A peer-reviewed version of this preprint was published in PeerJ on 14 November 2017.

View the peer-reviewed version (peerj.com/articles/4030), which is the preferred citable publication unless you specifically need to cite this preprint.

Gan HM, Thomas BN, Cavanaugh NT, Morales GH, Mayers AN, Savka MA, Hudson AO. 2017. Whole genome sequencing of *Rhodotorula mucilaginosa* isolated from the chewing stick (*Distemonanthus benthamianus*): insights into *Rhodotorula* phylogeny, mitogenome dynamics and carotenoid biosynthesis. PeerJ 5:e4030 https://doi.org/10.7717/peerj.4030

Whole genome sequencing of *Rhodotorula mucilaginosa* isolated from the chewing stick (*Distemonanthus benthamianus*): insights into *Rhodotorula* phylogeny, mitogenome dynamics and carotenoid biosynthesis

Han Ming Gan [Corresp., 1, 2, 3], **Bolaji N Thomas** [4], **Nicole T Cavanaugh** [5], **Grace H Morales** [5], **Ashley N Mayers** [4], **Michael A Savka** [5], **Andre O Hudson** [Corresp. 5]

[1] Centre for Integrative Ecology-School of Life and Environmental Sciences, Deakin University, Victoria, Australia

[2] Genomics Facility, Monash University, Selangor, Malaysia

[3] School of Science, Monash University, Selangor, Malaysia

[4] College of Health Science and Technology, Rochester Institute of Technology, Rochester, New York, United States

[5] Thomas H. Gosnell School of School of Life Sciences, Rochester Institute of Technology, Rochester, New York, USA

Corresponding Authors: Han Ming Gan, Andre O Hudson
Email address: han.gan@deakin.edu.au, aohsbi@rit.edu

In industry, the yeast *Rhodotorula mucilaginosa* is commonly used for the production of carotenoids. The production of carotenoids is important because they are used as natural colorants in food and some carotenoids are precursors of retinol (vitamin A). However, the identification and molecular characterization of the carotenoid pathway/s in species belonging to the genus *Rhodotorula* is scarce due to the lack of genomic information thus potentially impeding effective metabolic engineering of these yeast strains for improved carotenoid production. In this study, we report the isolation, identification, characterization and the whole nuclear genome and mitogenome sequence of the endophyte *R. mucilaginosa* RIT389 isolated from *Distemonanthus benthamianus,* a plant known for its anti-fungal and antibacterial properties and commonly used as chewing sticks. The assembled genome of *R. mucilaginosa* RIT389 is 19 Mbp in length with an estimated genomic heterozygosity of 9.29%. Whole genome phylogeny supports the species designation of strain RIT389 within the genus in addition to validating the monophyly of the genus *Rhodotorula*. Further, we report for the first time, the recovery of the complete mitochondrial genome of *R. mucilaginosa* using the genome skimming approach. The assembled mitogenome is at least 7,000 bases larger than that of *Rhodotorula taiwanensis* which is largely attributed to the presence of large intronic regions containing open reading frames coding for homing endonuclease from the LAGLIDADG and GIY-YIG families. Furthermore, genomic regions containing the key genes for carotenoid production were identified in *R. mucilaginosa* RIT389, revealing differences in gene synteny that may play a role in the regulation of the biotechnologically important carotenoid synthesis pathways in

yeasts.

Running Head: *Rhodotorula mucilaginosa* RIT389 isolated from *Distemonanthus benthamianus*

Whole genome sequencing of *Rhodotorula mucilaginosa* isolated from the chewing stick (*Distemonanthus benthamianus*): insights into *Rhodotorula* phylogeny, mitogenome dynamics and carotenoid biosynthesis.

Han Ming Gan[1,2,3]*, Bolaji N. Thomas[4], Nicole T. Cavanaugh[5], Grace H. Morales[5], Ashley N. Mayers[4], Michael A. Savka[5] and André O. Hudson[5]*

[1] Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia
[2]Genomics Facility, Tropical and Medicine Biology Platform, Monash University Malaysia, Bandar Sunway, Selangor, Malaysia
[3] School of Science, Monash University Malaysia, Bandar Sunway, Selangor, Malaysia
[4] The College of Health Science and Technology, Rochester Institute of Technology, Rochester, NY, USA.
[5] Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, USA

*** Corresponding authors:**

Han Ming Gan, Ph.D.
Centre for Integrative Ecology,
School of Life and Environmental Sciences,
Deakin University
75 Pigdons Road, Waurn Ponds,
3216 VIC Australia
E-mail: han.gan@deakin.edu.au

André O. Hudson, Ph.D.
Thomas H. Gosnell School of Life Sciences
Rochester Institute of Technology
85 Lomb Memorial Dr.
Rochester, NY 14623 U.S.A.
Telephone: 585-475-4259
FAX: 585-475-5766
E-mail: aohsbi@rit.edu

Abstract

In industry, the yeast *Rhodotorula mucilaginosa* is commonly used for the production of carotenoids. The production of carotenoids is important because they are used as natural colorants in food and some carotenoids are precursors of retinol (vitamin A). However, the identification and molecular characterization of the carotenoid pathway/s in species belonging to the genus *Rhodotorula* is scarce due to the lack of genomic information thus potentially impeding effective metabolic engineering of these yeast strains for improved carotenoid production. In this study, we report the isolation, identification, characterization and the whole nuclear genome and mitogenome sequence of the endophyte *R. mucilaginosa* RIT389 isolated from *Distemonanthus benthamianus,* a plant known for its anti-fungal and antibacterial properties and commonly used as chewing sticks. The assembled genome of *R. mucilaginosa* RIT389 is 19 Mbp in length with an estimated genomic heterozygosity of 9.29%. Whole genome phylogeny supports the species designation of strain RIT389 within the genus in addition to validating the monophyly of the genus *Rhodotorula*. Further, we report for the first time, the recovery of the complete mitochondrial genome of *R. mucilaginosa* using the genome skimming approach. The assembled mitogenome is at least 7,000 bases larger than that of *Rhodotorula taiwanensis* which is largely attributed to the presence of large intronic regions containing open reading frames coding for homing endonuclease from the LAGLIDADG and GIY-YIG families. Furthermore, genomic regions containing the key genes for carotenoid production were identified in *R. mucilaginosa* RIT389, revealing differences in gene synteny that may play a role in the regulation of the biotechnologically important carotenoid synthesis pathways in yeasts.

Keywords

Rhodotorula mucilaginosa, Distemonanthus benthamianus, chew sticks

mitogenome, phylogenomics, carotenoid, next-generation sequencing, endophyte

Introduction

Rhodotorula mucilaginosa is a common saprophytic fungus that is a part of the Basidiomycota phylum. The organism is typically found in soils, lakes, ocean water, milk and fruit juice (Wirth & Goldani 2012). Of the numerous species in the genus *Rhodotorula,* only *Rhodotorula mucilaginosa*, *Rhodotorula glutinis*, and *Rhodotorula minuta* have been known to be pathogenic to humans (Wirth & Goldani 2012; Zaas et al. 2003). Despite being categorized as an opportunistic and emerging pathogen, *R. mucilaginosa* from natural environments appear to possess interesting biological traits ranging from indole acetic acid production (plant growth-promoting), bacterial quorum sensing signal degradation (quorum quenching) to carotenoid production (Ghani et al. 2014; Ignatova et al. 2015; Libkind et al. 2004). Despite its genomic potential, genomic resource for *R. mucilaginosa* is surprisingly scarce in public database. To date, the only genomic resource available for this species is from *R. mucilaginosa* strain C2.5t1 that was isolated from the seeds of the cacao plant in Cameroon (Deligios et al. 2015).

