
# Functional genome annotation and transcriptome analysis of Pseudozyma hubeiensis BOT-O, an oleaginous yeast that utilizes glucose and xylose at

Fungal Genetics and Biology 166 (2023) 103783

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

# Functional genome annotation and transcriptome analysis of *Pseudozyma hubeiensis* BOT-O, an oleaginous yeast that utilizes glucose and xylose at equal rates

Friederike Mierke [a,b,1], Daniel P. Brink [b,c,1], Joakim Norbeck [b], Verena Siewers [b,*], Thomas Andlid [a]

[a] Food and Nutrition Science, Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden
[b] Systems and Synthetic Biology, Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden
[c] Applied Microbiology, Department of Chemistry, Lund University, Lund, Sweden

ARTICLE INFO

*Keywords:*
*Pseudozyma hubeiensis*
Genome sequencing
RNAseq
Differential gene expression
Nitrogen-starvation
Lipid accumulation

ABSTRACT

*Pseudozyma hubeiensis* is a basidiomycete yeast that has the highly desirable traits for lignocellulose valorisation of being equally efficient at utilization of glucose and xylose, and capable of their co-utilization. The species has previously mainly been studied for its capacity to produce secreted biosurfactants in the form of mannosylerythritol lipids, but it is also an oleaginous species capable of accumulating high levels of triacylglycerol storage lipids during nutrient starvation. In this study, we aimed to further characterize the oleaginous nature of *P. hubeiensis* by evaluating metabolism and gene expression responses during storage lipid formation conditions with glucose or xylose as a carbon source.

The genome of the recently isolated *P. hubeiensis* BOT-O strain was sequenced using MinION long-read sequencing and resulted in the most contiguous *P. hubeiensis* assembly to date with 18.95 Mb in 31 contigs. Using transcriptome data as experimental support, we generated the first mRNA-supported *P. hubeiensis* genome annotation and identified 6540 genes. 80% of the predicted genes were assigned functional annotations based on protein homology to other yeasts. Based on the annotation, key metabolic pathways in BOT-O were reconstructed, including pathways for storage lipids, mannosylerythritol lipids and xylose assimilation. BOT-O was confirmed to consume glucose and xylose at equal rates, but during mixed glucose-xylose cultivation glucose was found to be taken up faster. Differential expression analysis revealed that only a total of 122 genes were significantly differentially expressed at a cut-off of $|\log_2 \text{fold change}| \geq 2$ when comparing cultivation on xylose with glucose, during exponential growth and during nitrogen-starvation. Of these 122 genes, a core-set of 24 genes was identified that were differentially expressed at all time points. Nitrogen-starvation resulted in a larger transcriptional effect, with a total of 1179 genes with significant expression changes at the designated fold change cut-off compared with exponential growth on either glucose or xylose.

## 1. Introduction

Biobased production is foreseen as a key driver in the transition from fossil resources towards sustainable production of chemicals and materials from renewable feedstocks (Scarlat et al., 2015; Soetaert and Vandamme, 2006). Plant-derived triacylglycerols (TAGs), also known as vegetable oils (Durrett et al., 2008), are already used in a wide-range of products, including cosmetics, lubricants, coatings, polymers and production of biodiesel (Belgacem and Gandini, 2008; Pinzi et al., 2009; Tao, 2007). Although vegetable oils are indeed of biological and renewable origin, their production has sustainability issues since the cultivation of the crops directly competes with arable land for food production (Durrett et al., 2008; Luque et al., 2010). Their production can also have detrimental effects on the environment, e.g. in the case of oil-palm cultivations which lead to massive loss of rain-forest (Vijay et al., 2016). An alternative to plant-based lipids is to use fatty acids produced from single cell organisms such as fungi and algae (Blitzblau et al., 2021; Thorpe and Ratledge, 1972). These so-called microbial oils

* Corresponding author.
*E-mail address:* siewers@chalmers.se (V. Siewers).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fgb.2023.103783
Received 6 May 2022; Received in revised form 10 February 2023; Accepted 27 February 2023
Available online 3 March 2023

can have properties similar to plant oils (Jenkins et al., 2015; Ma et al., 2018) and the added benefit that they can be produced from non-edible renewable feedstocks such as lignocellulose (Jin et al., 2015). Lignocellulosic biomass is found in high abundance in agricultural and forestry waste products (Balat and Balat, 2009) and is rich in sugars such as glucose and xylose, up to ~ 60% and ~ 30%, respectively, depending on the origin of the biomass (Foyle et al., 2007; Hamelinck et al., 2005; Templeton et al., 2010). However, efficient utilization of all sugars in the feedstock will be vital in order to establish economically feasible bioprocesses capable of replacing the current fossil-based production (Nogue and Karhumaa, 2015). The ideal fatty acid-producing microbe would therefore need to be able to accumulate high amounts of TAGs while co-consuming glucose and xylose.

Yeasts are important players in large-scale bioprocesses since they generally have a high robustness to harsh process conditions and, unlike many bacteria, are insusceptible to phage infections (Hong and Nielsen, 2012; Mattanovich et al., 2014). Lipid accumulating yeasts are of a particular interest for microbial fatty acid production since there are several yeast species that qualify as oleaginous microbes, which are defined as capable of storing lipids $\geq$ 20% of their cell dry weight (Ratledge and Wynn, 2002; Thorpe and Ratledge, 1972). Accumulation of storage lipids, such as TAGs, is typically induced by cultivation of oleaginous yeasts in the presence of excess carbon levels while limiting the levels of another essential nutrient, such as a nitrogen, phosphorous or sulphur source (Papanikolaou and Aggelis, 2011; Ratledge and Wynn, 2002). Certain species, including *Lipomyces starkeyi*, *Rhodosporidium toruloides* and *Yarrowia lipolytica,* have, when cultivated in this manner, been reported to accumulate up to ~ 60% lipids (Papanikolaou and Aggelis, 2011). In a recent study, where close to 1200 yeast isolates were screened for their ability to co-consume glucose, xylose, and L-arabinose and produce lipids, only 12 oleaginous strains were capable of simultaneous sugar utilization (Tanimura et al., 2016). The most promising isolate was the basidiomycete *Pseudozyma hubeiensis* IPM1-10 that had the ability to utilize hexose and pentose sugars simultaneously, albeit with a preference for glucose over xylose and L-arabinose (Tanimura et al., 2016). These traits thus make *P. hubeiensis* a possible candidate for fatty acid production from sugars found in renewable biomass.

*P. hubeiensis* was first described by Wang and colleagues after isolation from plant leaves in China (Wang et al., 2006). Similar to its close relatives in the *Ustilaginales* order, such as *Moesziomyces antarcticus* (previously *P. antarctica*; Boekhout (1995); Li et al. (2019)), a majority of the research on *P. hubeiensis* has focused on its capacity to produce extracellular glycolipid biosurfactants such as mannosylerythritol lipids (MELs) when grown on various oil substrates (Fukuoka et al., 2007; Kitamoto et al., 1990; Konishi et al., 2008; Morita et al., 2007). *P. hubeiensis* is also able to accumulate high amounts of intracellular storage lipids (24.6 %; Tanimura et al. (2016)), which qualifies it as an oleaginous yeast, a trait that has been somewhat overshadowed in the literature by its MEL production. A few genomes in the *Pseudozyma* genus have been sequenced, including one annotated *P. hubeiensis* genome (strain SY62; Konishi et al. (2013)). Transcriptomics has been performed in *M. antarcticus* and *Moesziomyces aphidis* to investigate the mechanisms underlying the MEL production, but not with regard to the comparison of growth on different sugars (Gunther et al., 2015; Morita et al., 2014; Wada et al., 2020).

