


# Draft genome sequences of strains CBS6241 and CBS6242 of the basidiomycetous yeast *Filobasidium floriforme*

Marco Alexandre Guerreiro <sup>1</sup>, Steven Ahrendt,<sup>2</sup> Jasmyn Pangilinan,<sup>2</sup> Cindy Chen,<sup>2</sup> Mi Yan,<sup>2</sup> Anna Lipzen,<sup>2</sup> Kerrie Barry,<sup>2</sup> Igor V. Grigoriev,<sup>2,3</sup> Dominik Begerow,<sup>1</sup> and Minou Nowrousian <sup>4,\*</sup>

<sup>1</sup>Department of Evolution of Plants and Fungi, Ruhr-Universität Bochum, Bochum 44801, Germany

<sup>2</sup>U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

<sup>3</sup>Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720, USA, and

<sup>4</sup>Department of Molecular and Cellular Botany, Ruhr-Universität Bochum, Bochum 44801, Germany

\*Corresponding author: Lehrstuhl für Molekulare und Zelluläre Botanik, Ruhr-Universität Bochum ND 7/176, Universitätsstraße 150, Bochum 44801, Germany.  
Email: minou.nowrousian@rub.de

## Abstract

The Tremellomycetes are a species-rich group within the basidiomycete fungi; however, most analyses of this group to date have focused on pathogenic *Cryptococcus* species within the order Tremellales. Recent genome-assisted studies of other Tremellomycetes have identified interesting features with respect to biotechnological applications as well as the evolution of genes involved in mating and sexual development. Here, we report genome sequences of two strains of *Filobasidium floriforme*, a species from the order Filobasidiales, which branches basally to the Tremellales, Trichosporonales, and Holtermanniales. The assembled genomes of strains CBS6241 and CBS6242 are 27.4 Mb and 26.4 Mb in size, respectively, with 8314 and 7695 predicted protein-coding genes. Overall sequence identity at nucleic acid level between the strains is 97%. Among the predicted genes are pheromone precursor and pheromone receptor genes as well as two genes encoding homeodomain (HD) transcription factors, which are predicted to be part of the mating type (*MAT*) locus. Sequence analysis indicates that CBS6241 and CBS6242 carry different alleles for both the pheromone/receptor genes as well as the HD transcription factors. Orthology inference identified 1482 orthogroups exclusively found in *F. floriforme*, some of which were involved in carbohydrate transport and metabolism. Subsequent CAZyme repertoire characterization identified 267 and 247 enzymes for CBS6241 and CBS6242, respectively, the second highest number of CAZymes among the analyzed Tremellomycete species. In addition, *F. floriforme* contains five CAZymes absent in other species and several plant-cell-wall degrading CAZymes with the highest copy number in Tremellomycota, indicating the biotechnological potential of this species.

**Keywords:** *Filobasidium floriforme*; mating-type locus; basidiomycete; Filobasidiales; CAZymes

## Introduction

The basidiomycete group of Tremellomycetes is a species-rich group comprising both filamentous and yeast-like fungi (Liu *et al.* 2015a, 2015b; Spatafora *et al.* 2017). In-depth studies in this group have mostly focused on pathogenic *Cryptococcus* species (order Tremellales) (Sun *et al.* 2019a; Bahn *et al.* 2020). However, facilitated by increased genome-sequencing capacities in recent years, additional species within the Tremellomycetes have been investigated, often for their biotechnological potential or to study the evolution of sexual development in this group (Sharma *et al.* 2015; Bellora *et al.* 2016; Barredo *et al.* 2017; Bracharz *et al.* 2017; Coelho *et al.* 2017; Sun *et al.* 2019b; Aliyu *et al.* 2020).

Mating and sexual development in basidiomycetes is regulated by mating type (*MAT*) genes encoding pheromone precursors, pheromone receptors, and homeodomain (HD) transcription factors. The ancestral state in basidiomycetes is thought to be tetrapolar, with two nonlinked genetic loci containing pheromone and pheromone receptor genes (*P/R* locus) and HD transcription factor genes (*HD* locus), respectively (Kües *et al.* 2011;

Coelho *et al.* 2017). However, studies of pathogenic *Cryptococcus* species revealed a single, large *MAT* locus predicted to have arisen from genomic transitions leading to fusion of the formerly unlinked *P/R* and *HD* loci (Lengeler *et al.* 2002; Fraser *et al.* 2004), whereas in all other species of the order Tremellales that were analyzed so far, the ancient tetrapolar arrangement of unlinked *P/R* and *HD* loci is found (Sun *et al.* 2019a). In contrast, in the Trichosporonales, the sister order to the Tremellales (Liu *et al.* 2015b), a recent analysis of *MAT* loci revealed that in all analyzed species, *P/R* and *HD* loci are physically linked in a single *MAT* locus (Sun *et al.* 2019b).

For the Tremellomycete order Filobasidiales, which branches basally to the Tremellales, Trichosporonales, and Holtermanniales (Liu *et al.* 2015b), genomes have been published for the genera *Naganishia* and *Solicoccozyma*, but none were analyzed with respect to the mating type (Close *et al.* 2016; Vajpeyi and Chandran 2016; Yong *et al.* 2016; Bijlani *et al.* 2020; Han *et al.* 2020; Nizovoy *et al.* 2021). Here, we present the first draft genome sequences including

Received: September 20, 2021. Accepted: November 08, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

an analysis of MAT genes for two strains of the genus *Filobasidium*, *Filobasidium floriforme* strains CBS6241 and CBS6242.

Carbohydrate-active enzymes (CAZymes) are essential for fungi as heterotrophic organisms. These enzymes are responsible for the biosynthesis, modification, binding, and breakdown of carbohydrates and glycoconjugates (Cantarel et al. 2009). Different fungal species harbor different sets of CAZymes to meet their ecological needs as saprobes, symbionts, endophytes, parasites, or pathogens (Rytioja et al. 2014; Kameshwar and Qin 2018). CAZyme content and diversity is therefore suggested to reflect niche adaptation. CAZymes are widely applied in various biotechnological processes and industries, such as in food, wine, paper, pulp, textile, detergents, biofuels, biorefinery, and bioremediation (Mäkelä et al. 2014). Since Tremellomycetes are known for their ability to colonize and inhabit a vast diversity of substrates, characterizing their set of CAZymes provides a great opportunity to identify and characterize species with biotechnological potential.