Carotenoid production in fungi has been suggested as a natural mechanism to protect against photo-oxidative damage in light-intensive environments, given the known antioxidant property of these lipid-soluble pigments as attributed to their chemical structure (Avalos & Carmen Limon 2015; Cerdá-Olmedo 1989; Echavarri-Erasun & Johnson 2002). The biosynthetic pathway of beta-carotene from phytoene has been

91 elucidated in at least three different fungal species based on cDNA cloning and
92 enzymatic characterization and was shown to require two major proteins namely, a
93 dehydrogenase and a bifunctional enzyme, encoding both cyclase and phytoene synthase
94 activities (Sanz et al. 2011; Verdoes et al. 2003). However, molecular characterization of
95 this biosynthesis pathway is lacking in the genus *Rhodotorula* despite their demonstrated
96 potential in industrial-scale carotenoid production (Cutzu et al. 2013; Davoli et al. 2004;
97 Libkind et al. 2004; Marova et al. 2012; Taccari et al. 2012). A recent study has shown
98 that the heterologous expression of a 3-hydroxy-3-methylglutaryl coenzyme A reductase
99 from *Saccharomyces cerevisiae* substantially increased carotenoid production in *R.*
100 *mucilaginosa* strain KC8 (Wang et al. 2017), indicating the potential of metabolic
101 engineering as alternative and/or complementary approach to growth condition
102 optimization (Cutzu et al. 2013; Davoli et al. 2004; Marova et al. 2012) for improving
103 carotenoid production in *Rhodotorula* species.

104 The plant *Distemonanthus benthamianus* is a semi-deciduous perennial tree
105 commonly found in second-growth forests in Nigeria, Cameroon and Ghana (Adeniyi et
106 al., 2011). *D. benthamianus* is of interest given that the plant is used as chewing sticks for
107 dental and oral hygiene by members of Yoruba community in Nigeria. A relatively recent
108 study showed that extracts from the bark of the stems exhibit bactericidal activity against
109 *Staphylococcus aureus* and *Streptococcus mutans*, two bacteria that are often associated
110 with skin and dental infections, respectively (Adeniyi & Odumosu 2012).

111 In this study, an initial screen for endophytic bacteria that are resistant to the
112 extracts of *D. benthamianus* led to the isolation of a pink-pigmented strain subsequently
113 identified as a fungal strain belonging to the species *Rhodotorula mucilaginosa*. Given

the intriguing property of this fungal species and its lack of genomic resources, we sequenced its whole genome on the Illumina platform and performed comparative genomic analysis to gain insight into the carotenoid biosynthetic pathway of this species and more generally the genus *Rhodotorula*. Notably, we also recovered the complete mitochondrial genome of *R. mucilaginosa,* the first for its species and the second for its genus, using genome skimming approach.

Materials and Methods

Strain Isolation

Two grams of internal tissue obtained from surfaced sterilized stem of *Distemonanthus benthamianus* plant was used to inoculate 25mL of half strength tryptic soy broth (TSB) medium and grown overnight at 30°C. Microorganisms were isolated by plating 100 µL of 10 fold serial dilutions from 10^{-5}-10^{-10} of the overnight culture on half strength tryptic soy agar.

Scanning Electron Microscopy (SEM)

To fix the organism, 100 µL of cells from 10^{-7} dilution from an overnight grown culture was suspended in 3% glutaraldehyde in 0.1M phosphate buffer pH 7.2 for 30 minutes. Following fixing, the cells were washed 3 times and pelleted in sterile water followed by a secondary fixation in 2% osmium tetroxide (in H_20) for 30 minutes. The cells were washed 3 more times in sterile water followed by dehydration of the cells in 25%, 50%, 75%, 95% and 100% ethanol for 5 minutes in each ethanol concentration. The cells were filtered through a 0.22 micron polyethersulfone membrane and incubated at room temperature for 1 hour followed by SEM stub mounting and sputter-coating using 10 nm gold/palladium.

Whole genome sequencing

Total DNA was extracted from a 3-day-old half strength tryptic soy agar culture of *R. mucilaginosa* RIT389 using the MolBio DNA extraction kit according to the manufacturer's instructions. The gDNA was sheared to 500 bp fragment using the Covaris ultrasonicator and subsequently prepared for whole genome sequencing using NEBNext Ultra™ DNA Library Prep kit for Illumina (New England BioLabs, Ipswich, MA). The generated library was subsequently quantified using Qubit and sequenced on the MiSeq (Illumina, San Diego, CA) located at the Monash University Malaysia Genomics Facility using the run configuration of 2×250bp.

Genome assembly and annotation

Genome size, heterozygosity rate and repeat content were initially estimated using GenomeScope (Vurture et al. 2017). Based on the observed high genome heterozygosity of strain RIT389, dipSPAdes version 3.10.1 was used to assemble the whole genome with the additional option of "-expect-rearrangements" activated (Bankevich et al. 2012). Genome completeness was calculated using BUSCO3 based on the Basidiomycota odb9 ortholog dataset (Simao et al. 2015). Then, gene prediction was performed using GeneMark-ES fungal version (Borodovsky & Lomsadze 2011) with the enhanced intron submodel that can better accommodate sequences with and without branch point sites in the fungal genomes.

Complete mitogenome was recovered by randomly sub-sampling 1/10 of the pair-end reads and assembling them using SPAdes version 3.10.1 (Bankevich et al. 2012). The contig corresponding to the whole mitogenome was re-circularized manually, as previously described and annotated automatically using MFannot

(http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl). Additional genes coding for homing endonucleases commonly found in fungal mitogenomes were identified based on the presence of protein domains corresponding to the GIY-YIG catalytic domain (PF01541.23) and LAGLIDADG endonuclease (PF00961.18, PF03161.12 and PF14528.5) using hmmsearch3 with an E-value cutoff of 1e-5 (Eddy 2011).

Phylogenomics and Comparative Genomics

Pair-wise average nucleotide identity (ANIm) was calculated using JSpecies (Richter et al. 2016) and subsequently visualized with the library package pheatmap in Rstudio. Single-copy genes present in all selected fungal genomes were identified using BUSCO3 (Simao et al. 2015). The protein sequences for each ortholog were aligned and trimmed using Muscle and trimAl (-automated1), respectively (Capella-Gutierrez et al. 2009; Edgar 2004). The final trimmed alignments were concatenated and used to construct a maximum likelihood tree using FastTreeMP (Price et al. 2010). The reconstructed tree was visualized and annotated using TreeGraph2 (Stöver & Müller 2010).