In this study, we investigated the *P. hubeiensis* strain BOT-O, which we previously isolated from plant leaves and twigs in a greenhouse in the botanical garden in Gothenburg, Sweden. The strain was initially screened for growth and lipid production on xylose and showed higher lipid production than any other strain isolated from the same environment (Qvirist et al., 2022). The aim of the study was to further characterise the oleaginous nature of *P. hubeiensis* and its ability to naturally co-utilise glucose and xylose, with a focus on physiology and gene expression. Specifically, we wanted to investigate if *P. hubeiensis* BOT-O has comparable lipid accumulation patterns to other oleaginous yeasts, and if the observed equally efficient utilisation of glucose and xylose could be attributed to any specific gene expression patterns. To do this, BOT-O was cultivated on glucose or on xylose, and samples from exponential growth and from lipid accumulation during nitrogen-starvation were analysed using RNA sequencing (RNAseq). In relation to this, we also present the first long-read Nanopore MinION genome sequence of a *Pseudozyma* strain and the first mRNA-guided genome annotation for the species, to our knowledge.

## 2. Materials and methods

### 2.1. Strains and cultivation conditions

The oleaginous yeast strain *P. hubeiensis* BOT-O was previously isolated from the Gothenburg Botanical Garden, Sweden (Qvirist et al., 2022) and was cultivated at 30 °C. The strain was maintained on YPD (20 g/L yeast extract, 10 g/L peptone, 20 g/L glucose) plates with 20 g/L agar. The BOT-O cryostock solution was stored in 25 % (w/w) glycerol at −80 °C. Growth on xylan was assessed after incubation at 30 °C for three days on YP-plates (YPD without glucose) supplemented with 10 g/L of different xylans: beech wood xylan, birch wood xylan and wheat arabinoxylan.

A defined medium consisting of YNB (yeast nitrogen base, without ammonium sulphate) supplemented with glucose and ammonium sulphate to a carbon-to-nitrogen (C/N) ratio of 80 was used to study differential gene expression before and during nitrogen-starvation. The C/N ratio was calculated based on mass, as previously described for *Rhodotorula* yeasts (Braunwald et al., 2013); in short, the g carbon to g nitrogen ratio in the medium was calculated by the assumption that glucose contains 40.0 % carbon (72.06 g carbon • $mol^{-1}$ / 180.16 g glucose • $mol^{-1}$) and ammonium sulphate 21.2% nitrogen (28.02 g nitrogen • $mol^{-1}$ / 132.14 g ammonium sulphate • $mol^{-1}$). The final medium, YNB-CN80, contained 1.7 g/L YNB (without ammonium sulphate), 20 g/L glucose or xylose, or 10 g/L glucose and 10 g/L xylose for the mixed cultivation, 0.47 g/L ammonium sulphate, and a potassium buffer (2.299 g/L $K_2HPO_4$; 11.83 g/l $KH_2PO_4$) at pH 5.5.

Pre-cultures were prepared by inoculating single colonies in 250 mL shake flasks containing 50 mL YPD and incubating overnight (~16 h) at 30 °C in a rotary shaker at 210 rpm. The pre-cultures were harvested by centrifugation, washed with sterile de-ionised water, and used to inoculate 500 mL shake flasks using 120 mL YNB-CN80 medium inoculated to an initial optical density at 600 nm ($OD_{600nm}$) of 0.1. A Genesys 20 spectrophotometer (Thermo Fisher, Waltham, MA, USA) was used for the OD measurements. The shake flasks were incubated at 30 °C and 210 rpm and sampled for differential gene expression at two consecutive time points: at 8 h ($t_1$), corresponding to exponential growth (non-depleted nitrogen levels); and at 25 h ($t_2$), corresponding to nitrogen-depletion. The experiment was repeated for two different carbon sources (glucose and xylose), totalling four conditions: glucose $t_1$ and $t_2$ and xylose $t_1$ and $t_2$. The YNB-CN80 cultivations were performed in biological triplicates. Cultures were sampled for $OD_{600nm}$, cell dry weight (CDW), ammonium levels, pH, total RNA and total lipid content. The CDW was determined in duplicates for all RNAseq time points. 1 mL of each sample was filtered through a pre-weighed 0.45 μm polyethersulfone membrane filter (Sartorius AG, Göttingen, Germany). The filter was then washed with water and dried for 15 min in a microwave oven at medium power. The filters were weighed with a micro scale (Precisa Gravimetrics AG, Dietikon, Switzerland) immediately after drying. Ammonium levels were determined using the Rapid Ammonia Assay Kit (Megazyme, Ireland). RNA, sugar and lipid determinations are described in separate sections below.

### 2.2. Extraction and sequencing of genomic DNA and total RNA

Genomic DNA (gDNA) from an overnight culture of *P. hubeiensis* BOT-O grown in YPD medium was extracted using the Invitrogen Easy-DNA gDNA Purification Kit (Thermo Fisher) by following *Protocol #6* –

*Large Scale Isolation of DNA from Yeast Cells*. The extracted gDNA was quantified with a Qubit v3 fluorometer. The length of the DNA fragments and DNA quality were checked on a 0.5% agarose gel. Long-read sequencing was performed with a MinION sequencer (Mk1B, Oxford Nanopore Technologies), using the Ligation Sequencing Kit (SQK-LSK110, Oxford Nanopore Technologies, Oxford, UK) and a R9.4.1 flow cell. Thirty-three fmol of gDNA was used for sample preparation with the Ligation Sequencing Kit. Complementary short-read sequencing for polishing the long-read assembly was done on an Ilumina MiSeq (2 × 250 bp; 'Version2' chemistry) platform at SciLifeLab National Genomics Infrastructure, Sweden.

Total RNA was extracted from the different YNB-CN80 cultivations described above. A total of 12 samples were processed: two different sugars at two different time points, with three biological replicates per condition. Since the biomass concentration increased throughout the cultivation starting from $OD_{600nm}$ of 0.1, a larger sample volume was collected at the earliest time point to obtain an adequate biomass amount for RNA extraction; 2x 7.5 mL was sampled at $t_1$, and 2x 4 mL at $t_2$, which was estimated to be enough for two technical replicates for the RNA extraction, if needed. Samples were centrifuged immediately (5 min, 3,800 g, 4 °C). The supernatant was quickly discarded, and the samples were quenched in liquid nitrogen and stored at −80 °C until extraction. Cells were disrupted with a Precellys evolution bead mill (Bertin Instruments, Montigny-le-Bretonneux, France) for four cycles that were each run for 45 s at 6,000 rpm. After taking the different biomass concentration in the samples from the different time points into account, it was found that a volume of 7.5 mL of $t_1$ samples ($OD_{600nm}$ ~ 0.6) and 1 mL of the $t_2$ samples ($OD_{600nm}$ ~ 8) resulted in suitable RNA yields. RNA was extracted using the RNeasy Mini Kit (Qiagen, Germany; *Purification of Total RNA from Yeast* protocol). The optional DNase clean-up step suggested by the manual of the kit was performed. RNAzap was used to clear the workspace from any RNases and DNases. The RNA quality was analysed using an Agilent 2100 Bioanalyzer. All samples had a RIN value of > 5 with the exception of the $t_2$ glucose samples with RIN values around 4. The RNA samples were sequenced at SciLifeLab National Genomics Infrastructure, Sweden using the Illumina TruSeq Stranded mRNA library kit with Poly-A selection and sequencing on an Illumina NovaSeq6000 (2 × 150 bp).