## Materials and methods

### Strains and culture conditions

Strains CBS6241 and CBS6242 were obtained from the Westerdijk Fungal Biodiversity Institute (Utrecht, The Netherlands). Strains were kept on YPD agar medium at 25°C.

### DNA extraction and sequencing

For DNA extraction, strains were grown as pre-cultures for 4 days at 25°C on YPD agar medium. From single colonies, 30 ml liquid YPD medium was inoculated and cultures were incubated at 25°C and 100 rpm on a shaker for 24 h ( $OD_{600} > 1$ ). DNA was extracted as described previously (Kourist et al. 2015).

The genome of strain CBS6241 was sequenced using Pacific Biosciences Sequel sequencing platform. One microgram of genomic DNA was sheared to 10 kb using Covaris g-TUBE. The sheared DNA was treated with DNA damage repair mix followed by end repair and ligation of blunt adapters using SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). The library was purified with AMPure PB beads. PacBio Sequencing primer was then annealed to the SMRTbell template library and sequencing polymerase was bound to them using Sequel Binding kit 3.0. The prepared SMRTbell template libraries were then sequenced on a Pacific Biosystem's Sequel sequencer using v3 sequencing primer, 1M v3 SMRT cells, and Version 3.0 sequencing chemistry with  $1 \times 360$  and  $1 \times 600$  sequencing movie run times. A total of 3,594,174 reads were obtained with an average length of 3.8 kb and an N50 of 5.7 kb.

Library preparation and Illumina sequencing of strain CBS6242 was performed by Eurofins (Konstanz, Germany). Paired-end reads of 151 nt were sequenced from a library with an average insert size of 330 nt.

### RNA extraction, sequencing, and transcriptome assembly

For RNA extraction, CBS6241 was grown as pre-culture for 4 days at 25°C on YPD agar medium. From single colonies, 30 ml of three different liquid media were inoculated (V8: 50 ml/l vegetable juice, pH 5.2; V8-YPD1: 50 ml/l vegetable juice, 10 g/l tryptone, 5 g/l yeast extract, 10 g/l glucose, pH 5.2; V8-YPD2: 25 ml/l vegetable juice, 20 g/l tryptone, 5 g/l yeast extract, 10 g/l glucose, pH 5.2). Cultures were incubated at 25°C and 100 rpm on a shaker for 24 h. RNA was extracted as described previously (Kourist et al. 2015).

The transcriptome of strain CBS6241 was sequenced using Illumina  $2 \times 150$  paired-end reads. Stranded cDNA library was generated using the Illumina Truseq Stranded RNA LT kit. mRNA was purified from 1  $\mu$ g of total RNA using magnetic beads containing poly-T oligonucleotides. mRNA was fragmented and reversed transcribed using random hexamers and SSII (Invitrogen) followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and eight cycles of PCR. The prepared library was quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. Sequencing of the library was performed on the Illumina NovaSeq sequencer using NovaSeq XP V1 reagent kits, S4 flow-cell, following a  $2 \times 150$  dependent indexed run recipe. Raw reads were filtered and trimmed. Using BBDuk (<https://sourceforge.net/projects/bbmap/>, Accessed: 2021 November 19), raw reads were evaluated for artifact sequence by kmer matching (kmer = 25), allowing 1 mismatch and detected artifact was trimmed from the 3' end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, following trimming, reads under the length threshold were removed (minimum length 25 bases or 1/3 of the original read length—whichever is longer). Filtered reads were assembled into consensus sequences using Trinity (v2.3.2) (Grabherr et al. 2011), run with the `-normalize_reads` (In-silico normalization routine) and `-jaccard_clip` (Minimizing fusion transcripts derived from gene dense genomes) options. The transcriptome data were used for annotation.

### Genome assembly and annotation

Filtered Pacific Biosciences subread data for strain CBS6241 was filtered for artifacts and assembled with Falcon version `pb-assembly = 0.0.2|falcon-kit = 1.2.3|pyflow = 2.1.0` (<https://github.com/PacificBiosciences/FALCON>, Accessed: 2021 November 19) and polished with Arrow version SMRTLink v7.0.1.66975 (<https://www.pacb.com/support/software-downloads>, Accessed: 2021 November 19), improved with finisherSC version 2.1 (Lam et al. 2015), and polished with Arrow version SMRTLINK v7.0.1.66975 (<https://www.pacb.com/support/software-downloads>, Accessed: 2021 November 19). The genome was annotated using the JGI Annotation pipeline (Grigoriev et al. 2014).

For strain CBS6242, Illumina reads were quality-trimmed with Trimmomatic v0.36 (Bolger et al. 2014) and assembled with SPAdes v3.14.0 (Prjibelski et al. 2020). Contigs  $> 1$  kb were kept for downstream analyses. Contigs were error-corrected with Pilon v1.22 (Walker et al. 2014) based on the Illumina reads mapped to the assembly with Bowtie2 v2.2.6 (Langmead and Salzberg 2012). For CBS6242, genes were predicted with Maker v2.31.8 (Cantarel et al. 2008) based on the predicted genes of CBS6241. Pheromone genes in both strains were identified with a custom-made Perl script (Supplementary Text S1) to search for the consensus sequence M-X(15-60)-CAAX-Stop, with X representing any amino acid and A representing the amino acids valine, leucine, isoleucine, methionine, threonine or serine. Putative telomeric repeats (sequence TTAGGGG occurring consecutively at least three times) were identified with a custom-made Perl script (Supplementary Text S2).

### Phylogenetic analysis and functional annotation

Published genome assembly data of tremellomycetous species were collected from NCBI and JGI databases (Supplementary

Table S1). The respective proteomes were predicted by Augustus version 3.3.3 (Stanke et al. 2008), with *Cryptococcus neoformans* as reference organism. Orthologous protein sequences were identified with OrthoFinder v2.5.2 (Emms and Kelly 2019). Single-copy orthologous sequences present in all species were individually aligned with MAFFT v7.273 (Katoh and Standley 2013) and concatenated. The maximum likelihood phylogenetic tree was calculated based on a single alignment of 142 single-copy orthologous genes by RAxML v8.2.12 (Stamatakis 2014) using 500 bootstrap replications, the PROTGAMEWAG model, 123 as seed number for the parsimony inferences and a random seed of 321.

The strains *Cryptococcus deneoformans* JEC21 (Loftus et al. 2005), *Cutaneotrichosporon oleaginosum* IBC0246 (Kourist et al. 2015) and *Cystofilobasidium capitatum* CBS7420 (David-Palma et al. 2020) were selected for further analyses as representatives of Tremellales, Trichosporonales and Cystofilobasidiales, respectively (one species per order). Orthology comparisons were calculated with ComplexUpset (v1.3.1) package for R v4.1.0 and “exclusive intersection” mode.