Identification of proteins involved in the carotenoid biosynthesis pathway was done by scanning the whole predicted proteome for protein domain hits (NC cutoff for TIGRfam and 1e-5 cutoff for Pfam) to lycopene cyclase (TIGR03462, CrtY), phytoene desaturase/dehydrogenase (TIGR02734, CrtI), squalene/phytoene synthase (PF0494, CrtB) and isopentenyl-diphosphate delta-isomerase (TIGR02150). Visualization and comparison of gene neighborhoods were performed using EasyFig with the default

BlastN setting (Sullivan et al. 2011). Proteins coded in each genomic sub-region were functionally annotated using Interproscan5 (Jones et al. 2014).

Results and Discussion

We noticed an organism that was pink/red in color from the screen on tryptic soy agar (Figure 1A). Based on our previous studies of isolating endophytic organisms, we initially thought the organism belonged to genus *Serratia* or a related genus based on the color of the colonies. However, based on SEM analysis, it was initially determined that the organism was eukaryotic and not a bacterium based on the size and the morphology depicting cell division (Figure 1B). The identification of the organism was subsequently confirmed using nucleotide sequencing.

GenomeScope estimated a genome size of 18.6 mega base pairs (Mbp) with an estimated heterozygosity of 9.29% for strain RIT389 (Figure 2). The predicted genome size is fairly close to the *de novo* assembled genome length of 19.6 Mbp contained in 250 contigs. The assembled genome has a GC content of 60.28% with an estimated completeness of 89.70%. In addition, *de novo* assembly using sub-sampled reads enabled the recovery of the complete mitogenome of strain RIT389. Approximately 4.55% of the total pair-end reads mapped to the complete mitogenome with an estimated coverage of 400× (Table 1). The complete mitogenome length is 47,023 bp with a GC content of 40.43% which is substantially lower than that of the nuclear genome.

205 **Genomic and genetic approaches support the species identification of strain RIT389**
206 **as *Rhodotorula mucilaginosa***

207 The ITS region of strain RIT389 exhibits a 100% identity with the sequences of
208 various *Rhodotorula mucilaginosa* strains including the type strain *R. mucilaginosa*
209 ATCC 201848 (data not shown). At the whole genome level, it exhibits the highest
210 average nucleotide identity of 94.10% to strain C2.5t1, the only other genome-sequenced
211 strain of this species at the time of this study (Deligios et al. 2015) (Figure 3). Similar to
212 strain RIT389, strain C2.5t1 is also a plant-associated and was isolated from a cacao
213 seeds (*Theobroma cacao* L) in Cameroon and shown to produce high carotenoid when
214 grown in medium supplemented with glycerol (Cutzu et al. 2013). Although the plant
215 growth-promoting activity of both strains RIT389 and C2.5t1 has not been studied, a
216 third strain of *R. mucilaginosa*, YR07 isolated from legume plant rhizophere, has been
217 shown to synthesize up to 45.3 μg of indole acetic acid (IAA) per mL of culture medium
218 (Ignatova et al. 2015). In addition, strain YR07 also exhibits antifungal activity as
219 evidenced by the formation of large inhibition zone against *Fusarium graminearum*, a
220 phytopathogenic fungi.
221
222 **Whole genome phylogeny supports the monophyly of the genus *Rhodotorula***
223 A total of 789 single-copy genes universally present in all 19 fungal strains
224 (Figure 4) were used to generate a concatenated amino acid alignment consisting of
225 539,792 sites (234,641 informative sites). By rooting *Microbotyrum violaceum* and
226 *Microbotyrum saponariae* belonging to a different order e.g. Microbotryales, as the
227 outgroup, members of the genus *Rhodotorula* formed a monophyletic group cluster with

maximal SH-like support with *R. graminis* WP1 being basal to the rest of the *Rhodotorula* strains (Figure 4). The lack of strong SH-like support at the shallow relationship especially for members of the species *R. toluides* is most likely due to the lack of genomic differences which is expected given that some of the strain names are the alternative strain name of the identical type strain. For example, the type strain designations ATCC10788, IFO0559 and JCM10020 for *R. toluides* were all derived from the original strain CBS 14. Interestingly, phylogenomic analysis indicates that the currently sequenced strains of *Rhodotorula toluides* consist of two major clades with a Jspecies-calculated intraclade and interclade average pair-wise ANI difference of 0.4% and 13%, respectively (Figure 3). *R.* sp. JG1b together with *R. mucilaginosa* strains RIT389 and C2.5t1 formed a monophyletic group that is sister taxa to the major *R. toluides* group. The close affinity of *R.* sp. JG-1b to *R. mucilaginosa* is interesting as it is an eurypsychrophilic yeast isolated from ~150,000-year-old ice-cemented permafrost soils (Goordial et al. 2016a). Given the close genomic affinity of *R.* sp. JG-1b to the currently sequenced *R. mucilaginosa* strains and their diverse isolation source, comparative genomics of these strains may assist in the future identification of novel cold adaptive traits at the molecular level in the genus *Rhodotorula* (Goordial et al. 2016b).

Homing endonuclease-mediated mitogenome expansion in *Rhodotorula mucilaginosa* RIT389

The mitogenome of *R. mucilaginosa* RIT389 is 93% similar to that of *R. taiwanensis* (Accession Number: HF558455), the only other publicly available complete *Rhodotorula* mitogenome (Zhao et al. 2013). The lack of complete and annotated

mitogenome sequences from *R. mucilaginosa* is surprising given the availability of various *R. mucilaginosa* whole genome sequences. Given the high abundance of mitochondrial organelle in an actively dividing cell, the depth of mitochondrial-derived sequencing reads will be substantially higher than that of the nuclear genome. While this genomic feature has been used to rapidly recover complete mitogenome via the genome skimming approach (Gan et al. 2014; Grandjean et al. 2017), overly high read depth of the mitogenome obtained from a full genome sequencing data is likely to be detrimental to *de novo* assembly due to the accumulation of sequencing errors, leading to the generation of fragmented and/or incomplete mitogenome assembly (Mirebrahim et al. 2015). In this work, we demonstrate that a simple subsampling approach followed by *de novo* assembly could be an effective way to recover complete mitogenome from fungal genome sequencing data.

Despite exhibiting a similar mitochondrial gene arrangement and a relatively high nucleotide sequence similarity to *R. taiwanensis* RS1, the assembled complete mitogenome of strain RIT389 is at least 7,000 bp larger than that of *R. taiwanensis*. Gene neighborhood analysis indicates that a majority of the length difference was largely due to the presence of intronic regions containing homing endonuclease genes(Figure 5) which function in the splicing of RNA transcripts eventually leading to the generation of functional host proteins. (Salvo et al. 1998; Wolters et al. 2015). In strain RIT389, approximately 3 kilobases of the large ribosomal RNA gene consist of intronic regions coding for LAGLIDADG-type endonuclease. Another type of endonuclease e.g. GIY-YIG endonuclease could also be identified within the 1.5 kilobases intronic region of RIT389 *nad5* gene which is absent in that of *R. taiwanensis* RS1.