### 2.3. Lipid extraction and analysis

The fatty acids were extracted using the KOH/ethanol extraction method described by Andlid et al. (1995), using the slight alterations described in Qvirist et al. (2022). Samples were freeze-dried, and weighed with a micro scale (Precisa Gravimetrics AG, Dietikon, Switzerland). An internal C17:0 TAG standard (triheptadecanoic acid) at a concentration of 4 mg/mL in toluene was added to the dry biomass samples, with volumes adjusted to the CDW of the original sample, with 10 μL for $t_1$ and 50 μL for $t_2$. Then, 2.5 mL of a 2.14 M KOH in 12 % EtOH solution was added and incubated for 2 h at 70 °C using a heat block. To acidify samples to a pH of 2, 1.25 mL of 5 M HCl were added, followed by an extraction with 2 mL of hexane. In an effort to minimize losses in each sample, the extraction step was repeated two more times with 1.5 mL of hexane each. Pooled extracts were evaporated under a flow of nitrogen gas at 40 °C. For methylation, 1 mL of 10% acetyl chloride in methanol and 1 mL toluene were added, followed by another incubation for 2 h at 70 °C. Samples were resuspended in 0.4 mL milliQ and 2 mL petroleum ether / diethyl ether (80:20) and thoroughly mixed. The collected upper phase was then evaporated under a flow of nitrogen gas at 40 °C. Samples were resuspended in 1 mL isooctane before analysis and storage. The resulting fatty acid methyl esters (FAME) were analysed using a GC–MS (Agilent technologies, USA: GC 7890A, MSD 5975C) with a DB-WAX column (0.25x30 mm, 0.25 μm film thickness). The flow was constant at 1 mL/min and helium was used as carrier gas. The fatty acid content was calculated using the internal C17:0 standard.

### 2.4. Chromatographic analysis of glucose, xylose and xylitol

The levels of glucose, xylose and xylitol in the cultivation broth were quantified using high-performance liquid chromatography (HPLC). Samples were prepared by collecting 1 mL culture broth and stored at −20 °C. Before injection onto the column, all samples were filtered through 25 mm nylon filters with a 0.2 μm pore size (VWR International, Radnor, PY, USA) and diluted twice with Milli-Q ultrapure water. The chromatography analysis was performed with a JASCO UV/RI HPLC system (JASCO, Easton, MD, USA) fitted with a Rezex ROA-Organic Acid H+ (8%) column (Phenomenex, Torrance, CA, USA) and a guard column (Phenomenex, Torrance, CA, USA). The system was run at 80 °C at 46 bar with a 5 mM $H_2SO_4$ eluent solution, a 5% methanol wash buffer, an eluent flow rate of 0.8 mL/min, and a 5 μL injection volume. The HPLC was equilibrated for approximately one hour before each run, and samples consisting of only Milli-Q water were injected at the start and at the end of each run. Compounds were detected using the refractive index (RI) detector of the system and quantified using a five-point dilution series of a mixed standard containing glucose, xylose and xylitol. The standard dilutions were injected twice in each run: once at the start, and once at the end. Peaks were analysed with the ChromNAV software (JASCO, Easton, MD, USA).

### 2.5. Genome assembly

The MinION reads from strain BOT-O were *de novo* assembled using the following workflow. Reads were basecalled using Guppy (v4.2.2 + effbaf8; Oxford Nanopore Technologies) with the dna_r9.4.1_450bps_hac.cfg config, and read quality was analysed with FastQC (v0.11.9; Andrews (2010)). Porechop (v0.2.4; Wick et al. (2017)) with *–discard_middle* settings, and Nanofilt (v2.7.1; De Coster et al. (2018)) were used for read trimming and adapter removal. Five different assembly algorithms were assessed for their performance on the BOT-O long-read data: miniasm (v0.3_r179; Li (2016)) with the minimap2 mapper (v2.11; Li (2018)); Canu (v1.5; Koren et al. (2017)); Flye (v2.8.2; Kolmogorov et al. (2019)), Shasta (v0.6.0; Shafin et al. (2020)) and SMARTdenovo (v20180219-5 cc1356; Liu et al. (2021)). SMARTdenovo was run either with its wrapper script (smartdenovo.pl), or by manually calling on its sub-algorithms (wtpre, wtzmo, wtclp, wtlay, wtcns) and specifying input parameters as previously suggested (Jia et al., 2020). Assembly quality was evaluated with Quast (v5.0.2; Mikheenko et al. (2018)). The best assembly from this round (miniasm) was chosen for further polishing. Error-correction using the base-called reads were performed with Racon (v1.4.13; Vaser et al. (2017)) followed by Medaka (v0.5.2; Oxford Nanopore Technologies). A last round of long-read error-correction was done using Nanopolish (v0.13.2; Loman et al. (2015)) with the non-basecalled raw signal (fast5) as indata; the software was run together with bwa (v0.7.17; Li and Durbin (2009)) and SAMtools (v1.10; Danecek et al. (2021); Li et al. (2009)). MUMmer-dnadiff (v4.0.0rc1; Marcais et al. (2018)) was used to evaluate the outcome of each error-correction step. As a final step, Illumina short-read data were used to polish the assembly, which was done using POLCA from the MaSuRCA package (v.3.4.2; Zimin et al. (2013); Zimin and Salzberg (2020)). Contigs shorter than 10 kb were removed from the assembly. BWA (v 0.7.17-r1188; Li and Durbin (2009)) and SAMtools (v1.6; Li et al. (2009)) was used to map the Illumina short-read data to the final, polished assembly to asses read depth across the contigs. Read mapping of the Illumina reads to the assembly was used to assess the uniformity of the coverage and possible aneuploidy across the contigs, as per Borneman et al. (2011).

### 2.6. Phylogenetic analysis

18S phylogenetic analysis was performed by comparing the 18S sequence identified in the BOT-O assembly with select 18S sequences available from NCBI. Sequences from the *Fungal 18S Ribosomal RNA*

*(SSU) RefSeq* project (NCBI accession: PRJNA39195) was used when available for the selected species. The 18S sequences were aligned using MUSCLE (v5.1; Edgar (2004)) and the ends of the resulting Multiple Sequence Alignments were trimmed to equal length with AliView (v1.28; Larsson (2014)). Genome-wide phylogeny was analysed with RealPhy (v1.12; Bertels et al. (2014)) using bowtie2 (v2.3.3.1; Langmead and Salzberg (2012)) and SAMtools (v1.6; Li et al. (2009)). Genome assemblies of a select number of Basidiomycota strains were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genome/) and analysed together with the final polished BOT-O assembly. A total of 24 different Basidiomycota genomes were used in the analysis, and the final number of 16 strains was reached in an iterative manner by removing too similar genomes. The phylogenetic trees were constructed by RAxML (v.8.2.10; Stamatakis (2014)), which calculates phylogeny based on the Maximum Likelihood method. RaxML was run with the GTRGAMMA model and trees underwent 100 bootstrap iterations. The final, bootstrapped trees were visualized with Dendroscope (v3.5.10; Huson and Scornavacca (2012)) and were selected based on a previous study suggesting that phylogram branches that are supported in > 70% of the bootstrap iterations have a 95% probability of representing a true clade (Hillis and Bull, 1993).

### 2.7. RNAseq data processing and transcriptome assembly

The sequence data from the 12 RNAseq samples (4 conditions with 3 biological replicates each) were pre-processed to prepare them for downstream analysis, i.e. transcriptome assembly and differential gene expression analysis. FastQC (v0.11.9; Andrews (2010)) was used to assess read quality and MultiQC (v1.10.1; Ewels et al. (2016)) to overview the results from all 12 samples. The reads were treated with TrimGalore (v0.6.1; https://github.com/FelixKrueger/TrimGalore) running cutadapt (v2.3; Martin (2011)) for quality trimming and adapter removal. The trimmed reads were mapped to the final BOT-O genome assembly with Hisat2 (v2.2.1; Kim et al. (2019)) using the *–dta* option, and the alignments were sorted and indexed with SAMtools (v1.10; Li et al. (2009)). Mapping statistics were assessed with RseQC (v2.6.4; Wang et al. (2012)).

Transcriptomes from each of the 12 samples were assembled using StringTie (v2.1.4; Pertea et al. (2015)) with the Hisat2 alignments as in-data. A non-redundant transcriptome was then generated using all 12 transcriptome assemblies using the StringTie *–merge* option. The final transcriptome was used as transcript-level evidence for the genome annotation pipeline.

