequencing, quality control of the raw reads was performed, which involved trimming the reads using Trimmomatic (version 0.20) [37] by removing the Nextera adapter and linker sequences (for the mate-pair libraries) and TruSeq adapters (for the pair-end libraries); removing reads containing more than 10% of unknown nucleotides (N); removing low quality reads containing more than 50% of low quality (Q-value≤10) bases. For the trimmed reads, the online program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to plot quality score and sequence length distribution. Finally, the software ABySS (version 1.3.5) [38] was used to visualize the library complexity by plotting the *k*-mer profile of the reads. With these data, the genome size of the NGR was estimated by *k*-mer distribution (15 depth frequency) through the program KmerGenie (version 1.5621) with default parameters (inspired by FastQC) [39].

Genome assembly

The filtered reads were assembled by SOAPdenovo2 (http://soap.genomics.org.cn/soapdenovo.html, version 2.0) under *k*-mer size of 15 [40–42] to generate scaffolds. The assembler SOAPdenovo2 follows the classic De Bruijn graph representation [43]. All reads were used for further gap closure using SOAPdenovo GapCloser Module as described in previous studies [42, 44]. Standard assembly statistics were obtained including: number of scaffolds, N50 (length N for which 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value), N90 (same as N50 but using 90% instead), GC-content (%), the longest length scaffold, the shortest length scaffold, and total assembly length considering only scaffolds > 500 bp.

Gene prediction and annotation

After obtaining the whole-genome sequence of the NGR, genes were predicted using the "Training module" and "Prediction module" of WebAUGUSTUS Service (http://bioinf.uni-greifswald.de/webaugustus/) with default parameters (strand = both; single strand = true; noInFrameStop = true) [45]. Repetitive sequences and distribution of transposable element (TEs) annotation were performed by using the program RepeatMasker (http://www.repeatmasker.org/, version v4.0.7) with default parameters based on libraries generated by different strategies: de novo-based, signature-based, and homology-based methods [46]. Tandem repeats were analyzed using the software Tandem Repeat Finder (http://tandem.bu.edu/trf/trf.html, version 4.09) with default parameters [47]. Ribosome RNA (rRNA) genes were predicted using the program rRNAmmer (http://www.cbs.dtu.dk/services/RNAmmer/, version 1.2) [48] with default parameters. Transfer RNA (tRNA) genes were predicted using the program tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/, version 2.0.4) with default parameters [49]. Non-coding RNAs were predicted by BLAST against Rfam (RNA families) database (http://rfam.xfam.org/, version 13.0) with default parameters [50, 51]. Functional annotation of the predicted genes was performed by similarity using BLAST against diverse public databases, including: the NCBI's non-redundant protein (Nr) database, UniProt/Swiss-Prot, Cluster of Orthologous Groups of proteins (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Protein Families (Pfam) [52]. BLASTN searches were carried out using an E-value less than $1e^{-5}$, minimal alignment length percentage larger than 40% (identity ≥40%, coverage ≥40%). The Gene Ontology (GO) annotations was performed using the program Blast2GO with default parameters [53] and GO-term classification was conducted based on the Nr annotations.

Identification of orthologous genes

Annotations of coding sequences and proteins of *Rhodotorula toruloides* NP11 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000320785.1/) and *Saccharomyces cerevisiae* S288C (https://www.ncbi.nlm.nih.gov/assembly/GCF_000146045.2/) were downloaded from NCBI Assembly database. The NGR, NP11 and S288C genome sequence alignments were performed in an all-against-all comparison using the MUMmer 3 package (http://mummer.sourceforge.net/, version 3.2.2) with default parameters [54]. Comparative genome analyses were performed at protein level. The software OrthoMCL (https://orthomcl.org/orthomcl/, version 2.0) was used to generate core-orthologs for the NGR, NP11 and S288C whole proteomes datasets with default parameters [55]. Subsequently, all the putative proteins of the three yeast species and core-orthologs were aligned (all against all) using BLASTP (http://www.ncbi.nih.gov/BLAST/) and a score for each pair of proteins with significant matches was assigned with a cut-off value of 1×10^{-7} [56]. These species-specific genes of the NGR were used for screening the candidates for the formation and dissemination of ballistospores. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed using DAVID functional annotation tool (https://david.ncifcrf.gov/tools.jsp, version 6.8) with default parameters [57, 58]. Subsequently, we used the GO/KEGG enrichment results to further screen the candidate genes for ballistospores-shooting of the NGR.