Identification genomic region associated with carotenoid biosynthesis

Essential genes required for the biosynthesis of carotenoid could be identified in strain RIT389 which is consistent with its red coloration (Figure 1A, Supplemental Table 1), a visual evidence for carotenoid production. The genes coding for phytoene synthase (*crtB*), lycopene cyclase (*crtY*) and phytoene desaturase (*crtI*) are located in relatively close proximity with one another while the gene coding for the enzyme geranyl pyrophosphate synthase which is crucial for the production of an early precursor for carotenoid is located on separate contig (Supplemental Table 1). As observed in several fungal species, the *crtB* and *crtY* genes are fused and thus code for a bifunctional protein containing both lycopene cyclase and phytoene synthase activities (Arrach et al. 2001; Sanz et al. 2011). Within the genus *Rhodotorula*, the gene coding for carotenoid oxygenase (*crtX*) responsible for the cleavage of carotenoid to retinal (Vitamin A) and *crtBY* are located in close proximity and are convergently transcribed except in the species *R. mucilaginosa*, whereby *crtX* and *crtBY* are divergently transcribed and are separated by a large gene coding for OPT family small oligopeptide transporter (Figure 6). In *Fusarium fujikuroi*, mutation in the *crtX* gene led to the overproduction of carotenoid (Prado-Cabrero et al. 2007), suggesting that carotenoid oxygenase is involved in the regulation of the carotenoid synthesis through a negative feedback mechanism. The notable difference in gene arrangement and transcription orientation involving *crtX* can therefore affect the regulation of carotenoid synthesis and accumulation in *R. mucilaginosa* (Noble & Andrianopoulos 2013). It is also worth noting that the *crtI* gene in strain RIT389 was predicted as two separate genes which is unexpected given that the gene region exhibits high nucleotide homology and coverage to its respective orthologs in

297 strains WP1 and Ct2.5. Whole transcriptome analysis of strain RIT389 will be necessary
298 to validate the observed spliced *crtI* gene in the future.
299

300 **Conclusion**

301 We demonstrate the feasibility of reconstructing the whole genome and complete
302 mitogenome of *Rhodotorula mucilaginosa* using only Illumina short reads. The whole
303 genome of *R. mucilaginosa* is the second to be reported to date for its species. Despite the
304 availability of various whole genome sequences of *Rhodotorula* in public database, the
305 complete and annotated mitogenome of *Rhodotorula mucilaginosa* strain RIT389 is the
306 first to be successfully reconstructed via genome skimming and annotated for its species.
307 In addition to confirming the monophyly of the genus *Rhodotorula*, we highlight the
308 considerable dissimilarity in the syntheny of carotenoid synthesis gene cluster among
309 *Rhodotorula* strains with potential implications in the regulation of carotenoid
310 production.
311

312 **Data Deposition**

313 This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank
314 under the accession NIUW00000000. The version described in this paper
315 is version NIUW01000000. Bioproject, Biosample and SRA accession numbers are
316 shown in Table 1.
317

318 **Competing interests**

319 The authors declare that they have no competing interests.

Funding

Funding for this work was provided by the following: College of Science (COS) and The Gosnell School of Life Sciences (GSoLS) at Rochester Institute of Technology (RIT), a Research Laboratory and Faculty Development Award from the College of Health Sciences and Technology (RIT) and the Monash University Malaysia Tropical Medicine and Biology Multidisciplinary Platform. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Acknowledgments

The authors thank Dr. Richard Hailstone from the Chester F. Carlson Center for Imaging Science (RIT) for assistance with SEM analysis.

Figure Legends

Figure 1. (A) Color/morphology of *Rhodotorula mucilaginosa* RIT389 grown on half-strength tryptic soy agar (B) Scanning electron microscopy of *Rhodotorula mucilaginosa* RIT389 at 52.8K magnification.

Figure 2. GenomeScope estimation of genome size, repeat content and heterozygosity (Kmer length = 21, Read length = 251 bp and Max kmer coverage = 1,000).

Figure 3. Pairwise average nucleotide identity calculation of *Rhodotorula* genomes. Genomes with the superscript “T” are type strains.

Figure 4. Maximum likelihood tree of a concatenated amino acid alignment consisting of 537,792 sites that represent 798 universally present single-copy genes from 19 fungal strains. Labels on branches indicate shimodaira-hasegawa (SH)-like local branch support values. The scale bar indicates the average number of amino acid substitutions per site.

Figure 5. Complete mitochondrial genome of *R. mucilaginosa* RIT389 compared against that of *R. taiwanensis* RS1. Orange frames indicate coding sequences commonly found in a typical mitochondrial genome. Red and Blue arrows indicate transfer and ribosomal RNAs, respectively. Arrow direction represents transcriptional orientation. Dotted lines indicate intronic regions.

Figure 6. Comparison of genomic sub-region containing the gene cluster associated with carotenoid biosynthetic pathway. Orange frames within the teal arrows indicate the coding sequences in the exonic regions of the corresponding genes.

Author Contributions

Han Ming Gan and André O. Hudson performed the experiments, analyzed the data, prepared figures and tables and wrote the first draft of the manuscript. Bolaji N. Thomas, and Michael A. Savka were involved in experimental design and manuscript writing/ editing. Nicole T. Cavanaugh, Grace H. Morales,and Ashley N. Mayers were involved in the screening and isolation of *R. mucilaginosa* RIT389.

References

Adeniyi CBA, and Odumosu BT. 2012. Antibacterial and antifungal properties of Distemonanthus benthamianus (Fabaceae) crude extract. *The global journal of pharmaceutical research* 1:567-574.

Arrach N, Fernández-Martín R, Cerdá-Olmedo E, and Avalos J. 2001. A single gene for lycopene cyclase, phytoene synthase, and regulation of carotene biosynthesis in Phycomyces. *Proceedings of the National Academy of Sciences* 98:1687-1692. 10.1073/pnas.98.4.1687

Avalos J, and Carmen Limon M. 2015. Biological roles of fungal carotenoids. *Curr Genet* 61:309-324.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, and Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455-477.

Borodovsky M, and Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinformatics* 4:1-10.

Capella-Gutierrez S, Silla-Martinez JM, and Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-1973.

Cerdá-Olmedo E. 1989. Production of Carotenoids with Fungi. In: Vandamme EJ, ed. *Biotechnology of Vitamins, Pigments and Growth Factors*. Dordrecht: Springer Netherlands, 27-42.

Cutzu R, Coi A, Rosso F, Bardi L, Ciani M, Budroni M, Zara G, Zara S, and Mannazzu I. 2013. From crude glycerol to carotenoids by using a Rhodotorula glutinis mutant. *World J Microbiol Biotechnol* 29:1009-1017.

Davoli P, Mierau V, and Weber RWS. 2004. Carotenoids and Fatty Acids in Red Yeasts Sporobolomyces roseus and Rhodotorula glutinis. *Applied Biochemistry and Microbiology* 40:392-397. 10.1023/B:ABIM.0000033917.57177.f2

Deligios M, Fraumene C, Abbondio M, Mannazzu I, Tanca A, Addis MF, and Uzzau S. 2015. Draft Genome Sequence of Rhodotorula mucilaginosa, an Emergent Opportunistic Pathogen. *Genome Announcements* 3:e00201-00215. 10.1128/genomeA.00201-15

Echavarri-Erasun C, and Johnson EA. 2002. Fungal carotenoids. *Applied Mycology and Biotechnology* 2:45-85. http://dx.doi.org/10.1016/S1874-5334(02)80006-5

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* 7:20.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Gan HM, Schultz MB, and Austin CM. 2014. Integrated shotgun sequencing and bioinformatics pipeline allows ultra-fast mitogenome recovery and confirms substantial gene rearrangements in Australian freshwater crayfishes. *BMC Evol Biol* 14:1471-2148.