### 2.8. Genome annotation

Gene models for the final BOT-O assembly were built using the MAKER pipeline (v3.01.2-beta; Holt and Yandell (2011)). MAKER was run iteratively in three rounds. The initial gene model was built with the following in-data: 1) the final BOT-O assembly; 2) the StringTie-assembled transcripts from BOT-O (4447 transcripts including isoforms) 3) *in silico*-inferred protein sequences from the *P. hubeiensis* SY62 gene model (7472 sequences; Uniprot proteome: UP000014071; Konishi et al. (2013)), and 4) filtered repeat sequences identified in the BOT-O assembly (described below; used to soft-mask the genome during the MAKER runs).

Repeat sequences were *de novo* identified in the final BOT-O assembly in order to reduce false discovery rates caused by e.g. transposon insertions in gene regions. This was done with RepeatModeler (2.0.1; Flynn et al. (2020)) calling on RepeatMasker (v4.1.1) and RMBlast (v2.9.0-p2). RepeatModeler was run with the option *-engine ncbi*, the RepBaseRepeatMaskerEdition-20181026 repetitive DNA elements database (Bao et al., 2015) and the curated portion of Dfam (Hubley et al., 2016) that came packaged with this version of RepeatMasker. The identified repeat sequences were used to mine the *P. hubeiensis* SY62 proteome for transposons and remove any hits before inputting them to MAKER. Mining was done with TransposonPSI (v1.0.0; https://transposonpsi.sourceforge.net) and transposon hits were removed from the proteomes with a small pipeline consisting of fasta_removeSeqFromIDlist.pl (GAAS v1.2.0; https://github.com/NBISweden/GAAS), blastx (v2.2.29+; Altschul et al. (1990)) and ProtExcluder (v1.2; Campbell et al. (2014)).

The MAKER pipeline was run with three different *ab initio* gene predictors: SNAP (v2013_11_29; Korf (2004)), GeneMark (v4.62; Lomsadze et al. (2014)) and Augustus (v3.2.3; Stanke et al. (2006)). Augustus was trained in a Braker2 environment (v2.1.5–20210115-e98b812; Bruna et al. (2021); Hoff et al. (2019)) using an intermediate GeneMark (v4.62; Lomsadze et al. (2014)) model, the option *-fungus*, and the Hisat2 mapped RNAseq data. Diamond (v0.9.31; Buchfink et al. (2015)) was used for the filtering step of Braker2. tRNAs were detected by tRNAscan-SE (v2.0.9; Chan et al. (2021)). Round one of MAKER was run on just the in-data listed above without any *ab initio* models; for round two, SNAP was trained on the MAKER gene model produced by round one, whereas Augustus and GeneMark were trained on the RNAseq data mapped to the complete genome assembly by Hisat2. For the third round, the MAKER model from round 2 was used to train SNAP and was used together with the Augustus and GeneMark models from round 2 as in-data to generate a third iteration MAKER gene model. A schematic overview of the gene prediction workflow can be found in Figure S4 in Supplemental File S1. The gene models from the three iterative rounds were assessed by the total number of predicted genes, AED score (Eilbeck et al., 2009) and BUSCO completeness (v5.0.0; basidiomycota_odb10 (v2020-09–10) database; Seppey et al. (2019)). The model from round two was found to be the best in terms of these metrics and was used for functional annotation.

Proteins were assigned Gene Ontology, Pfam, Superfamily and Interpro annotations using InterproScan (v5.30–69.0; Jones et al. (2014)). Functional annotations were added using a custom pipeline centred around BLASTp (v2.11.0+; Altschul et al. (1990)). In an attempt to harmonize the putative gene annotations with other yeasts by using names from the model yeast *Saccharomyces cerevisiae*, an annotation priority was established where the BOT-O proteins first were queried against *S. cerevisiae* proteins (Uniprot proteome: UP000002311), and remaining unannotated proteins were then in turn queried against Basidiomycota proteins (Uniprot KB query "taxonomy:basidiomycota"). The BOT-O proteins were blasted with local BLASTp databases built from the Uniprot data using makeblastdb; the best hit for each query protein was saved and the resulting list was threshold filtered to only include hits with a blast e-value >=1e-06 and blast score of > 100. AGAT (v0.6.0; Dainat (2021)) was used to apply the filtered best hits to the annotation file. The results of the BLASTp rounds were validated by analysing the protein sequences from BOT-O against *S. cerevisiae* (Uniprot proteome: UP000002311) and non-conventional yeasts *Yarrowia lipolytica* (Uniprot proteome: UP000001300), *Ustilago maydis* (Uniprot proteome: UP000000561) and *Rhodosporidium toruloides* NBRC0880 (Uniprot proteome: UP000239560) with Orthofinder (v 2.5.2; Emms and Kelly (2019)). Based on the results, we decided to continue with the *S. cerevisiae* Orthofinder results. Genes that resulted in the same homologous database hit were manually curated.

BOT-O proteins with homology to known secreted pathogenicity proteins in *U. maydis* were analysed for signal peptides with Signal-P (v5.0; Armenteros et al. (2019)). Analysis of the presence of proteins involved in the infection of plants was done by homology analysis using amino acid sequences from *U. maydis* virulence-related proteins compiled from: Djamei et al. (2011); Doehlemann et al. (2011); Doehlemann et al. (2009); Kämper et al. (2006); Redkar et al. (2015); Wahl et al. (2010). The HMMER web server (https://www.ebi.ac.uk/Tools/hmmer/; Potter et al. (2018)) was used to analyse candidate fatty acid synthase genes from BOT-O and compare them to their homologs in *U. maydis*.

### 2.9. Differential gene expression

The differential expression levels of the predicted BOT-O genes were analysed in a total of four different comparisons: nitrogen depletion on glucose compared to exponential growth on glucose (g2g1); nitrogen depletion on xylose compared to exponential growth on xylose (x2x1); exponential growth on xylose compared to exponential growth on glucose (x1g1); nitrogen depletion on xylose compared to nitrogen depletion on glucose (x2g2). Using the Hisat2 RNA-read mapping results (described in Section 2.7 above), the read counts per gene were quantified using subread-featureCounts (v2.0.0; Liao et al. (2013)). Differential expression was calculated with DESeq2 (v1.26.0; Love et al. (2014)) using Rstudio v1.1.456 running R (v3.6.1, 2019–07-05; R Core Team (2022)). Genes that did not fulfil the condition to have at least 5 reads in at least three of the RNAseq 12 samples (i.e. the four conditions in three biological replicates each) were removed from the analysis. DESeq2′s built-in median-of-ratios normalization was used to normalize the expression counts before the differential expression analysis. For the principal component analysis (PCA) analysis, expression count data were normalized with VST (variance-stabilizing-transformation). Gene dispersion was analysed with *DESeq2::estimateDispersions* and tested with Walds's test (*nbinomWaldTest*). Multiple testing corrections were performed using the Benjamini-Hochberg method for adjustments of p-values, which is the default setting of DESeq2 (Love et al., 2014). Volcano plots of the differential expression data were generated with EnhancedVolcano (v1.4.0; Blighe et al. (2019)).

### 2.10. Metabolic pathway analysis

The KEGG mapper (Kanehisa and Sato, 2020) was used for reconstruction of metabolic pathways. BOT-O genes were assigned K-numbers for KEGG Orthology (KO) using the Assign KO tool (Kanehisa et al., 2016) using *Ustilaginaceae* as a reference dataset. Due to size limitations, the FASTA file containing all genes from the BOT-O assembly had to be split into four separate files, each of which were subsequently used as input for the KO assignment. The resulting K-number files were concatenated and used as input for the pathway mapper. The results were compared to the BOT-O functional annotation master table (Supplemental File S2) and used to create a reconstruction of the central metabolic pathways of *P. hubeiensis*. When gaps in the pathways were found, manual curation of the BOT-O assembly was performed using blastp and protein sequences corresponding to the missing reaction steps from yeast species other than *S. cerevisiae* (mainly other oleaginous yeasts). The overlay feature of the KEGG Mapper was used (Kanehisa and Sato, 2020) to compare the results of the differential gene expression analysis of the two nitrogen-starvation conditions. The list of up- and downregulated genes from the g2g1 and x2x1 comparisons were assigned K-numbers as described above and used as input for KEGG Mapper.

