**Supplementary Information**

Lipid and carotenoid production by the *Rhodosporidium toruloides* mutant in cane molasses

Xuehai Zheng1, Ronghui Hu1, Duo Chen, Jiannan Chen, Wenjin He, Luqiang Huang, Chentao Lin, Huibin Chen, Youqiang Chen, Jinmao Zhu, Feng Qi, Ting Xue\*

The Public Service Platform for Industrialization Development Technology of Marine Biological Medicine and Products of the State Oceanic Administration, Fujian Key Laboratory of Special Marine Bioresource Sustainable Utilization, Southern Institute of Oceanography, College of Life Sciences, Fujian Normal University, Fuzhou, China

\*Correspondence: College of Life Sciences, Fujian Normal University, Fuzhou 350117, Fujian, China.

1. mail address:xueting@fjnu.edu.cn (T. Xue).

1Xuehai Zheng, Ronghui Hu and Duo Chen contributed equally to this work.

1. **Methods**

**1.1. Extraction and analytical methods for lipids and carotenoids**

*R. toruloides* cells were collected and washed twice to remove the culture medium. Cell-free extracts were prepared with a grinding mill at 65 Hz for 3 min, and 1 mL acetone was subsequently added to dissolve the pigment. The supernatant was obtained by filtration for HPLC analyses. 3 mL of chloroform and methanol (2:1, v/v) was put into the cell-free extracts to dissolve the lipids. The supernatant was collected and transferred to preweighed vials and oven-dried at 65°C to attain a constant weight. After dissolving the lipid into 1 mL of toluene solution containing 200 µL of 1 mg/mL butylated hydroxytoluene (BHT), 2 mL of HCl in methanol was added, and incubated overnight at 50°C. The mixture was processed using 1 mL *n*-hexane containing 1 mg/mL methyl nonadecanate. The supernatant was obtained by centrifugation and filtration, and further utilized for GC analyses. The extraction process for pigments and lipids was repeated two times.

Stock solutions of β-carotene, torulene and torularhodin were treated by dissovlving 1 mg β-carotene, torulene and torularhodin in 10 mL methanol solution. Stock solutions of torulene and torularhodin standards were gradient diluted to concentrations varying from 10 µg/L to 100 µg/L, and a stock solution of β-carotene standard was gradient diluted to concentrations varying from 20 µg/L-200 µg/L, and the standard curves of torulene (Y = 1.32e+005X-8.12e+005, R2 = 0.993), torularhodin (Y = 6.09e+004X+3.31e+005, R2 = 0.993) and β-carotene (Y = 2.07e+003X+ 1.23e+004, R2 = 0.994), respectively, were obtained. Torulene, torularhodin and β-carotene were all detected and quantified using HPLC with a Sun Fire C18 column and an FDA refractive index detector at 28°C (Waters Corp., USA). The mobile phases were acetonitrile:methanol (v/v, 15:85; solvent A) and acetonitrile:methanol:ethyl acetate (v/v/v, 20:60:20; solvent B). The lipid content was measured by using GC analysis at a gradient flow rate of 1 mL/min with 100% A at 0 min, 100% B at 8 min, 100% B at 25 min, and 100% A at 40 min. Cell dry weight (CDW) was measured by the gravimetric method (Ayuwaningsih et al., 2018).

**1.2. Genome sequencing, assembly and annotation**

The full PacBio subreads (9.66 Gb) were assembled by the CANU pipeline using default parameters (Koren et al., 2017). Primary contigs were polished using 3.24 Gb Illumina clean data by the BWA and Pilon programs (Walker et al., 2014). BUSCO by eukaryotic models were used to assess the completeness of the mutant *R. toruloides* M18 genome (Seppey et al., 2019). Tandem Repeats Finder, LTR\_FINDER and RepeatMasker were utilized to identify the repeat sequences in the *R. toruloides* M18 genome, as previously described (Benson, 1999; Tarail-Graovac and Chen, 2009; Xu, 2007). Next, we do annotation of the *R. toruloides* M18 genome assembly with homology-based, *ab initio* annotation and transcriptome-based. The MAKER pipeline was used to predict coding genes (version 2.31.9) (None, 2011). Functional annotation was performed in the databases (eggNOG, GO, COG and KEGG) by BLASTP with an E-value ≤ 1e-5 (Ashburner et al.; Kanehisa et al., 2014). Small RNAs and ncRNA were predicted using alignment to the miRNA and Rfam databases using tRNAsan-SE and BLASTN, respectively (Gardner et al., 2009; Lowe and Chan, 2016).