Ghani NA, Sulaiman J, Ismail Z, Chan X-Y, Yin W-F, and Chan K-G. 2014. Rhodotorula Mucilaginosa, a Quorum Quenching Yeast Exhibiting Lactonase Activity

Isolated from a Tropical Shoreline. *Sensors (Basel, Switzerland)* 14:6463-6473. 10.3390/s140406463

Goordial J, Davila A, Lacelle D, Pollard W, Marinova MM, Greer CW, DiRuggiero J, McKay CP, and Whyte LG. 2016a. Nearing the cold-arid limits of microbial life in permafrost of an upper dry valley, Antarctica. *The ISME Journal* 10:1613-1624. 10.1038/ismej.2015.239

Goordial J, Raymond-Bouchard I, Riley R, Ronholm J, Shapiro N, Woyke T, LaButti KM, Tice H, Amirebrahimi M, Grigoriev IV, Greer C, Bakermans C, and Whyte L. 2016b. 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 Announcements* 4. 10.1128/genomeA.00069-16

Grandjean F, Tan MH, Gan HM, Lee YP, Kawai T, Distefano RJ, Blaha M, Roles AJ, and Austin CM. 2017. Rapid recovery of nuclear and mitochondrial genes by genome skimming from Northern Hemisphere freshwater crayfish. *Zoologica Scripta*.

Ignatova LV, Brazhnikova YV, Berzhanova RZ, and Mukasheva TD. 2015. Plant growth-promoting and antifungal activity of yeasts from dark chestnut soil. *Microbiological Research* 175:78-83. http://dx.doi.org/10.1016/j.micres.2015.03.008

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, and Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236-1240. 10.1093/bioinformatics/btu031

Libkind D, Brizzio S, and van Broock M. 2004. Rhodotorula mucilaginosa, a carotenoid producing yeast strain from a Patagonian high-altitude lake. *Folia Microbiol* 49:19-25.

Marova I, Carnecka M, Halienova A, Certik M, Dvorakova T, and Haronikova A. 2012. Use of several waste substrates for carotenoid-rich yeast biomass production. *J Environ Manage* 95:8.

Mirebrahim H, Close TJ, and Lonardi S. 2015. De novo meta-assembly of ultra-deep sequencing data. *Bioinformatics* 31.

Noble LM, and Andrianopoulos A. 2013. Fungal Genes in Context: Genome Architecture Reflects Regulatory Complexity and Function. *Genome Biology and Evolution* 5:1336-1352. 10.1093/gbe/evt077

Prado-Cabrero A, Scherzinger D, Avalos J, and Al-Babili S. 2007. Retinal biosynthesis in fungi: characterization of the carotenoid oxygenase CarX from Fusarium fujikuroi. *Eukaryot Cell* 6:650-657.

Price MN, Dehal PS, and Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:0009490.

Richter M, Rossello-Mora R, Oliver Glockner F, and Peplies J. 2016. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32:929-931.

Salvo JL, Rodeghier B, Rubin A, and Troischt T. 1998. Optional introns in mitochondrial DNA of Podospora anserina are the primary source of observed size polymorphisms. *Fungal Genet Biol* 23:162-168.

Sanz C, Velayos A, Álvarez MI, Benito EP, and Eslava AP. 2011. Functional Analysis of the Phycomyces carRA Gene Encoding the Enzymes Phytoene Synthase and Lycopene Cyclase. *PLoS ONE* 6:e23102. 10.1371/journal.pone.0023102

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, and Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-3212.

Stöver BC, and Müller KF. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7. 10.1186/1471-2105-11-7

Sullivan MJ, Petty NK, and Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* 27:1009-1010.

Taccari M, Canonico L, Comitini F, Mannazzu I, and Ciani M. 2012. Screening of yeasts for growth on crude glycerol and optimization of biomass production. *Bioresour Technol* 110:488-495.

Verdoes JC, Sandmann G, Visser H, Diaz M, van Mossel M, and van Ooyen AJJ. 2003. Metabolic Engineering of the Carotenoid Biosynthetic Pathway in the Yeast Xanthophyllomyces dendrorhous (Phaffia rhodozyma). *Applied and Environmental Microbiology* 69:3728-3738. 10.1128/aem.69.7.3728-3738.2003

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, and Schatz MC. 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 24.

Wang Q, Liu D, Yang Q, and Wang P. 2017. Enhancing carotenoid production in Rhodotorula mucilaginosa KC8 by combining mutation and metabolic engineering. *Annals of Microbiology* 67:425-431. 10.1007/s13213-017-1274-2) contains supplementary material, which is available to authorized users.

10.1007/s13213-017-1274-2

Wirth F, and Goldani LZ. 2012. Epidemiology of Rhodotorula: An Emerging Pathogen. *Interdisciplinary Perspectives on Infectious Diseases* 2012:7. 10.1155/2012/465717

Wolters JF, Chiu K, and Fiumera HL. 2015. Population structure of mitochondrial genomes in Saccharomyces cerevisiae. *BMC Genomics* 16:451. 10.1186/s12864-015-1664-4

Zaas AK, Boyce M, Schell W, Lodge BA, Miller JL, and Perfect JR. 2003. Risk of fungemia due to Rhodotorula and antifungal susceptibility testing of Rhodotorula isolates. *J Clin Microbiol* 41:5233-5235.

Zhao XQ, Aizawa T, Schneider J, Wang C, Shen RF, and Sunairi M. 2013. Complete mitochondrial genome of the aluminum-tolerant fungus Rhodotorula taiwanensis RS1 and comparative analysis of Basidiomycota mitochondrial genomes. *MicrobiologyOpen* 2:308-317. 10.1002/mbo3.74

500

Table 1(on next page)

Strain RIT389 genome statistic and strain information

Table 1. Strain RIT389 genome statistic and strain information

Organism	*Rhodotorula mucilaginosa*
Strain Name	RIT389
SRA	SRR5860569
Bioproject	PRJNA390458
Biosample	SAMN07235707
Whole genome:	
Accession Number	NIUW01000000
Assembled genome length	19,664,434 bp
N_{50} length	194,287 bp
Number of contigs	250
GC %	60.28%
Predicted protein-coding gene	7,065
BUSCO Completeness (Basidiomycota odb9)	
Complete BUSCOs	89.70%
Complete and single-copy BUSCOs	86.70%
Complete and duplicated BUSCO	3.00%
Fragmented BUSCO	1.60%
Missing BUSCO	8.70%
Total BUSCO groups searched	1,335
Mitochondrial Genome	
Accession Number	MF694646
Genome Size	47,023 bp
GC %	40.43%
Coverage	400×
Alignment Rate	4.55%

Figure 1(on next page)

(A) Color/morphology of *Rhodotorula mucilaginosa* RIT389 grown on half-strength tryptic soy agar (B) Scanning electron microscopy of *Rhodotorula mucilaginosa* RIT389 at 52.8K magnification.

(A) Color/morphology of *Rhodotorula mucilaginosa* RIT389 grown on half-strength tryptic soy agar (B) Scanning electron microscopy of *Rhodotorula mucilaginosa* RIT389 at 52.8K magnification.