### 2.11. Gene set analysis

Gene set analysis was performed with the Gene Set Enrichment Analysis (GSEA) method (Subramanian et al., 2005) running in Piano (v2.2.0; Varemo et al. (2013)) using the R setup described above. The analysis was performed using the t-values from the four differential expression datasets and the Gene Ontology terms from the functional annotation. The method was run in 15 iterations for each differential expression condition and a consensus result was compiled based on median p-values.

## 3. Results

### 3.1. Long-read genome sequencing and assembly of P. hubeiensis BOT-O

The genome of the recently isolated *P. hubeiensis* BOT-O strain was sequenced and annotated to serve as a foundation for the subsequent RNAseq experiments. The genome sequence was obtained using Oxford Nanopore MinION long-read sequencing and the final assembly was polished with Illumina short-read sequencing data to compensate for the known high error rate of long-read sequencing (Zimin and Salzberg, 2020). A total of 309,260 reads with a mean read length of 28.5 kb were generated during the MinION run, which corresponds to a sequencing coverage of ~ 400x, assuming a *P. hubeiensis* genome size of 18.44 Mb (Konishi et al. (2013). Since genome assemblers have been reported to perform differently well on data from various species (Fournier et al., 2017), five different long-read assemblers were benchmarked for their performance with the BOT-O genome reads (Table S1 in Supplemental File S1). Standard genome assembly metrics (total length, largest contig, N50, N75, L50 and L75) were in general comparable among most assemblers (Table S1 in Supplemental File S1), but two assemblers stood out as the best in terms of assembly completeness, i.e. lowest number of contigs: miniasm with 33 contigs and Flye with 32 contigs. The miniasm assembly was eventually chosen for further polishing based on its higher number of contigs $\geq$ 50,000 bp (31 compared to 23) and lack of uncalled bases (N's), see Table S1 in Supplemental File S1.

The final, polished *P. hubeiensis* BOT-O assembly had a size of 18.95 Mb and a GC content of 56.27% (Table 1), which is comparable to the previously sequenced strain *P. hubeiensis* SY62 that had a size of 18.44 Mb and a 56.50% GC-content (Konishi et al., 2013). However, we report a much higher contiguity: 31 versus 74 contigs for the BOT-O and SY62 assemblies, respectively. No obvious aneuploid regions were detected, after mapping of the Illumina reads to the BOT-O assembly (Figure S1 in Supplemental File S1). Contig completeness was assayed by looking for the occurrence of repeating telomeric motifs, using the assumption that contigs which capture full chromosomes should have several repeat motifs at the 5′ and 3′ ends of contigs. The TTAGGG telomere motif has been found to be evolutionary conserved in basidiomycetes (Guzman and Sanchez, 1994; Ramírez et al., 2011) and several BOT-O contigs indeed had several TTAGGG repeats at one end of the sequence (Table S2 in Supplemental File S1). Two contigs (utg000015l and utg000024l) had multiple TTAGGG repeats in both ends (6–9 repeats at each contig end), indicating that these contigs might each represent a full chromosome. An additional 13 contigs had $\geq$ 5 repeats in only one end.

### 3.2. Establishing the taxonomic relationship of BOT-O to other yeasts by 18S rRNA and genome-wide phylogenetic analyses

The polished BOT-O assembly was analysed for its taxonomic relationship to other yeast species in two complementary ways: by extracting the 18S rRNA sequence from the assembly (Fig. 1A), and by a whole-genome phylogeny approach (Fig. 1B). The BOT-O 18S sequence was used to place *P. hubeiensis* in a larger yeast taxonomy context by comparing it to different Ascomycota and Basidiomycota species (Fig. 1A). The whole-genome phylogeny, which is based on identifying

**Table 1**
Metrics of the final BOT-O Assembly, compared to the previous genome annotation for *P. hubeiensis* SY62.

| *P. hubeiensis* strain | BOT-O | SY62 |
|---|---|---|
| Number of contigs | 31 | 74 |
| Largest contig [bp] | 2,382,866 | 1,398,671 |
| Contigs $\geq$ 50,000 bp | 31 | 53 |
| Total length [Mb] | 18.95 | 18.44 |
| GC content [%] | 56.27 | 56.51 |
| N50 [bp] | 813,625 | 445,580 |
| N75 [bp] | 560,939 | 275,346 |
| L50 | 8 | 13 |
| L75 | 15 | 25 |
| # N's per 100 kbp | 0.00 | 39.88 |

The lengths of all BOT-O contigs are listed in Table S2 in Supplemental File S1.