Genes comprised in orthogroups exclusively found in *F. floriforme* were extracted and functionally annotated with PFAM v34.0 online database (Mistry et al. 2021) and eggNOG-mapper v2.1.0-1 (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021).

## Analysis of the MAT regions

Homologs to MAT genes were identified with BLAST analyses (Altschul et al. 1997) using MAT proteins from *Cryptococcus neoformans* (Lengeler et al. 2002) as queries. Phylogenetic trees of Ste3 proteins from *F. floriforme* strains and from published sequences of Trichosporonales and Tremellales species (Lengeler et al. 2000, 2002; Kourist et al. 2015; Takashima et al. 2015, 2018; Cho et al. 2016; Sriswasdi et al. 2016; Sun et al. 2019b) were generated with PAUP version 4.0b10 for Windows (D.L. Swofford, distributed by Sinauer Associates, copyright 2001 Smithsonian Institution) for Neighbor joining analyses or with MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) based on multiple alignments generated with CLUSTALX (Thompson et al. 1997). Comparisons of genomic regions at nucleic acid level were performed with nucmer from the MUMmer package (Kurtz et al. 2004).

## Prediction of CAZymes

Proteomes of published datasets in Filobasidiales and of representative members of Tremellales, Trichosporonales and Cystofilobasidiales were scanned against dbCAN2 v9.0 database (Zhang et al. 2018) using HMMER v3.3.2 (Eddy 2011). Matches with an e-value lower than  $1e^{-15}$  and a coverage higher than 0.35 were used for further analyses.

## Results and discussion

### Genome assembly and assessment

The genome of the *F. floriforme* strain CBS6241 was sequenced as part of the 1000 Fungal Genomes project (<http://1000.fungalgenomes.org>) (Grigoriev et al. 2011, 2014) using PacBio sequencing, while strain CBS6242 was sequenced with Illumina sequencing. With assembly sizes of 26–27 Mb and 7695 and 8314 predicted genes (Table 1), the genomes of the two *F. floriforme* strains are in the same range as the previously sequenced 24.8 Mb genomes of two strains of the Filobasidiales species *Naganishia albida*, for which 7375 and 8637 genes were predicted (Vajpeyi and Chandran 2016; Yong et al. 2016).

**Table 1** Genome assembly statistics for CBS6241 and CBS6242

	CBS6241	CBS6242
Assembly size (Mb)	27.5	26.4
No. of scaffolds	42	700
N50 (kb)	9388	120
GC content (%)	52.9	52.9
Predicted genes	8314	7695
completeness (%) <sup>a</sup>	94.9	92.6
Coding regions (%)	46.0	44.3

<sup>a</sup> Completeness was analyzed with BUSCO v5.2.1 (Manni et al. 2021).

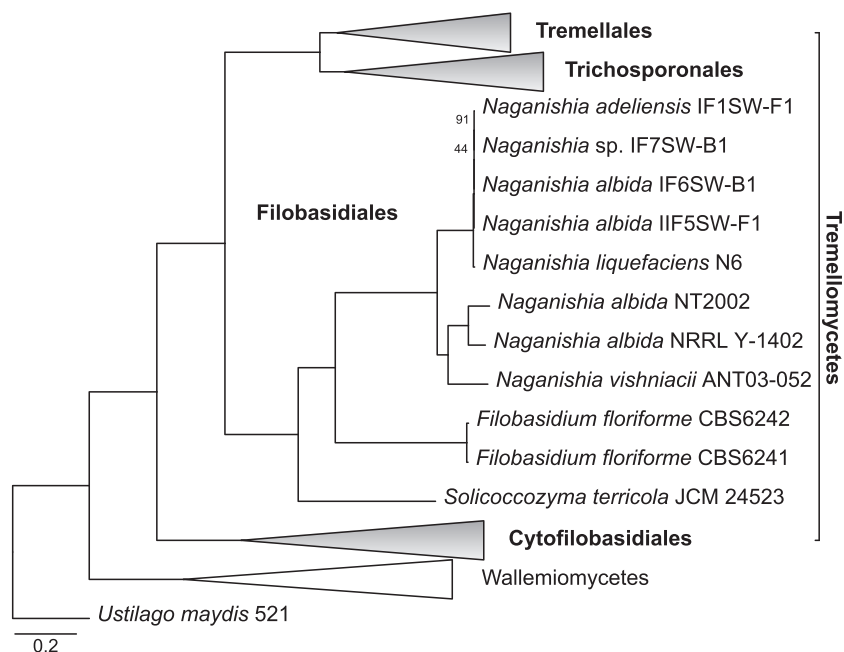
Searches for putative telomeric repeats identified seven contigs in CBS6241 with putative telomeric repeats at both ends, and 23 contigs with putative telomeric repeats at one end (Supplementary Table S2). This suggests that the genome of CBS6241 consists of (at least) 19 chromosomes. In the Illumina-sequenced genome of CBS6242, no putative telomeric repeats were identified, as is to be expected in a genome assembled from short reads.