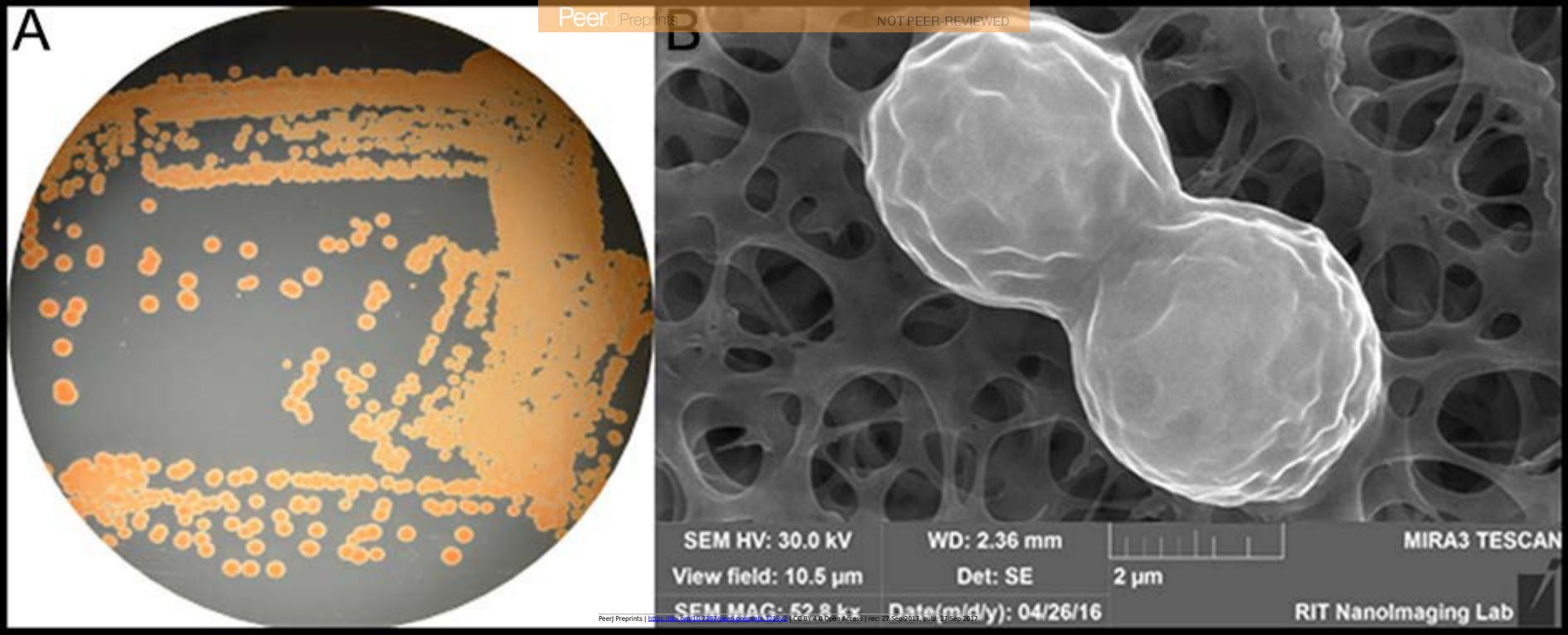
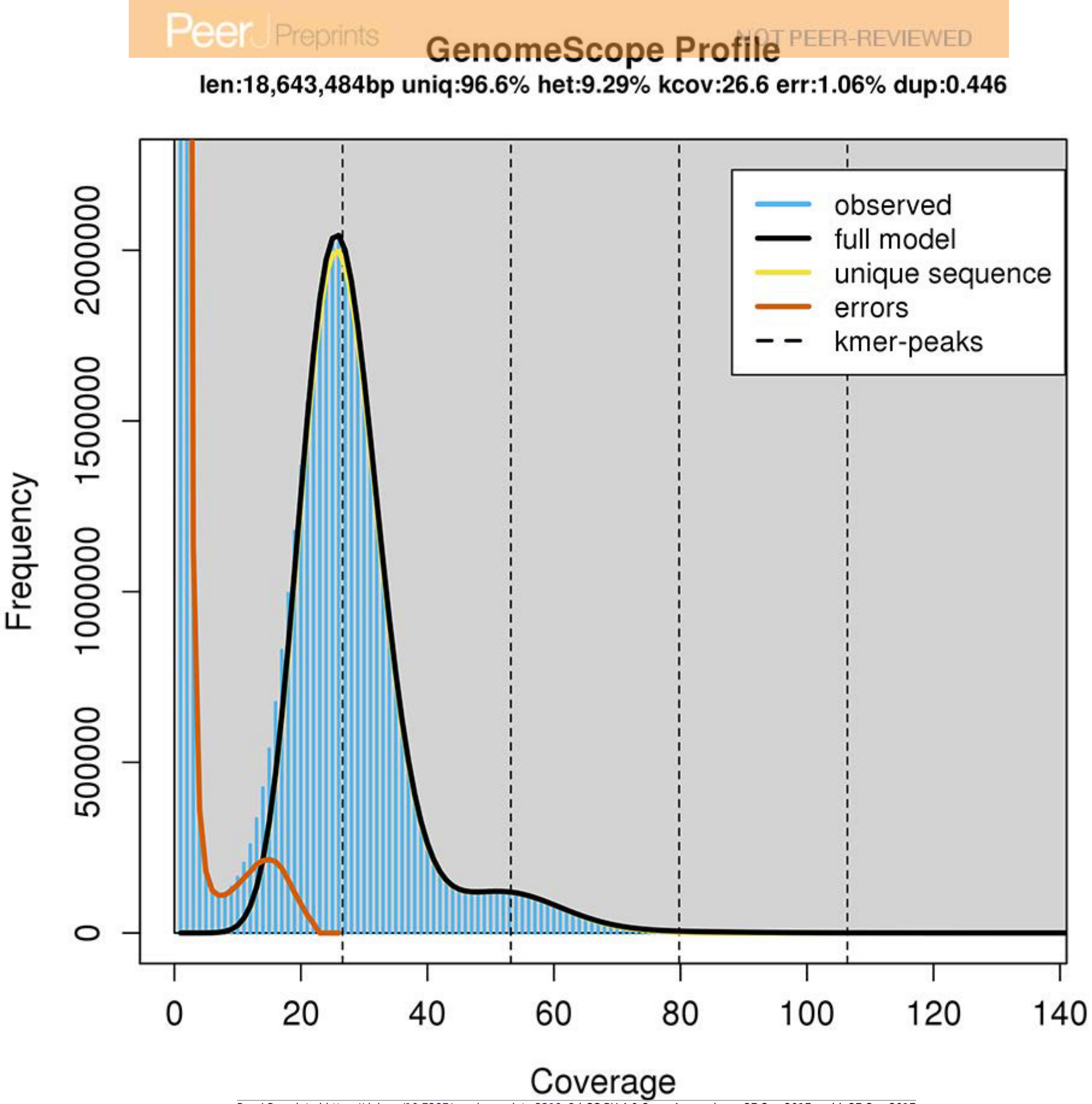


Figure 2(on next page)

GenomeScope estimation of genome size, repeat content and heterozygosity (Kmer length = 21, Read length = 251 bp and Max kmer coverage = 1,000).