RNA-seq and gene models prediction

RNA was extracted from the NGR cells using Trizol Reagent Kit (Invitrogen, USA) and then checked by 1% agarose gel electrophoresis and a NanoPhotometer spectrophotometer (IMPLEN, CA, USA). The complementary DNA (cDNA) libraries were constructed using NEB Next Ultra RNA Library Prep Kit for Illumina (NEB, USA) following manufacturer's recommendations. The constructed products were purified and then amplified with PCR to obtain the high-quality cDNA library and sequenced at an Illumina Hiseq 2000 platform using the SE100 strategy at the Novogene Bioinformatics Technology Co., Ltd. (Beijing, China). Raw reads of fastq format were firstly processed through the fastp program (version 0.18.0) (https://github.com/OpenGene/fastp) with default parameters [59]. Clean reads were obtained by the fastp program with following parameters: 1) removing reads containing adapters; 2) removing reads containing more than 10% of unknown nucleotides (N); 3) removing low quality reads containing more than 50% of low quality bases [59]. Q20, Q30, GC-content and sequence duplication level of the clean reads were calculated to assess the quality of clean reads. All the downstream analyses were based on clean reads with high quality.

These high-quality clean reads were first mapped to the NGR genome using the software HISAT2 (version 2.1.0) with "-rna-strandness RF" and other parameters set as a default [60, 61]. The mapped reads were assembled to reconstruct transcripts by using StringTie (version 1.3.1) in a "reference-based" approach [61, 62]. The reconstructed transcripts that are not annotated in the NGR genome are defined as novel genes [62]. Gene expression was then measured in fragments per kilobase of exon per million fragments mapped (FPKM) using the program StringTie (version 1.3.1) with default parameters [62]. After mapping reads to the NGR genome, the software HISAT2 (version 2.1.0) was used in reconstruction of transcripts which may extend the 5′ untranslated region (5′ UTR) or 3′ untranslated region (3′ UTR) of gene to optimize the gene structure. Calling variants of transcripts were carried out using the Genome Analysis Toolkit (GATK, version 4.1.4.1) with default parameters [63]. The software ANNOVAR (by default parameters) was used for single nucleotide polymorphisms (SNPs) and insertion-deletion (InDel) annotation [64].

Assessment of genome completeness

Both orthologous gene alignment and RNA-seq data mapping were employed to evaluate our genome completeness. The program BUSCO (version 3.0.1) was applied to align the orthologs of the NGR to a reference gene set of basidiomycota_odb9 with default parameters [65].

Substitution rate estimation and selection analyses

The substitution rates (Ka/Ks, the ratio of nonsynonymous to synonymous substitutions) for each orthologous gene were used to evaluate the overall differences in the selective pressure at the gene level within the three yeast lineages, using the free ratio model in KaKs_ Calculator Toolbox (version 2.0) software with default parameters [66, 67]. The genes with p-value< 0.05 and a higher Ka/Ks value (Ka/Ks > 1) were considered to be the positively selected genes, as described in a previous study [68].

Phylogenetic analysis

All 26S rDNA nucleotide sequences for the phylogenetic analyses were from NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore). All sequences were processed with the software MEGA (version 7.0) using Muscle alignment with UPGMB clustering method [69, 70]. Subsequently, all sequences were trimmed manually to remove the unaligned sequences of 3′ and 5′ end. The phylogeny was tested by applying Bootstrap method [71], Bootstrap values expressed as percentages of 1000 replications, are given at the branching points. Phylogenetic tree was constructed using these trimmed sequences by the Neighbor-Joining method. All analysis was carried out using default parameters without bootstrapping in the software MEGA 7.0.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6593-1.

Additional file 1. Genome completeness analysis through applied BUSCO software (version 3.0.1) to align the orthologs of the NGR to a reference gene set of basidiomycota_odb9. (XLS 71 kb)

Additional file 2. All genes expression profile and functional annotations resulted from RNA-seq analysis. (XLS 1904 kb)

Additional file 3. SNP/InDel annotations resulted from RNA-seq data. (XLS 48 kb)

Additional file 4. Gene structure optimization resulted from RNA-seq data.

Additional file 5. GO categories of 4057 genes resulted from genome analysis.

Additional file 6. Mini-satellite DNA annotation file of NGR.

Additional file 7. Micro-satellites DNA annotation file of NGR.

Additional file 8. All gene annotation information resulted from genome analysis.