### Phylogenetic analysis

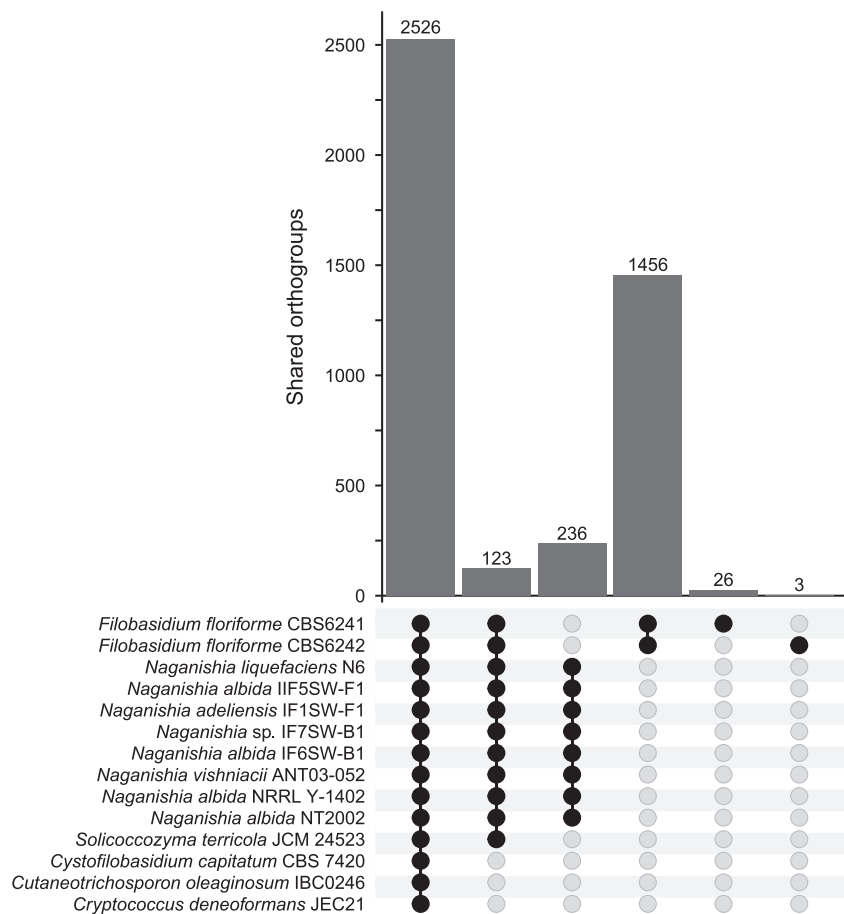
The maximum likelihood analysis of the 142 single-copy orthologous protein sequences produced a highly resolved phylogeny of Tremellomycota, in which most clades were supported by bootstrap values of 100% (Figure 1, Supplementary Figure S1). The Filobasidiales order is monophyletic and basal to the orders Tremellales and Trichosporonales. Cystofilobasidiales is the most basal order in Tremellomycota. The calculated phylogeny is consistent with previous reconstructions (Liu et al. 2015a, 2015b). Currently, classifications suggest that the *Filobasidium* genus is likely monophyletic and closely related to *Naganishia* and *Solicocozyma* genera (Kwon-Chung 2011; Liu et al. 2015a), supporting our results. Due to their placement in the phylogeny (Figure 1, Supplementary Figure S1), a taxonomic revision of available *Naganishia* genomes might be advisable for future studies.

### Orthologous genes and functional assignment

Orthology inference analysis revealed that 2526 orthogroups were shared among all the selected Tremellomycetes species (Figure 2). Furthermore, 123 were exclusively found among members of Filobasidiales, while 236 were exclusive to the *Naganishia* genus. A total of 1485 unique orthogroups were exclusively found in the two strains of *F. floriforme*, from which 83 orthogroups (comprising 224 orthologous sequences) were successfully functionally assigned by both PFAM and eggNOG databases (Supplementary Table S3). The most represented eggNOG functional categories included unknown function (31 orthogroups), carbohydrate transport and metabolism (13), post-translational modification, protein turnover, and chaperones (10), and replication, recombination, and repair (8). Functional description resulted, among others, in glycosyl hydrolases, dehydrogenases, kinases and proteases, that were specific to *F. floriforme*. These unique features might indicate a unique evolutionary adaptation to the environment (Rytioja et al. 2014; Nizovoy et al. 2021). Both strains comprised similar copy numbers, with a few exceptions in some domains, which could be explained by the different assembly quality of both genomes (Table 1). Orthology inference analysis and functional assignment suggest that *F. floriforme* might comprise unique genomic traits that might be valuable for further functional studies (Nizovoy et al. 2021).



**Figure 1** Phylogenetic analysis of Tremellomycetes. A maximum likelihood analysis of 142 single-copy orthologous protein sequences was performed with 500 bootstrap replicates using *Ustilago maydis* as an outgroup. All depicted nodes showed 100% bootstrap support except when noted. Accession numbers for the genome assembly data is provided in Supplementary Table S1. Groups outside of the Filobasidiales are collapsed, for the full phylogeny, see Supplementary Figure S1. The scale bar gives substitutions per site.



**Figure 2** Orthogroup analysis to identify shared orthogroups between *F. floriforme* strains and other Tremellomycetes. Black dots indicate species/strains that share the indicated number of orthogroups. Indicated orthogroups are exclusively found in the indicated set of species [shared orthogroups were compared among orders (Tremellales, Trichosporonales, Filobasidiales, and Cystofilobasidiales) and within the *Naganishia* genus].



## Analysis of MAT genes

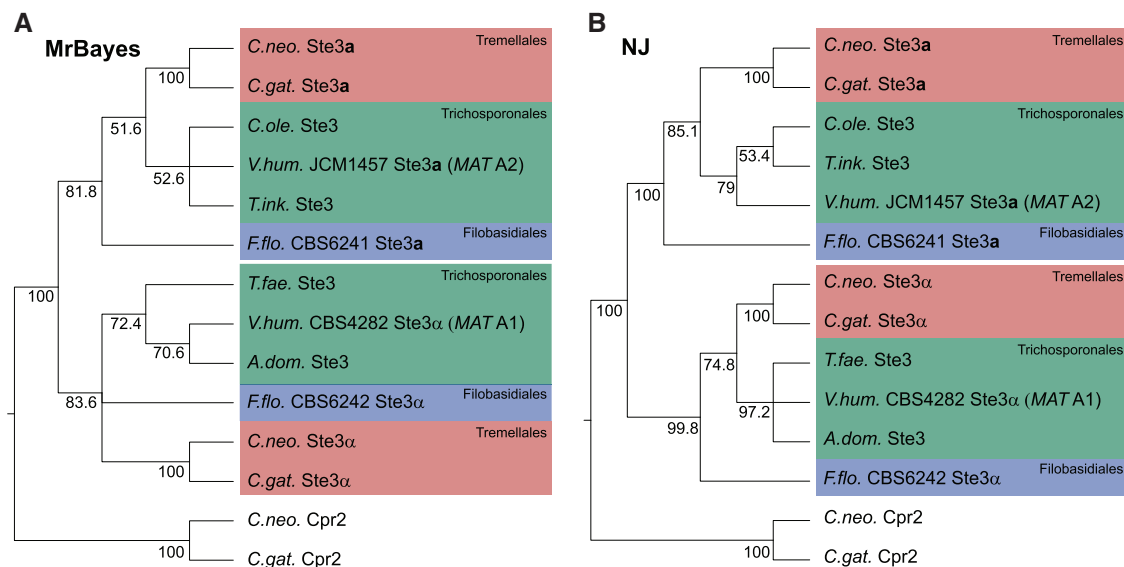
Putative pheromone receptor genes and *HD* genes were identified by BLAST searches with the corresponding *C. neoformans* genes, whereas putative pheromone precursor genes were identified through searches for a consensus sequence with a custom-made Perl script. In each of the *F. floriforme* strains, one *STE3* pheromone receptor gene, one pheromone precursor gene, and two *HD* transcription factor genes (*SXI1* and *SXI2*) were identified (Supplementary Figures S2 and S3). The pheromone receptor gene and pheromone precursor gene of CBS6241 are located in a 21 kb region on contig 14, whereas the *HD* genes are located in a 4 kb region on contig 3. Both contigs have telomeric repeats at both ends (Supplementary Table S2), making it likely that they represent different chromosomes. This suggests that the *P/R* and *HD* loci of CBS6241 are genetically unlinked and that therefore the mating type configuration of *F. floriforme* might be tetrapolar.