1. **Results and discussion**
   1. **Genome sequencing, assembly and annotation for *R. toruloides* mutant M18**

The *de novo* genome assembly demonstrated that *R. toruloides* mutant M18 genome size was 21.31 Mb (contig N50:1.42 Mb), a GC content of 61.79% and the longest contig length of 1.84 Mb. Approximately 86.4% of the plant orthologs were included, and 93.7% of the Illumina clean reads can mapped to the genome. Sequence annotation predicted 7,539 protein-coding genes in *R. toruloides* mutant M18. We functionally annotated 5,453 genes and 170 ncRNAs in *R. toruloides* mutant M18 genome, including 125 tRNAs, 36 rRNAs and 9 snRNAs. Repetitive sequences of the *R. toruloides* mutant M18 was 0.75% of the genome, and LTR retrotransposons was 0.14%.

* 1. **Evolution and phylogeny of the *R. toruloides* mutant M18 genome**

We constructed a phylogenetic tree of *R. toruloides* mutant M18 and 15 other species based on 162 single-copy families. The phylogenetic tree showed that *R. toruloides* mutant M18 belongs to different species in the *Rhodosporidium* genus. In total, 1,1613 families of homologous genes were identified among 16 plant species by CAFE (version 2.1). Evolution analysis demonstrated that 233 gene families were expanded, whereas 1,062 gene families were contracted in the *R. toruloides* mutant M18genome. KEGG analysis resulted that most expanded genes were enriched in signal transduction, lipid metabolism (GPD1), energy metabolism and development. GO analysis indicated that most expanded genes were related to growth, metabolic processes, cellular process, localization, stimulus responses, reproduction and reproduction. For the 1,062 contracted gene families, we conducted GO enrichment for the GO terms primarily involved in the cell, organelles, catalytic activity, binding, signaling and transporter activity. KEGG enrichment of the contracted genes suggested that they were enriched in carbohydrate metabolism, amino acid metabolism, energy metabolism. By comparing *R. toruloides* mutant M18 with 11 other species, we observed that 510 (39.38%) of 1,295 *R. toruloides* mutant M18 gene families were shared by these 12 species, while 14.74% (191) of the gene families were unique to *R. toruloides* mutant M18. Functional analysis performed through a KEGG analysis demonstrated that 276 genes in these 191 unique families were enriched in pyruvate metabolism (PDH) and glycerophospholipid metabolism (GPD1). It indicated that the cytochrome P450 gene family regulate the carotenoid biosynthesis, including CYP71, CYP97A, CYP97B, and CYP97C in previous research (Cui et al., 2019; Kim & Dellapenna, 2006; Tamaki et al., 2018). Additionally, we found that CYP97A3 was significantly expanded and specific to the mutant M18 genome. It is surmised that the these expanded gene families may play an important affection on the secondary metabolites pathways, resulting in a higher carotenoid content of mutant M18 compared with other strains.

**2.3. qRT-PCR validation**

Four target genes (RtMYB111, RtbHLH4, Δ-4 desaturase, Δ-12 desaturase) expression levels were analyzed using qRT-PCR. The relative levels of the amplified mRNAs were assessed by using the 2−ΔΔCt method. In comparation with FPKM values, the levels of all four genes relative expression evaluated by qRT-PCR were in agreement with the RNA-seq results. Cane molasses, as a substitute for glucose in based media, is appropriate for achieving low-cost lipids and carotenoids production in *R. toruloides*, and this study may represent a valuable reference for future efforts attempting to obtain cost-efficient biodiesel production in industrial application.

**References**

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., et al., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25-29.

Ayuwaningsih, I.P., Fachrul, M.F., Rinanti, A., Abdullah, A.G., Nandiyanto, A.B., 2018. Increasing lipid content from biomass of microalgae to produce biofuels with optimization of nitrogen source. [MATEC Web of Conferences](https://www.researchgate.net/journal/2261-236X_MATEC_Web_of_Conferences) 197(5), 13011.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2), 573-80.

Cui, H., Ma, H., Cui, Y., Zhu, X., Li, R.J., 2019. Cloning, identification and functional characterization of two cytochrome P450 carotenoids hydroxylases from the diatom *Phaeodactylum tricornutum*. J. Biosci. Bioeng. 128(6), 755-765.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., 2009. Rfam: updates to the RNA families database. Nucleic Acids Res. 37(Database issue), 136-140.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42 ,D199-205.

Kim, J., Dellapenna, D., 2006. Defining the primary route for lutein synthesis in plants: The role of Arabidopsis carotenoid β-ring hydroxylase CYP97A3. Proc. Natl. Acad. Sci. USA.103(9), 3474-3479.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5),722-736.

Lowe, T.M., Chan, P., 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 44 (W1), W54-W57.

None., 2011. Molecular biology: Protein maker and gene regulator. Nature 473(7346), 127.

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. 1962, 227-245.

Tamaki, S., Kato, S., Shinomura, T., Ishikawa, T., Imaishi, H., Physiological role of β-carotene monohydroxylase (CYP97H1) in carotenoid biosynthesis in *Euglena gracilis*. Plant Sci. 278,80-87.

Tarailo‐Graovac, M., Chen, N., 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinformatics 25(1), 1-14.

Xu, Z., Wang, H., 2007. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35(Web Server issue), W265-268.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11), e112963.