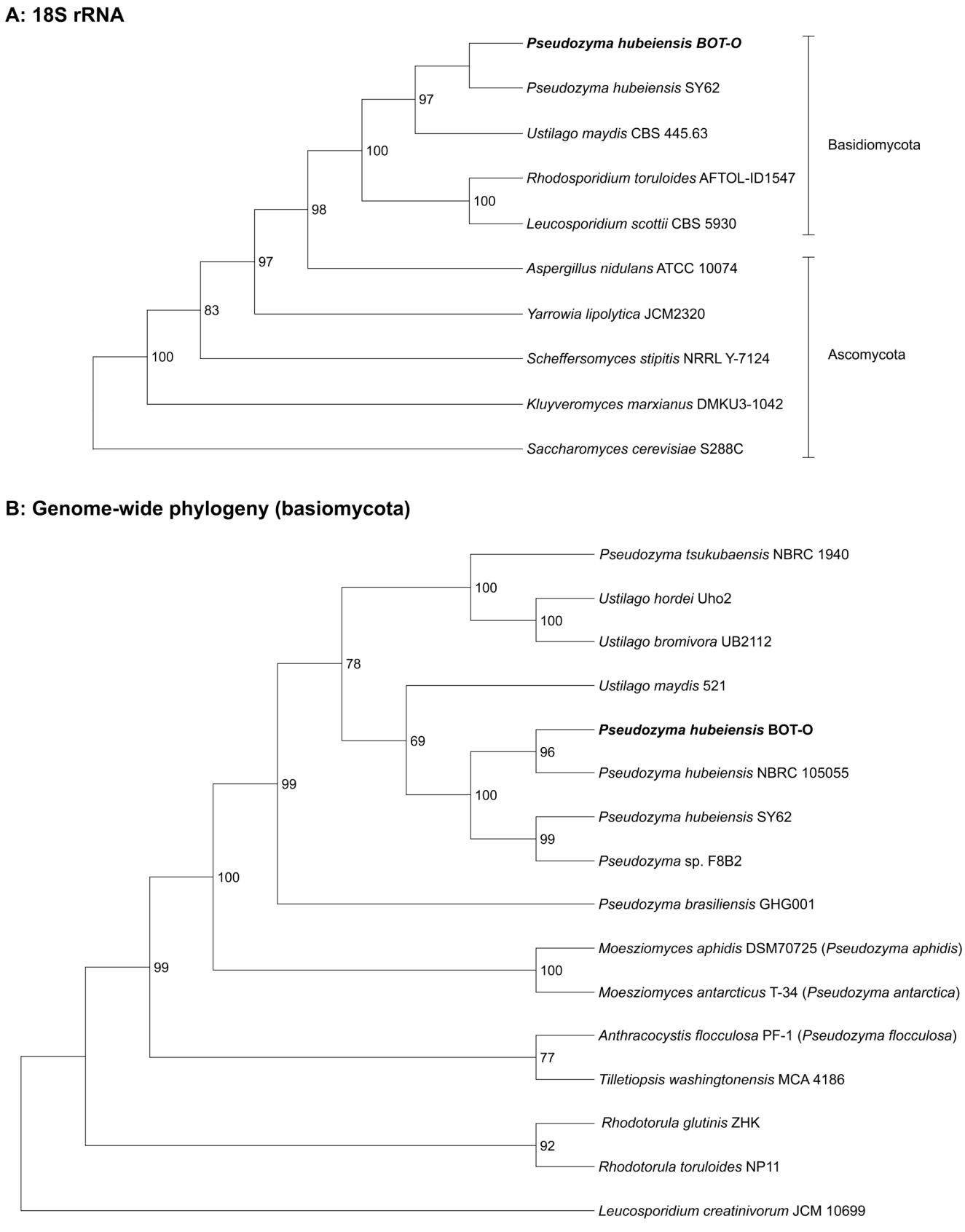



**Fig. 1.** Phylogram of *P. hubeiensis* BOT-O in relation to other basidiomycete yeasts. Two different types of indata were used to establish the phylogeny: A: 18S rRNA sequences from Asco- and Basidiomycota species; B: Whole-genome sequences from Basidiomycota species. The trees were built from whole-genome data using RealPhy and the RAxML Maximum Likelihood method. A total of 100 bootstrap iterations were used to construct the final trees and the numbers at each branch are the bootstrap values.

local polymorphisms in shared loci across genomes (Bertels et al., 2014), was unsuccessful at resolving longer evolutionary distances with acceptable statistical support, and was in the end found to work best when only Basidiomycota genomes were included (Fig. 1B). *P. hubeiensis* has previously been shown to be closely related to the model *Ustilaginales* species *U. maydis* using 28S phylogeny (Kurtzman et al., 2011), and this was corroborated by both our phylogeny approaches (Fig. 1). *U. maydis* was therefore selected for use as a reference genome for *P. hubeiensis* in the downstream bioinformatics analyses when possible.

### 3.3. Physiological characterisation and identification of RNA sequencing conditions

BOT-O was physiologically characterized in shake flask cultures (Fig. 2) to identify relevant time points for collecting RNAseq-samples and to confirm if the strain consumes glucose or xylose at equal rates as has been reported for *P. hubeiensis* IPM1-10 (Tanimura et al., 2016). A commonly used process condition to induce lipid accumulation in oleaginous yeasts is cultivation in nitrogen limitation and carbon excess (Ratledge and Wynn, 2002) and thus defined medium (YNB) with ammonium sulphate as nitrogen source was used. To investigate effect of different sugars on the gene expression, the medium was supplemented with either 20 g/L glucose or 20 g/L xylose at a carbon:nitrogen ratio of 80 g/g (CN80). Cultivation on glucose or xylose resulted in similar growth rates ($\mu_{max}$) and similar sugar consumption rates (0–72 h); the growth rates were 0.23 $h^{-1}$ and 0.21 $h^{-1}$ on glucose and on xylose, respectively, and the sugar consumption rates were 0.20 $g{\cdot}L^{-1}{\cdot}h^{-1}$ and 0.19 $g{\cdot}L^{-1}{\cdot}h^{-1}$ on glucose and on xylose, respectively (Fig. 2A and 2B). In both conditions, the sugars were not depleted at the end of the cultivation (72 h) with ~ 3.9 g/L glucose, and ~ 5.6 g/L xylose remaining. No xylitol was detected in the samples from the xylose cultivations (data not shown).

For each sugar, samples for RNAseq analysis and determination of lipid content were taken at 8 h at which the cells were growing exponentially ($t_1$; time point 1), and at 25 h ($t_2$; time point 2) at which time the nitrogen was depleted for either cultivation, as determined by an enzymatic kit (Fig. 2C). The total lipids measured at the different time points further confirmed that the nitrogen-depleted samples at $t_2$ were representative of lipid accumulation, reaching levels of around 300 mg lipids/g CDW (Fig. 2D). Lipid composition profiles can be found in Figure S2 in Supplemental File S1. There was no significant difference in the amount of total lipids between the glucose and xylose cultivations at the chosen time points (Fig. 2D; Student's *t*-test). Intracellular storage lipids were also visible with light microscopy (Figure S3 in Supplemental File S1). In total, 12 samples were taken for RNAseq analysis (two different sugars at two time points, in three biological replicates).

### 3.4. Genome annotation and reconstruction of key metabolic pathways

#### 3.4.1. Gene prediction and functional annotation

Before differential gene expression analysis could be performed, the RNAseq data was used to predict and annotate open reading frames (ORFs) in the BOT-O genome assembly with the MAKER pipeline (Holt and Yandell, 2011). The reads from the 12 RNAseq samples were used to generate a transcriptome assembly for use as mRNA evidence in the prediction pipeline. The assembly resulted in 4447 transcripts, including isoforms. Repeat regions in the genome assembly were identified and masked, as they can introduce bias during the gene prediction (Cantarel