The genomic regions containing the *HD* genes are syntenic in CBS6241 and CBS6242 (Supplementary Figure S4). The predicted pheromone precursor gene of CBS6242 is located toward one end of contig 18, and the region is syntenic to the corresponding region in CBS6241 (Supplementary Figure S4). However, the pheromone receptor genes *STE3* is located on a separate, short contig that shows two inversions compared to CBS6241 (Supplementary Figure S4). Furthermore, a repeat-rich region of 15 kb that is present close to *STE3* in CBS6241 was not assembled in CBS6242, probably due to the short read-based assembly. Repeat expansions and sequence divergence have been observed previously in the *MAT* regions of other basidiomycetes and are often associated with sex- or mating type-determining regions with reduced recombination (Lengeler et al. 2002; Loftus et al. 2005; Branco et al. 2017, 2018; Coelho et al. 2017).

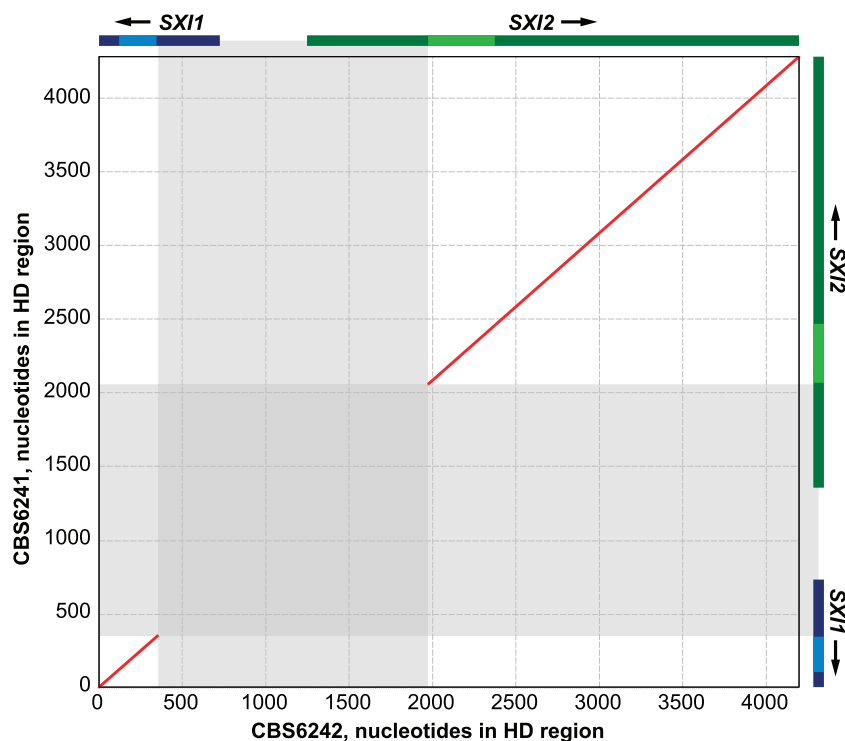
The *STE3* gene of CBS6241 belongs to the *MATa* group of *STE3* pheromone receptor genes, whereas the *STE3* gene of CBS6242 belongs to the *MATα* group of *STE3* genes (Figure 3, Supplementary Figure S2). A phylogenetic analysis of *STE3* alleles from the two *F. floriforme* strains as well as *STE3* genes from Trichosporonales and Tremellales showed trans-species polymorphism already observed for Trichosporonales and Tremellales (Metin et al. 2010; Findley et al. 2012; Sun et al. 2019b), indicating that the trans-species polymorphism of *STE3* alleles was present already in the last common ancestor of the Filobasidiales and its sister orders (Figure 3).

Both strains carry two *HD* genes (*SXI1* and *SXI2*) next to each other but divergently transcribed, which is the typical genomic arrangement for basidiomycete *HD* genes (Coelho et al. 2017). Sequence comparison of the *HD* region of CBS6241 and CBS6242 showed that the region is similar in both strains except for a part encompassing the intergenic region and the N-terminal regions of *SXI1* and *SXI2*, which is highly divergent (Figure 4). This is similar to previous findings in other basidiomycetes, where it was shown that the divergent N-termini of the *SXI1* and *SXI2* proteins are relevant for the interactions of the two different types of *HD* transcription factors, and that allele specificity is conferred by these regions (Banham et al. 1995; Kämper et al. 1995, 2020; Metin et al. 2010; Findley et al. 2012).

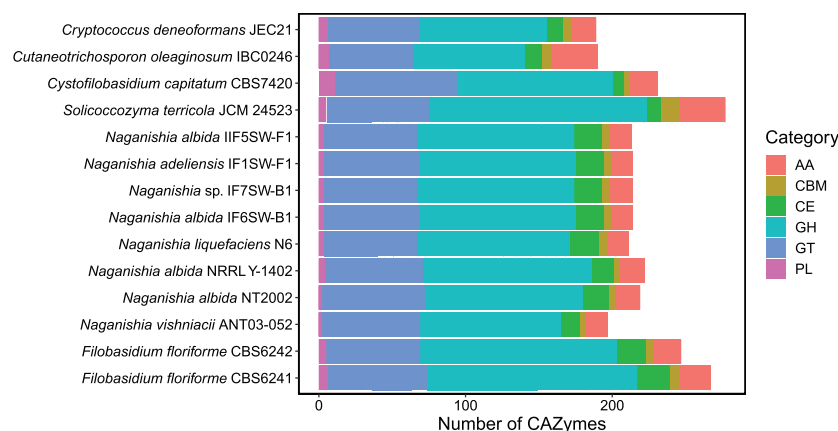
Thus, the two strains carry different and potentially mating-compatible alleles at the *P/R* as well as the *HD* locus, and one can conclude that there are at least two alleles for each *MAT* locus present in the population. Mating between CBS6241 and CBS6242 was observed in 1972 (Rodrigues De Miranda 1972). However, in our laboratories, we were not able to observe sexual structures in co-cultures of the two strains. It is possible that we were not able to recreate the conditions required for mating of the two strains, or that the strains have lost the ability for sexual reproduction