Additional file 9. Genome annotations file of NGR.

Additional file 10. CDS/cDNA sequences resulted from genome analysis.

Additional file 11. Protein sequences resulted from genome analysis.

Additional file 12. GO/KEGG enrichment results of species-specific genes of the NGR.

Additional file 13. Ka/Ks results of single-copy homologous gene (NGR vs NP11).

Abbreviations

BP: Biological process; CC: Cellular component; COG: Cluster of Orthologous Groups of proteins; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; MF: Molecular function; Nr: NCBI's non-redundant protein database; Pfam: Protein Families; Rfam: RNA families

Acknowledgements

The authors would like to express gratitude to all those who helped us during the writing of this paper. We also thank Guangzhou Genedenovo Biotechnology Co., Ltd. for assisting in bioinformatics analysis. Particularly, Chun-Ji Li thanks the priceless care, love and support of Die Zhao over the years.

Authors' contributions

CJL performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote and revised the paper, and prepared the figures and Tables. DZ provided multiple technical supports in bioinformatics analysis. CJL and BXL conceived the study ideas and designed the experiments. BXL and NZ were responsible for the funding. JYY prepared experimental materials for strain culture. NZ and HTZ supervised the written process of the paper. All authors have read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (2017YFD0200807), the National Natural Science Foundation of China (31271818), and the Natural Science Foundation Guidance Project of Liaoning Province (2019-ZD-0711).

Availability of data and materials

All the raw sequence data are available via GenBank under the SRA accessions SRX3638123-SRX3638125 (three raw data for the transcriptome) and SRR9733737- SRR9733739 (six raw data for the whole genome). This *S. pararoseus* strain NGR Whole Genome Shotgun project has been deposited at DDBJ/ ENA/ GenBank under the accession RSDY00000000; BioProject: PRJNA505991; BioSample: SAMN10440612. The version described in this article is version RSDY01000000.

Ethics approval and consent to participate

The strain NGR used in this study was obtained from strawberry fruit in the greenhouse of Shenyang Agricultural University, and it is not an endangered species. Its strain number is recorded in the China General Microbiological Culture Collection Center as CGMCC 2.5280. The collection of the microbial materials was complied with institutional and national guidelines of China.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Land and Environment, Shenyang Agricultural University, Shenyang 110866, People's Republic of China. [2]College of Agriculture and Biology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, People's Republic of China. [3]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, People's Republic of China. [4]College of Biological Science and Technology, Shenyang Agricultural University, Shenyang 110866, People's Republic of China.

Received: 27 October 2019 Accepted: 19 February 2020
Published online: 24 February 2020