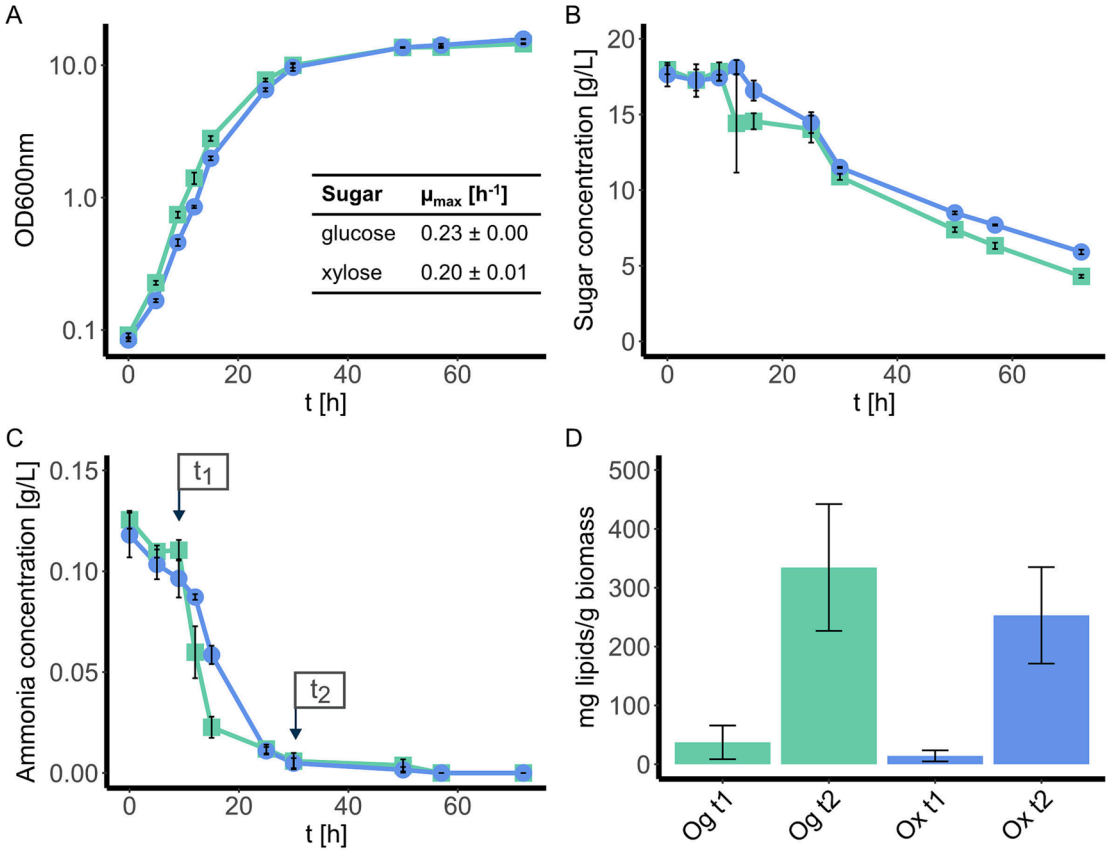



**Fig. 2.** Shake flask cultivation of BOT-O on glucose and xylose and RNAseq sampling time points. A: Growth measured as $OD_{600nm}$. B: Sugar concentration for glucose or xylose. C: Ammonia concentration. D: Lipid concentration in mg lipids/g CDW at $t_1$ and $t_2$ for both growth on glucose and on xylose. All experiments were performed in biological triplicates and error bars represent the standard deviations. Samples for RNAseq were taken at two time points with $t_1$ = time point 1 during exponential growth, and $t_2$ = time point 2 during nitrogen-starvation phase. Green and ■ = BOT-O on glucose, Blue and ● = BOT-O on xylose. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2008). Seven recurring repeat-rich regions were identified in the BOT-O assembly. The MAKER pipeline was trained and run in an iterative manner (Figure S4 in Supplemental File S1). The resulting gene models were assessed by three metrics: 1) the number of predicted genes and their average length; 2) the Annotation Edit Distance (AED), which is a gene prediction quality score measuring how well a given gene prediction is supported by the supplied evidence, e.g. mRNA and protein data (Eilbeck et al., 2009); and 3) the gene model completeness in terms of evolutionary conserved genes, so called BUSCOs (Seppey et al., 2019). All metrics improved from the first to the second round (Table 2; Figure S5 in Supplemental File S1) which shows the importance of training the subsequent iterations with the gene models produced by the previous iteration of the pipeline. The third iteration resulted in very similar results as the second, and to not risk generating an over-predicted model, the second iteration was therefore chosen as the final gene model. This model contained 6525 mRNA genes and 164 tRNA genes, and 97.8% of the 1764 Basidiomycota BUSCOs could be identified among the predicted genes (Table 2). A total of 3301 introns were discovered, with an average of 0.5 introns per gene (Table S3 in Supplemental File S1). For comparison, we also subjected the *P. hubeiensis* SY62 genome assembly (Konishi et al., 2013) to the same analyses and found that the final BOT-O gene model had fewer ORFs than SY62 (6525 and 7472, respectively) and higher BUSCO completeness (97.8% vs 95.3%: Table 2).

The final gene model was functionally annotated based on homology of the predicted BOT-O protein sequences to proteins of previously annotated yeasts, using a custom automated annotation pipeline running BLASTp. To harmonize the gene names to standard yeast nomenclature, an annotation decision hierarchy was implemented where BOT-O proteins were first annotated by homology to the *Saccharomyces cerevisiae* reference genome (Engel et al., 2014). Proteins that had no significant homology hits in *S. cerevisiae* were subsequently annotated based on homology to all proteins listed in the Basidiomycota taxonomy in the Uniprot database (UniProt Consortium, 2019). A total of 5260 (or 79%) of the predicted ORFs were successfully assigned functional annotations, including 164 tRNA loci. The remaining 1429 genes were annotated as hypothetical proteins (Supplemental File S2).

### 3.4.2. Reconstruction of relevant metabolic pathways in P. hubeiensis BOT-O

Based on the functional annotation, we were able to reconstruct several central metabolic pathways in *P. hubeiensis* BOT-O in full (Fig. 3). Genes involved in glycolysis, pentose phosphate pathway (PPP) and TCA cycle were clearly identified (see Sheet "2. Pathway assignment" in Supplemental File S2). The genes involved in the well-studied pathway for production of MELs (Konishi et al., 2008; Morita et al., 2007; Rau et al., 2005) were identified based on homology to the MEL pathway in *U. maydis* (Hewald et al., 2006) and *P. tsukubaensis* (Saika et al., 2016) (Fig. 3). While the degree of homology to the best BOT-O candidate proteins varied between each MEL enzyme, the corresponding candidate genes were located back-to-back in a gene cluster on contig utg0000211 (PHBOTO_003088-92), which is also how the MEL pathway genes are organized in *U. maydis* (Hewald et al., 2006).

Candidate genes for the fungal xylose reductase/xylitol dehydrogenase (XR/XDH) pathway (Fig. 3) were identified by manual curation. Xylulokinase (XKS) is part of the PPP, and thus a BOT-O XKS (PHBOTO_001564) could be identified based on homology to the *S. cerevisiae* gene. The XR and XDH had to be identified by comparison to proteins from other xylose-utilizing yeast species, since *S. cerevisiae* cannot naturally utilize xylose (Hahn-Hägerdal et al., 2007). Three putative XRs (PHBOTO_000425, PHBOTO_002476, PHBOTO_004254) and four potential XDHs (PHBOTO_000427, PHBOTO_003491, PHBOTO_005823, PHBOTO_006563) were found when using the commonly studied Xyl1p and Xyl2p enzymes from *Scheffersomyces stipitis* (Uniprot: P31867 and P22144), but no single best candidate could be discriminated based on the results. Eight candidate XRs have been identified in *Ustilago bevomyces* (Lee et al., 2016), a species from a genus much closer related to *Pseudozyma* than *Scheffersomyces* (Fig. 1). Protein homology to the single xylose dehydrogenase proposed in *U. bevomyces* (Uniprot: A0A0B4ZXX9; Lee et al. (2016)) suggested that PHBOTO_003491 encodes a XDH. Homology to UbXR1 (Uniprot: A0A0B5A1B2), which has been verified to have a strictly NADPH-dependent XR activity (Lee et al., 2016), suggested that PHBOTO_004254 encodes an XR. UbXR3 has been demonstrated to have dual preference for xylose and arabinose and proposed to catalyse the first step of the arabinose assimilation (Lee et al., 2016); it had a homolog in PHBOTO_005965. BOT-O has been verified to grow on arabinose (data not shown).

A common trait among oleaginous yeasts is the presence of the ATP-citrate lyase (ACL) enzyme, and many non-oleaginous yeasts, such as *S. cerevisiae,* lack the *ACL1* gene (Boulton and Ratledge, 1981; Ratledge and Wynn, 2002). In BOT-O, the PHBOTO_001285 locus was identified as a likely *ACL1* gene based on homology to *R. toruloides*. Other key genes related to the oleaginous phenotype, such as *ACC1*, *CTP1* and *FAS* were also identified in BOT-O (Fig. 3). The fatty acid synthase complex has been shown to differ between different fungal species and is commonly distributed across two subunits, for instance in *S. cerevisiae* (Zhu et al., 2017). The established reference strain *U. maydis* 521 has been shown to have a single-chain *FAS* gene (Teichmann et al., 2007; Wernig et al., 2020), and upon further investigation we saw that *U. maydis* has two very similar genes encoding for *FAS*: UMAG_10339, a putative 3-hydroxyacyl-[acyl-carrier-protein] dehydratase that has been shown to have the FAS function (Uniprot: A0A0D1C5S0; Wernig et al. (2020)), and UMAG_06460 (Uniprot: A0A0D1DNX1; Teichmann et al. (2007)), which is listed as *FAS2* in Uniprot. These two *U. maydis FAS* genes were homologous to PHBOTO_002928 and PHBOTO_006517

**Table 2**
Assessment of genome assembly and annotation completeness by analysis of evolutionary conserved Basidiomycota genes, so-called BUSCOs (Benchmarking Universal Single-Copy Orthologs).

| Gene model | Number of predicted ORFs (excluding tRNA) | Average length of the predicted CDSs (bp) | Complete BUSCOs | Fragmented BUSCOs | Missing BUSCOs |
|---|---|---|---|---|---|
| BOT-O assembly (before annotation) | – | – | 1738 (98.5%) | 9 (0.5%) | 17 (1.0%) |
| BOT-O annotation, round 01 | 6392 | 1759.92 | 1601 (90.7%) | 31 (1.8%) | 132 (7.5%) |
| BOT-O annotation, round 02 (Final model) | 6525 | 1811.46 | 1725 (97.8%) | 12 (0.7%) | 27 (1.5%) |
| BOT-O annotation, round 03 | 6493 | 1819.63 | 1728 (97.9%) | 9 (0.5%) | 27 (1.6%) |
| *P. hubeiensis* SY62 annotation (reference genome) | 7472 | 1668.42 | 1681 (95.3%) | 33 (1.9%) | 50 (2.8%) |

BOT-O annotations were performed in an iterative manner using the MAKER pipeline. The "BOT-O assembly" contains all the genomic nucleotides, whereas the BOT-O and SY62 annotation rounds only include the nucleotides from the predicted Open Reading Frames (ORFs). The average length per gene was calculated using the coding sequences (CDSs), i.e. introns were excluded. BUSCO percentages were calculated based on the total 1764 genes of the basidiomycota_odb10 (2020–09-10) database. The gene model from BOT-O annotation round 02 was chosen for all downstream analysis, based on this benchmarking and on the AED score plots (Figure S5 in Supplemental File S1).