**Figure 3** Phylogenetic analysis of *Ste3* proteins from several Tremellomycetes. Analysis was done by MrBayes (A) or Neighbor joining (NJ, in B). Bayesian probabilities (A) or bootstrap percentages for 1000 bootstrap replications (B) are given at the branches. The pheromone-receptor-like *Cpr2* proteins from *C. neoformans* and *C. gattii* were used as outgroups. The phylogenetic trees show a deep trans-species polymorphism for the *Ste3* proteins within the Tremellomycetes that not only includes the sister orders Tremellales and Trichosporonales, but also includes the early-branching Filobasidiales represented by the two *F. floriforme* strains. Species abbreviations and accession numbers or locus tag numbers: *A. dom*, *Apiotrichum domesticum* (T. domesticum\_002\_745, genome accession BCFW01000000); *C. neo*, *Cryptococcus neoformans* (*Ste3a*: AAN75624.1, *Ste3α*: XP\_012049557.1, *Cpr2*: XP\_012047561.1); *C. gat*, *Cryptococcus gattii* (*Ste3a*: AEG78597.1, *Ste3α*: XP\_003196044.1, *Cpr2*: XP\_003191200.1); *C. ole*, *Cutaneotrichosporon oleaginosum* (XP\_018276494.1); *F. flo*, *Filobasidium floriforme* (CBS6241: gene\_1555, CBS6242: CBS6242\_07693 = FFLO\_06159), *T. fae*, *Trichosporon faecale* (T. faecale\_002\_949, genome accession JXYK01000000), *T. ink*, *Trichosporon inkin* (T. inkin\_003\_120, genome accession JXYM01000000); *V. hum*, *Vanrija humicola* (*Ste3a*: JCM1475\_001\_295, genome accession BCJF01000000, *Ste3α*: TXT13458.1).



**Figure 4** Dot plot of HD transcription factor gene regions of *F. floriforme* strains CBS6241 and CBS6242. The comparison was performed with nucmer from the MUMmer package (Kurtz et al. 2004). Nucleic acid sequence identity in the first aligned region (nt 1–356) is 96.4%, in the second aligned region (nt ~2000–4200) it is 84.0%. No sequence similarity was detected in the region shaded in gray, which comprises the intergenic region and the N-terminal regions of SXI1 and SXI2. The regions encoding the conserved homeodomains in Sxi1 and Sxi2 are shown in light blue and light green, respectively. Gene regions outside of the conserved homeodomain-encoding regions are given in dark blue and dark green for SXI1 and SXI2, respectively. The sequences used for comparison comprise the following genomic regions in the CBS6241 and CBS6242 assemblies, respectively: CBS6241 contig14 nt 669,525–673,807, CBS6242 contig45 nt 55,021–59,216.



**Figure 5** Overview of CAZymes identified in *F. floriforme* compared to other Tremellomycetes. Abbreviations of CAZyme categories: AA, auxiliary activities; CBM, carbohydrate binding modules; CE, carbohydrate esterases; GH, glycoside hydrolases; GT, glycosyltransferases; PL, polysaccharide lyases.

through accumulation of mutations after several decades of being cultured under laboratory conditions. The availability of genome sequences for two strains of *F. floriforme* should facilitate the analysis of MAT alleles of additional strains of this species that might be used in future mating experiments.

### Characterization of CAZymes repertoire

Characterization of CAZymes content revealed that *F. floriforme* contains a high diversity of CAZymes, which includes 267 and

247 genes for CBS6241 and CBS6242, respectively (Figure 5). *Solicoccozyma terricola*, a promising biotechnological strain (Tanimura et al. 2014; Close et al. 2016) contained the highest number of CAZymes (277), but a similar number (149) of glycoside hydrolases (GHs) compared to the *F. floriforme* strains (143 and 134). *Filobasidium floriforme* contained the highest number of carbohydrate esterases (CEs), while *Cystofilobasidium capitatum* contained the most glycosyltransferases (GTs) and polysaccharide lyases (PLs). *Filobasidium floriforme* comprised 5 CAZymes that

were absent in the other organisms (PL3, PL27, GH39, GT28, and GT34) and 8 with a higher number of copies than any of the other species (AA9, CE5, CE9, GH3, GH10, GH31, GH35, and GH43) (Supplementary Figures S5–S10). Interestingly, most of these (AA9, CE5, CE9, GH3, GH10, GH31, GH35, GH43, GT34, and PL3) are important plant-cell-wall degrading enzymes (Zhao et al. 2013; Chang et al. 2016). This highlights the promising vast potential of *F. floriforme* in industrial and biotechnological application (Bosetto et al. 2016).

## Data availability

PacBio reads from CBS6241 (genome sequencing) were deposited in the NCBI SRA database under accession number SRP256613. Illumina reads from CBS6241 (RNA-seq) were deposited at the NCBI SRA database under accession number SRP256618. Illumina reads from CBS6242 (genome sequencing) were deposited in the NCBI SRA database under accession number SRP258233. The CBS6241 genome sequence (BioProject ID PRJNA621322) was deposited at DDBJ/ENA/GenBank under the accession JAIFAB000000000, and the CBS6242 genome sequence (BioProject ID PRJNA627804) under the accession JABELV000000000. Supplementary material (Supplementary Figures S1–S10, Tables S1–S3, Supplementary Texts S1–S2) is part of the manuscript submission. Supplementary Figure S1: Phylogenetic analysis of Tremellomycetes. Supplementary Figure S2: Multiple alignment of Ste3 homologs in the MAT loci of several Tremellomycetes. Supplementary Figure S3: Analysis of Sxi proteins from Tremellomycetes. Supplementary Figure S4: MAT loci of CBS6241 and CBS6242. Supplementary Figure S5: Overview of CAZyme category AA. Supplementary Figure S6: Overview of CAZyme category CBM. Supplementary Figure S7: Overview of CAZyme category CE. Supplementary Figure S8: Overview of CAZyme category GH. Supplementary Figure S9: Overview of CAZyme category GT. Supplementary Figure S10: Overview of CAZyme category PL. Supplementary Table S1: Genome assemblies that were used in this study. Supplementary Table S2: Analysis of putative telomeric repeats at contig ends of CBS6241. Supplementary Table S3: Analysis of unique orthogroups in *F. floriforme*. Supplementary Text S1: Perl script for searching for putative pheromone genes. Supplementary Text S2: Perl script for searching for putative telomeric repeats.