GenomeScope estimation of genome size, repeat content and heterozygosity (Kmer length = 21, Read length = 251 bp and Max kmer coverage = 1,000).



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3219v2 | CC BY 4.0 Open Access | rec: 27 Sep 2017, publ: 27 Sep 2017

Figure 3(on next page)

Pairwise average nucleotide identity calculation of Rhodotorula genomes.

Pairwise average nucleotide identity calculation of Rhodotorula genomes. Genomes with the superscript “T” are type strains.

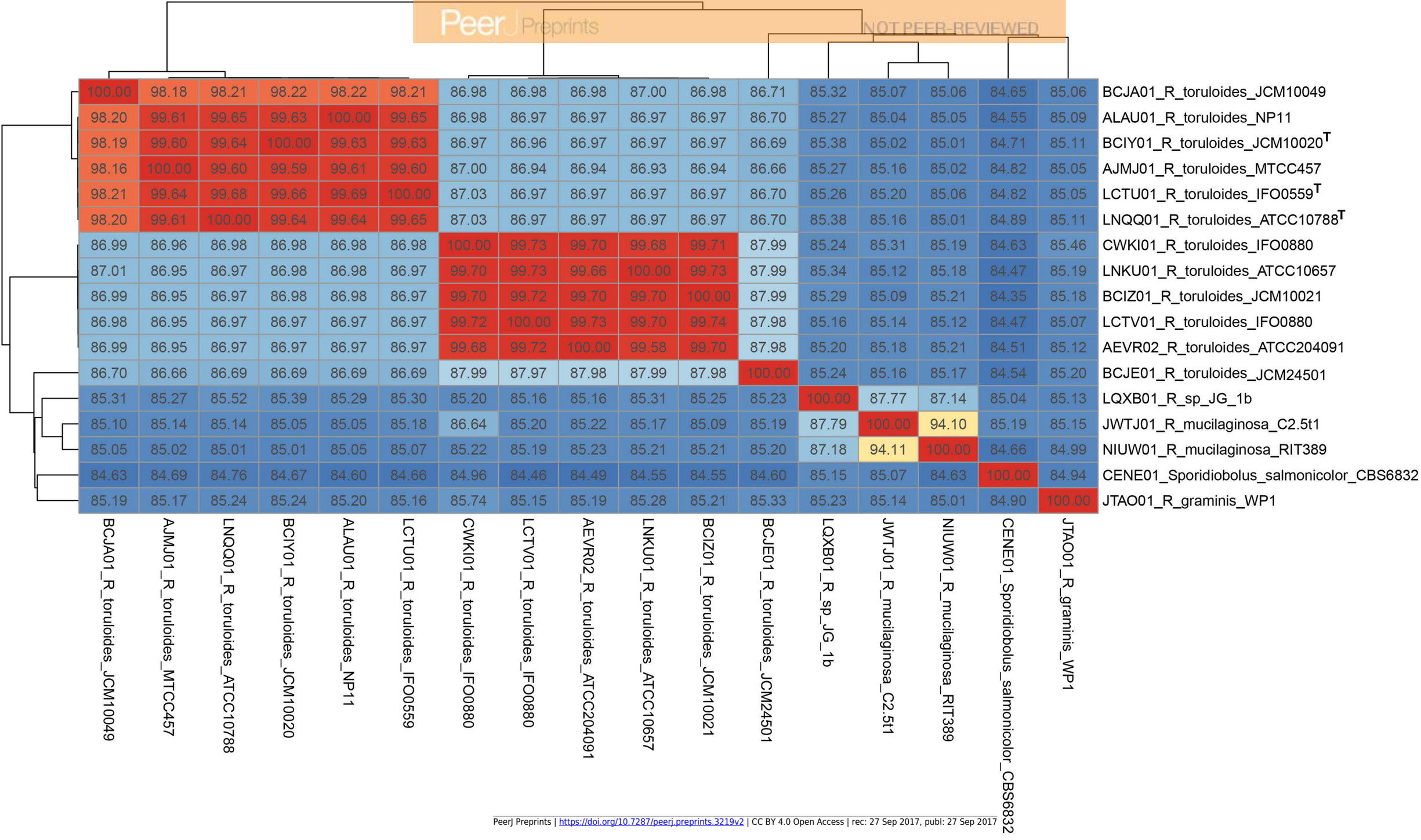


Figure 4(on next page)

Maximum likehood tree of a concatenated amino acid alignment

Maximum likelihood tree of a concatenated amino acid alignment consisting of 537,792 sites that represent 798 universally present single-copy genes from 19 fungal strains. Labels on branches indicate shimodaira-hasegawa (SH)-like local branch support values. The scale bar indicates the average number of amino acid substitutions per site.

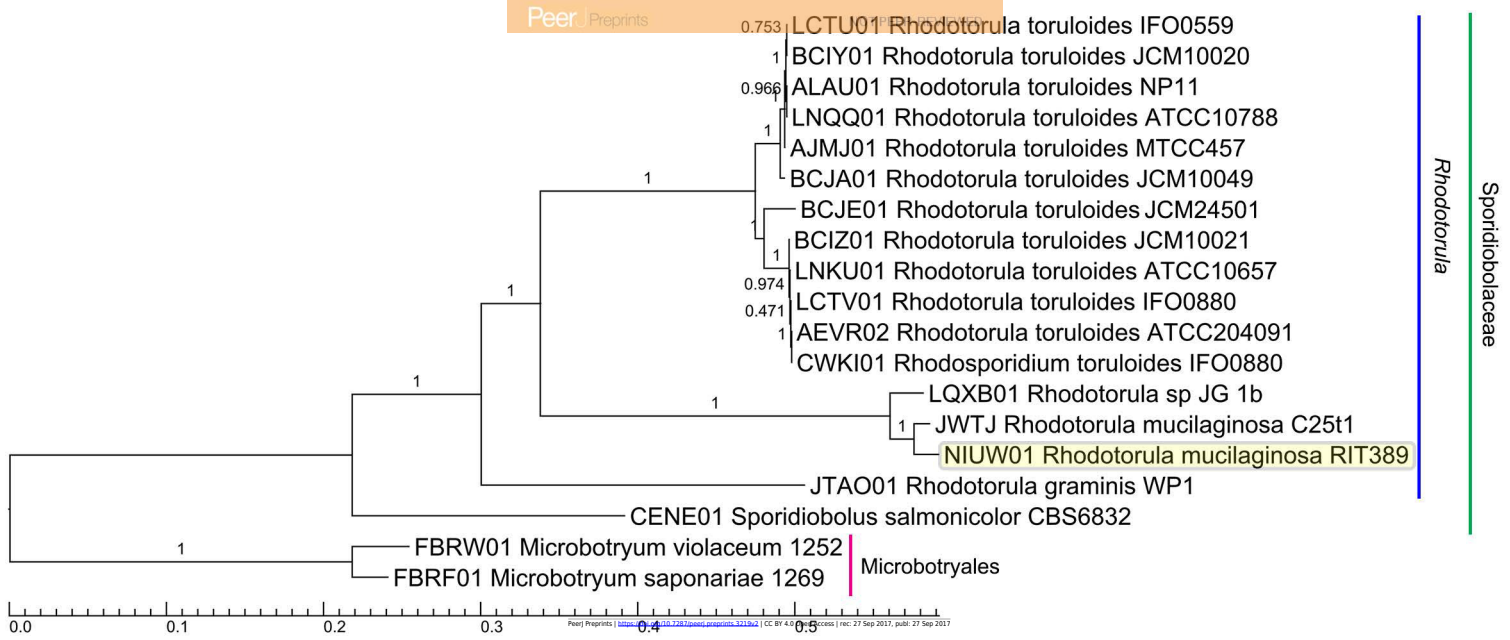


Figure 5(on next page)

Complete mitochondrial genome of *R. mucilaginosa* RIT389 compared against that of *R. taiwanensis* RS1.