References

1. Kumar S, Kushwaha H, Bachhawat AK, Raghava GPS, Ganesan K. Genome sequence of the oleaginous red yeast *Rhodosporidium toruloides* MTCC 457. Eukaryot Cell. 2012;11(8):1083–4.
2. Paul D, Magbanua Z, MA II, French T, Bridges SM, Burgess SC, et al. Genome sequence of the oleaginous yeast *Rhodotorula glutinis* ATCC 204091. Genome Announc. 2014;2(1):e00046–14.
3. Lin X, Wang Y, Zhang S, Zhu Z, Zhou YJ, Yang F, et al. Functional integration of multiple genes into the genome of the oleaginous yeast *Rhodosporidium toruloides*. FEMS Yeast Res. 2014;14(4):547–55.
4. Sambles C, Middelhaufe S, Soanes D, Kolak D, Lux T, Moore K, et al. Genome sequence of the oleaginous yeast *Rhodotorula toruloides* strain CGMCC 2.1609. Genom Data. 2017;13:1–2.
5. Urbina H, Aime MC. A closer look at Sporidiobolales: ubiquitous microbial community members of plant and food biospheres. Mycologia. 2018;110(1): 79–92.
6. Wang QM, Groenewald M, Takashima M, Theelen B, Han PJ, Liu XZ, et al. Phylogeny of yeasts and related filamentous fungi within *Pucciniomycotina* determined from multigene sequence analyses. Stud Mycol. 2015;81:27–53.
7. Han M, Du C, Xu Z, Qian H, Zhang W. Rheological properties of phosphorylated exopolysaccharide produced by *Sporidiobolus pararoseus* JD-2. Int J Biol Macromol. 2016;88:603–13.
8. Manowattana A, Techapun C, Watanabe M, Chaiyaso T. Bioconversion of biodiesel-derived crude glycerol into lipids and carotenoids by an oleaginous red yeast *Sporidiobolus pararoseus* KM281507 in an airlift bioreactor. J Biosci Bioeng. 2018;125(1):59–66.
9. Chaiyaso T, Manowattana A. Enhancement of carotenoids and lipids production by oleaginous red yeast *Sporidiobolus pararoseus* KM281507. Prep Biochem Biotechnol. 2018;48(1):13–23.
10. Li C, Zhang N, Li B, Xu Q, Song J, Wei N, et al. Increased torulene accumulation in red yeast *Sporidiobolus pararoseus* NGR as stress response to high salt conditions. Food Chem. 2017;237:1041–7.
11. Li C, Li B, Zhang N, Wei N, Wang Q, Wang W, et al. Salt stress increases carotenoid production of *Sporidiobolus pararoseus* NGR via torulene biosynthetic pathway. J Gen Appl Microbiol. 2019;65(3):111–20.
12. Du C, Guo Y, Cheng Y, Han M, Zhang W, Qian H. Torulene and torularhodin, protects human prostate stromal cells from hydrogen peroxide-induced oxidative stress damage through the regulation of Bcl-2/Bax mediated apoptosis. Free Radic Res. 2017;51(2):113–23.
13. Keceli TM, Erginkaya Z, Turkkan E, Kaya U. Antioxidant and antibacterial Eeffects of carotenoids extracted from *Rhodotorula glutinis* strains. Asian J Chem. 2013;25(1):42–6.
14. Latha BV, Jeevaratanm K. Thirteen-week oral toxicity study of carotenoid pigment from *Rhodotorula glutinis* DFR-PDY in rats. Indian J Exp Biol. 2012; 50(9):645–51.
15. Kot AM, Błażejak S, Gientka I, Kieliszek M, Bryś J. Torulene and torularhodin: "new" fungal carotenoids for industry? Microb Cell Factories. 2018;17(1):49.
16. Han M, Xu Z, Du C, Qian H, Zhang W. Effects of nitrogen on the lipid and carotenoid accumulation of oleaginous yeast *Sporidiobolus pararoseus*. Bioprocess Biosyst Eng. 2016;39(9):1425–33.
17. Li Q, Du W, Liu D. Perspectives of microbial oils for biodiesel production. Appl Microbiol Biot. 2008;80(5):749–56.
18. Vicente G, Bautista LF, Rodríguez R, Gutiérrez FJ, Sádaba I, Ruiz-Vázquez RM, et al. Biodiesel production from biomass of an oleaginous fungus. Biochem Eng J. 2009;48(1):22–7.
19. Subramaniam R, Dufreche S, Zappi M, Bajpai R. Microbial lipids from renewable resources: production and characterization. J Ind Microbiol Biot. 2010;37(12):1271–87.
20. Gujjala LKS, Kumar SPJ, Talukdar B, Dash A, Kumar S, Sherpa KC, et al. Biodiesel from oleaginous microbes: opportunities and challenges. Biofuels. 2019;10(1):45–59.
21. Jiang W, Lv Y, Cheng L, Yang K, Chao B, Wang X, et al. Whole genome sequencing of the giant devil catfish, *Bagarius yarrelli*. Genome Biol Evol. 2019;11(8):2071–7.
22. Li C, Zhang N, Song J, Wei N, Li B, Zou H, et al. A single desaturase gene from red yeast *Sporidiobolus pararoseus* is responsible for both four- and five-step dehydrogenation of phytoene. Gene. 2016;590(1):169–76.
23. Li C, Li B, Zhang N, Wang Q, Wang W, Zou H. Comparative transcriptome analysis revealed the improved β-carotene production in *Sporidiobolus pararoseus* yellow mutant MuY9. J Gen Appl Microbiol. 2019;65(3):121–8.
24. Mannazzu I, Landolfo S, Da Silva TL, Buzzini P. Red yeasts and carotenoid production: outlining a future for non-conventional yeasts of biotechnological interest. World J Microbiol Biotechnol. 2015;31(11):1665–73.
25. Wang QM, Yurkov AM, Göker M, Lumbsch HT, Leavitt SD, Groenewald M, et al. Phylogenetic classification of yeasts and related taxa within *Pucciniomycotina*. Stud Mycol. 2015;81:149–89.
26. Fonseca A, Sampaio JP. *Rhodosporidium lusitaniae* sp. nov., a novel homothallic Basidiomycetous yeast species from Portugal that degrades phenolic compounds. Syst Appl Microbiol. 1992;15(1):47–51.
27. Sampaio JP. Chapter 130 - *Sporidiobolus* Nyland (1949). In: Kurtzman CP, Fell JW, Boekhout T, editors. The yeasts. 5th ed. London: Elsevier; 2011. p. 1549–61.
28. Sampaio JP. Chapter 127 - *Rhodosporidium* Banno (1967). In: Kurtzman CP, Fell JW, Boekhout T, editors. The yeasts. 5th ed. London: Elsevier; 2011. p. 1523–39.
29. Stolze-Rybczynski JL, Cui Y, Stevens MH, Davis DJ, Fischer MW, Money NP. Adaptation of the spore discharge mechanism in the Basidiomycota. PLoS One. 2009;4(1):e4163.
30. Turner JCR, Webster J. Mass and momentum transfer on the small scale: how do mushrooms shed their spores? Chem Eng Sci. 1991;46(4):1145–9.
31. Ianiri G, Abhyankar R, Kihara A, Idnurm A. *Phs1* and the synthesis of very long chain fatty acids are required for ballistospore formation. PLoS One. 2014;9(8):e105147.
32. Han M, Xu J, Liu Z, Qian H, Zhang W. Co-production of microbial oil and exopolysaccharide by the oleaginous yeast *Sporidiobolus pararoseus* grown in fed-batch culture. RSC Adv. 2018;8(6):3348–56.
33. Coelho MA, Almeida JMGC, Hittinger CT, Gonçalves P. Draft genome sequence of *Sporidiobolus salmonicolor* CBS 6832, a red-pigmented basidiomycetous yeast. Genome announc. 2015;3(3):e44415.
34. Fischer MWF, Stolze-Rybczynski JL, Cui Y, Money NP. How far and how fast can mushroom spores fly? Physical limits on ballistospore size and discharge distance in the Basidiomycota. Fungal Biol. 2010;114(8):669–75.
35. Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998;46(4):409–18.
36. Yi S, Wang S, Zhong J, Wang W. Comprehensive transcriptome analysis provides evidence of local thermal adaptation in three loaches (genus: *Misgurnus*). Int J Mol Sci. 2016;17(12):1943.
37. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
38. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117–23.
39. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30(1):31–7.
40. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.
41. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24(5):713–4.
42. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience. 2012;1:18.
43. Pevzner PA, Tang H, Waterman MS. An eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98(17):9748–53.
44. Haddad NJ, Loucif-Ayad W, Adjlane N, Saini D, Manchiganti R, Krishnamurthy V, et al. Draft genome sequence of the Algerian bee *Apis mellifera intermissa*. Genom Data. 2015;4:24–5.
45. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32:W309–12.
46. Li SF, Guo YJ, Li JR, Zhang DX, Wang BX, Li N, et al. The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). Mob DNA. 2019;10:3.

47. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
48. Lagesen K, Hallin P, Rødland EA, Stærfeldt H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100–8.
49. Lowe TM, Chan PP. tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016;44(W1):W54–7.
50. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-coding RNA analysis using the Rfam database. Curr Protoc Bioinformatics. 2018;62(1):e51.
51. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46(D1):D335–42.
52. Li S, Tang Y, Fang X, Qiao T, Han S, Zhu T. Whole-genome sequence of *Arthrinium phaeospermum*, a globally distributed pathogenic fungus. Genomics. 2019;112(1):919–29.
53. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36(10):3420–35.
54. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
55. Li L, Stoeckert JCJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
56. Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature. 2001;411(6841):1046–9.
57. Dennis GJ, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4(5):3.
58. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
59. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
60. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
61. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11(9):1650–67.
62. Pertea M, Pertea GM, Antonescu CM, Chang T, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
63. Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy Moonshine A, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2018;43(1110):11. 10.1–33.
64. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164.
65. FAS O, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
66. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000; 17(1):32–43.
67. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 2006;4(4):259–63.
68. Wang Y, Yang L, Zhou K, Zhang Y, Song Z, He S. Evidence for adaptation to the Tibetan plateau inferred from tibetan loach transcriptomes. Genome Biol Evol. 2015;7(11):2970–82.
69. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
70. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.
71. Russo CADM, Selvatti AP. Bootstrap and rogue identification tests for phylogenetic analyses. Mol Biol Evol. 2018;35(9):2327–33.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.