Supplementary material is available at G3 online.

## Acknowledgments

The authors would like to thank Silke Nimtz for excellent technical assistance. MN would like to thank Christopher Grefen for support at the Department of Molecular and Cellular Botany.

## Funding

This work was funded by the German Research Foundation (DFG, grant NO407/7-2 to MN). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

## Conflicts of interest

The authors declare that there is no conflict of interest.

## Literature cited

- Aliyu H, Gorte O, Zhou X, Neumann A, Ochsenreither K. 2020. In silico proteomic analysis provides insights into phylogenomics and plant biomass deconstruction potentials of the Tremellales. *Front Bioeng Biotechnol.* 8:226.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- Bahn YS, Sun S, Heitman J, Lin X. 2020. Microbe profile: *Cryptococcus neoformans* species complex. *Microbiology (Reading).* 166:797–799.
- Banham AH, Asante-Owusu RN, Göttgens B, Thompson S. A. J, Kingsnorth CS, et al. 1995. An N-terminal dimerization domain permits homeodomain proteins to choose compatible partners and initiate sexual development in the mushroom *Coprinus cinereus*. *Plant Cell.* 7:773–783.
- Barredo JL, Garcia-Estrada C, Kosalkova K, Barreiro C. 2017. Biosynthesis of astaxanthin as a main carotenoid in the heterobasidiomycetous yeast *Xanthophyllomyces dendrorhous*. *J Fungi (Basel).* 3:44.
- Bellora N, Moliné M, David-Palma M, Coelho MA, Hittinger CT, et al. 2016. Comparative genomics provides new insights into the diversity, physiology, and sexuality of the only industrially exploited tremellomycete: *Phaffia rhodozyma*. *BMC Genomics.* 17:901.
- Bijlani S, Singh NK, Mason CE, Wang CCC, Venkateswaran K. 2020. Draft genome sequences of Tremellomycetes strains isolated from the International Space Station. *Microbiol Resour Announc.* 9:e00504-20.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–2120.
- Bosetto A, Justo PI, Zanardi B, Venzon SS, Graciano L, et al. 2016. Research progress concerning fungal and bacterial  $\beta$ -xylosidases. *Appl Biochem Biotechnol.* 178:766–795.
- Bracharz F, Beukhout T, Mehlmer N, Brück T. 2017. Opportunities and challenges in the development of *Cutaneotrichosporon oleaginosus* ATCC 20509 as a new cell factory for custom tailored microbial oils. *Microb Cell Fact.* 16:178.
- Branco S, Badouin H, Rodríguez De La Vega RC, Gouzy J, Carpentier F, et al. 2017. Evolutionary strata on young mating-type chromosomes despite the lack of sexual antagonism. *Proc Natl Acad Sci U S A.* 114:7067–7072.
- Branco S, Carpentier F, Rodríguez De La Vega RC, Badouin H, Snirc A, et al. 2018. Multiple convergent supergene evolution events in mating-type chromosomes. *Nat Commun.* 9:2000.
- Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *bioRxiv.* doi:10.1101/2021.1106.1103.446934.
- Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37:D233–D238.
- Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18:188–196.
- Chang HX, Yendrek CR, Caetano-Anolles G, Hartman GL. 2016. Genomic characterization of plant cell wall degrading enzymes and in silico analysis of xylanases and polygalacturonases of *Fusarium virguliforme*. *BMC Microbiol.* 16:147.
- Cho O, Ichikawa T, Kurakado S, Takashima M, Manabe R, et al. 2016. Draft genome sequence of the causative antigen of summer-type