Complete mitochondrial genome of *R. mucilaginosa* RIT389 compared against that of *R. taiwanensis* RS1. Orange frames indicate coding sequences commonly found in a typical mitochondrial genome. Red and Blue arrows indicate transfer and ribosomal RNAs, respectively. Arrow direction represents transcriptional orientation. Dotted lines indicate intronic regions.

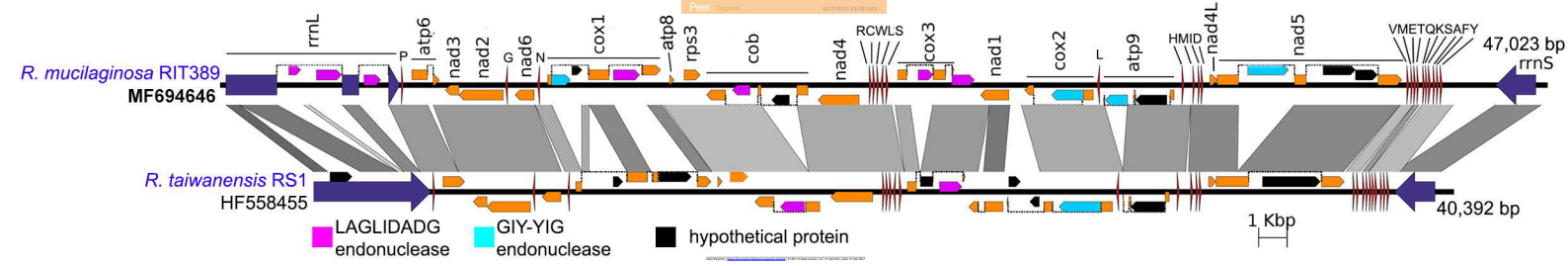
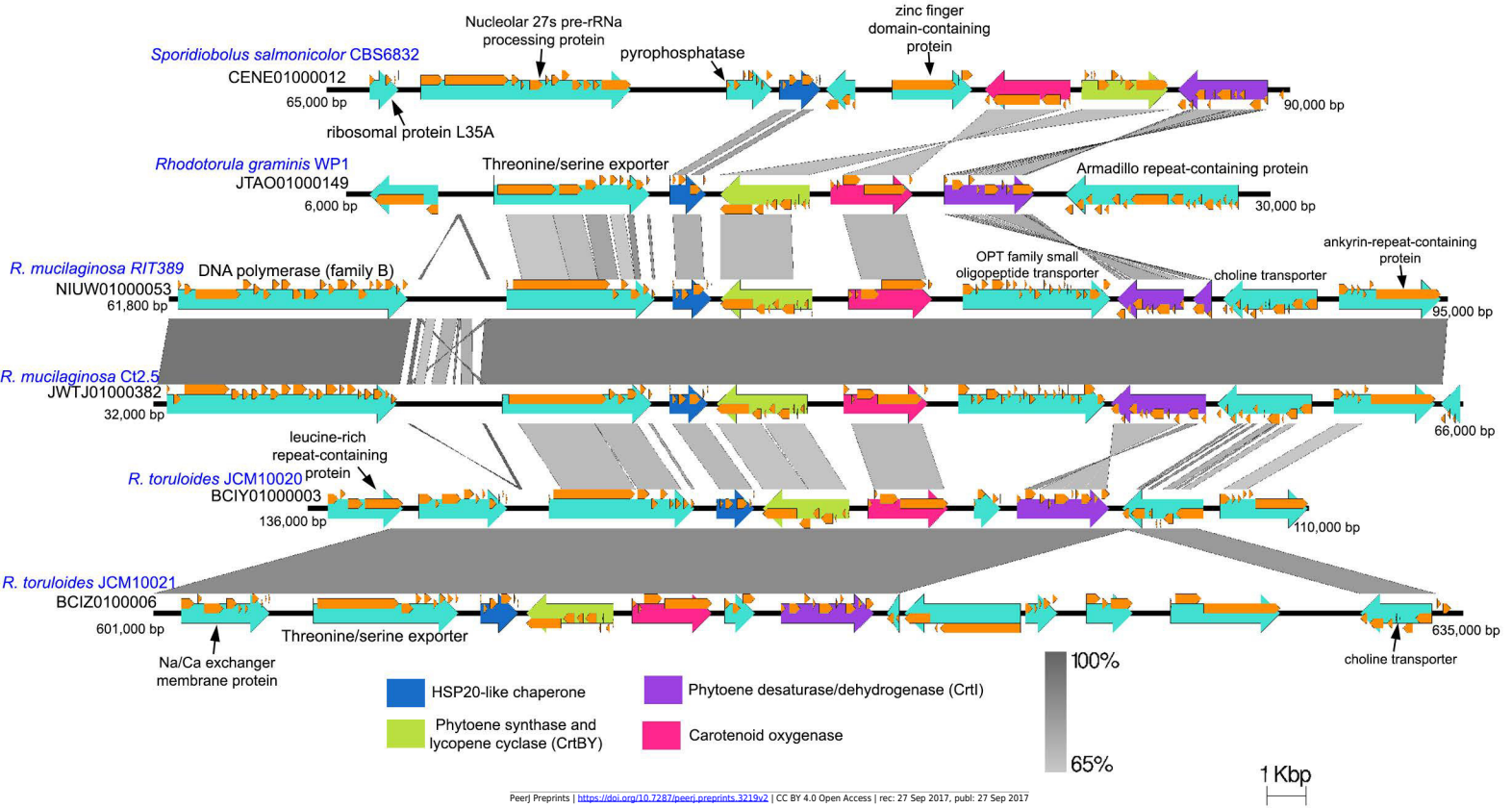

rrnL
P
atp6
nad3
nad2
G
nad6
N
cox1
atp8
rps3
cob
nad4
RCWLS
cox3
nad1
cox2
L
atp9
HMID
nad4L
nad5
VMETQKSAFY
47,023 bp
rrnS
R. mucilaginosa RIT389
MF694646
R. taiwanensis RS1
HF558455
40,392 bp
1 Kbp
LAGLIDADG endonuclease
GIY-YIG endonuclease
hypothetical protein

Figure 6(on next page)

Comparison of genomic sub-region containing the gene cluster associated with carotenoid biosynthetic pathway.

Comparison of genomic sub-region containing the gene cluster associated with carotenoid biosynthetic pathway. Orange frames within the teal arrows indicate the coding sequences in the exonic regions of the corresponding genes.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3219v2 | CC BY 4.0 Open Access | rec: 27 Sep 2017, publ: 27 Sep 2017