- hypersensitivity pneumonitis, *Trichosporon domesticum* JCM 9580. *Genome Announc.* 4:e00651-16.
- Close D, Ojumu J, Zhang G. 2016. Draft genome sequence of *Cryptococcus terricola* JCM 24523, an oleaginous yeast capable of expressing exogenous DNA. *Genome Announc.* 4:e01238-16.
- Coelho MA, Bakkeren G, Sun S, Hood ME, Giraud T. 2017. Fungal Sex: the Basidiomycota. *Microbiol. Spectr.* 5. <https://doi.org/10.1128/microbiolspec.FUNK-00462016>.
- David-Palma M, Libkind D, Brito PH, Silva M, Bellora N, et al. 2020. The untapped Australasian diversity of astaxanthin-producing yeasts with biotechnological potential—*Phaffia australis* sp. nov. and *Phaffia tasmanica* sp. nov. *Microorganisms.* 8:1651.
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* 7:e1002195.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.
- Findley K, Sun S, Fraser JA, Hsueh YP, Averette AF, et al. 2012. Discovery of a modified tetrapolar sexual cycle in *Cryptococcus amyloletus* and the evolution of MAT in the *Cryptococcus* species complex. *PLoS Genet.* 8:e1002528.
- Fraser JA, Diezmann S, Subaran RL, Allen A, Lengeler KB, et al. 2004. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. *PLoS Biol.* 2:e384.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.
- Grigoriev IV, Cullen D, Goodwin SB, Hibbett D, Jeffries TW, et al. 2011. Fueling the future with fungal genomics. *Mycology.* 2:192–209.
- Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42:D699–D704.
- Han YW, Kajitani R, Morimoto H, Palihati M, Kurokawa Y, et al. 2020. Draft genome sequence of *Naganishia liquefaciens* strain N6, isolated from the Japan Trench. *Microbiol Resour Announc.* 9:e00827-20.
- Huelsenbeck JP, Ronquist F. 2001. Bayesian inference of phylogeny. *Bioinformatics.* 17:754–755.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47:D309–D314.
- Kameshwar AKS, Qin W. 2018. Comparative study of genome-wide plant biomass-degrading CAZymes in white rot, brown rot and soft rot fungi. *Mycology.* 9:93–105.
- Kämper J, Friedrich MW, Kahmann R. 2020. Creating novel specificities in a fungal nonself recognition system by single step homologous recombination events. *New Phytol.* 228:1001–1010.
- Kämper J, Reichmann M, Romeis R, Bölker M, Kahmann R. 1995. Multiallelic recognition: nonself-dependent dimerization of the bE and bW homeodomain proteins in *Ustilago maydis*. *Cell.* 81:73–83.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.
- Kourist R, Bracharz F, Lorenzen J, Kracht ON, Chovatia M, et al. 2015. Genomics and transcriptomics of the oil-accumulating basidiomycete yeast *Trichosporon oleaginosus*: insights into substrate utilization and alternative evolutionary trajectories of fungal mating systems. *mBio.* 6:e00918.
- Kües U, James TY, Heitman J. 2011. Mating type in basidiomycetes: unipolar, bipolar, and tetrapolar patterns of sexuality. In: S Pöggeler, J Wöstemeyer, editors. *The Mycota XIV. Evolution of Fungi and Fungal-Like Organisms*. Berlin, Heidelberg: Springer. p. 97–160.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.
- Kwon-Chung KJ. 2011. *Filobasidium* Olive, 1968. In: CP Kurtzman, JW Fell, T Boekhout, editors. *The Yeasts*. Amsterdam: Elsevier. p. 1457–1465.
- Lam KK, Labutti K, Khalak A, Tse D. 2015. FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. *Bioinformatics.* 31:3207–3209.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9:357–359.
- Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, et al. 2002. Mating-type locus of *Cryptococcus neoformans*: a step in the evolution of sex chromosomes. *Eukaryot Cell.* 1:704–718.
- Lengeler KB, Wang P, Cox GM, Perfect JR, Heitman J. 2000. Identification of the MATa mating-type locus of *Cryptococcus neoformans* reveals a serotype A MATa strain thought to have been extinct. *Proc Natl Acad Sci USA.* 97:14455–14460.
- Liu XZ, Wang QM, Göker M, Groenewald M, Kachalkin AV, et al. 2015a. Towards an integrated phylogenetic classification of the *Tremellomycetes*. *Stud Mycol.* 81:85–147.
- Liu XZ, Wang QM, Theelen B, Groenewald M, Bai FY, et al. 2015b. Phylogeny of tremellomycetous yeasts and related dimorphic and filamentous basidiomycetes reconstructed from multiple gene sequence analyses. *Stud Mycol.* 81:1–26.
- Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, et al. 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science.* 307:1321–1324.
- Mäkelä MR, Donofrio N, De Vries RP. 2014. Plant biomass degradation by fungi. *Fungal Genet Biol.* 72:2–9.
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 38:4647–4654. doi:10.1093/molbev/msab1199.
- Metin B, Findley K, Heitman J. 2010. The mating type locus (MAT) and sexual reproduction of *Cryptococcus heveanensis*: insights into the evolution of sex and sex-determining chromosomal regions in fungi. *PLoS Genet.* 6:e1000961.
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49:D412–D419.
- Nizovoy P, Bellora N, Haridas S, Sun H, Daum C, et al. 2021. Unique genomic traits for cold adaptation in *Naganishia vishniacii*, a polyextremophile yeast isolated from Antarctica. *FEMS Yeast Res.* 21:foaa056.
- Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes *de novo* assembler. *Curr Protoc Bioinformatics.* 70:e102.
- Rodrigues De Miranda L. 1972. *Filobasidium capsuligenum* nov. comb. *Antonie Van Leeuwenhoek.* 38:91–99.
- Ronquist F, Huelsenbeck JP. 2003. Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 19:1572–1574.
- Rytioja J, Hildén K, Yuzon J, Hatakka A, De Vries RP, et al. 2014. Plant-polysaccharide-degrading enzymes from basidiomycetes. *Microbiol Mol Biol Rev.* 78:614–649.
- Sharma R, Gassel S, Steiger S, Xia X, Bauer R, et al. 2015. The genome of the basal agaricomycete *Xanthophyllomyces dendrorhous* provides insights into the organization of its acetyl-CoA derived pathways and the evolution of Agaricomycotina. *BMC Genomics.* 16:233.



- Spatafora JW, Aime MC, Grigoriev IV, Martin F, Stajich JE, et al. 2017. The fungal tree of life: from molecular systematics to genome-scale phylogenies. *Microbiol Spectr.* 5: 10.1128/microbiolspec.FUNK-0053-2016.
- Sriswasdi S, Takashima M, Manabe R, Ohkuma M, Sugita T, et al. 2016. Global deceleration of gene evolution following recent genome hybridizations in fungi. *Genome Res.* 26:1081–1090.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30: 1312–1313.
- Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* 24:637–644.
- Sun S, Coelho MA, David-Palma M, Priest SJ, Heitman J. 2019a. The evolution of sexual reproduction and the mating-type locus: links to pathogenesis of *Cryptococcus* human pathogenic fungi. *Annu Rev Genet.* 53:417–444.
- Sun S, Coelho MA, Heitman J, Nowrousian M. 2019b. Convergent evolution of linked mating-type loci in basidiomycete fungi. *PLoS Genet.* 15:e1008365.
- Takashima M, Manabe R-I, Iwasaki W, Ohyama A, Ohkuma M, et al. 2015. Selection of orthologous genes for construction of a highly resolved phylogenetic tree and clarification of the phylogeny of *Trichosporonales* species. *PLoS One.* 10:e0131217.
- Takashima M, Sriswasdi S, Manabe RI, Ohkuma M, Sugita T, et al. 2018. A *Trichosporonales* genome tree based on 27 haploid and three evolutionary conserved ‘natural’ hybrid genomes. *Yeast.* 35:99–111.
- Tanimura A, Takashima M, Sugita T, Endoh R, Kikukawa M, et al. 2014. *Cryptococcus terricola* is a promising oleaginous yeast for bio-diesel production from starch through consolidated bioprocessing. *Sci Rep.* 4:4776.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876–4882.
- Vajpeyi S, Chandran K. 2016. Draft genome sequence of the oleaginous yeast *Cryptococcus albidus* var. *albidus*. *Genome Announc.* 4:e00390-16.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 9:e112963.
- Yong X, Yan Z, Xu L, Zhou J, Wu X, et al. 2016. Genome sequence of a microbial lipid producing fungus *Cryptococcus albidus* NT2002. *J Biotechnol.* 223:6–7.
- Zhang H, Yohe T, Huang L, Entwistle S, Wu P, et al. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 46:W95–W101.
- Zhao Z, Liu H, Wang C, Xu JR. 2013. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics.* 14:274.

Communicating editor: A. Rokas