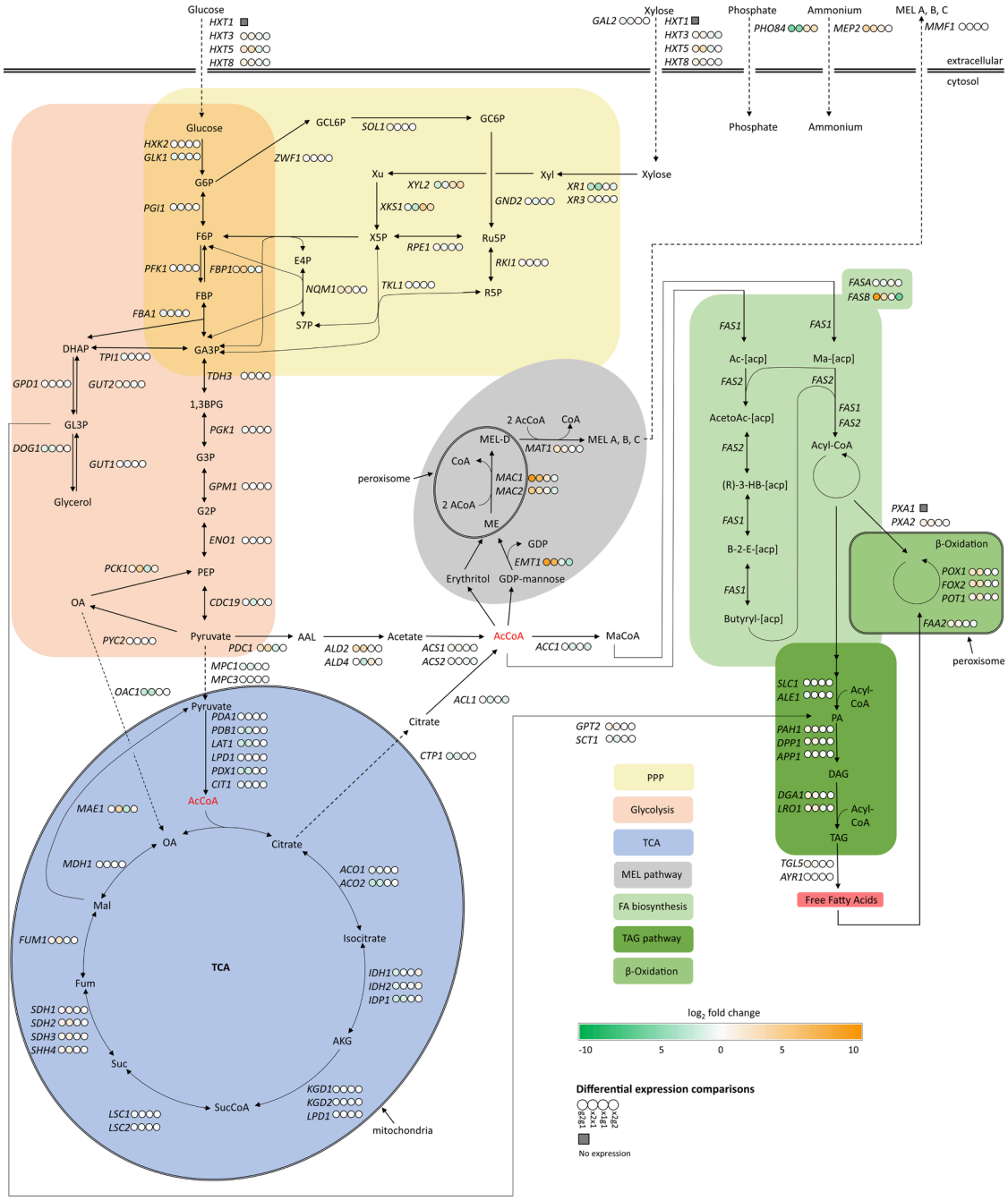



**Fig. 3.** Reconstruction of main metabolic pathways in *P. hubeiensis* BOT-O based on the functional annotation of the genome, and differential expression of the involved genes. The network, reconstructed using KEGG, the *Saccharomyces* Genome Database (SGD) and BLASTp, includes glycolysis (orange), the pentose phosphate pathway (yellow), the TCA cycle (blue), lipid biosynthesis, triacylglycerol biosynthesis and β-oxidation (all different shades of green), as well as MEL biosynthesis (grey) and specific transport processes. Differential gene expression data for each gene in all four analyzed comparisons are presented in the circles next to the gene names: g2g1 - nitrogen-starvation versus exponential growth on glucose; x2x1 - nitrogen-starvation versus exponential growth on xylose; x1g1 - exponential growth on xylose compared to glucose; x2g2 - nitrogen-starvation on xylose compared to glucose. Compound abbreviations: G6P: D-glucose-6P, F6P: D-fructose-6P, FBP: D-fructose-1,6$P_2$, DHAP: dihydroxyacetone phosphate, GL3P: glycerol-3P, GA3P: D-glyceraldehyde-3P, 1,3BPG: glycerate-1,3$P_2$, G3P: glycerate-3P, G2P: glycerate-2P, PEP: phosphoenolpyruvate, OA: oxaloacetate, AKG: α-ketoglutarate, SucCoA: succinyl-CoA, Suc: succinate, Fum: fumarate, Mal: (S)-malate, AAL: acetaldehyde, AcCoA: acetyl-CoA, MaCoA: malonyl-CoA, GCL6P: D-glucono-1,5-lactone-6P, GC6P: D-gluconate-6P, Xyl: xylitol, Xu: D-xylulose, Ru5P: D-ribulose-5P, R5P: D-ribose-5P, X5P: D-xylulose-5P, E4P: D-erythrose-4P, S7P: D-sedo-heptatose-7P, ME: mannosylerythritol, MEL: mannosylerythritol lipid, Ac-[acp]: acetyl-[acp], AcetoAc-[acp]: acetoacetyl-[acp], (R)-3-HB-[acp]: (R)-3-hydroxybutanoyl-[acp], B-2-E-[acp]: but-2-enoyl-[acp], PA: phosphaditic acid, DAG: diacylglycerol, TAG: triacylglycerol. Gene name abbreviations are according to SGD or other oleaginous yeasts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Figure S6 in Supplemental File 1), according to Blast and HMMR analyses. Based on the multitude of shared domains (Figure S6 Supplemental File S1) and how the closely related *U. maydis* has two *FAS* genes, we propose that both of these BOT-O genes encode a fatty acid synthase and thus name them *FASA* (PHBOTO_002928) and *FASB* (PHBOTO_006517). The putative enzymes required for production and degradation of the TAG storage lipids were also identified (Fig. 3), including the diacylglycerol O-acyltransferase (encoded by *DGA1*) and Lro1p acyltransferase known to be essential for TAG formation (Oelkers et al., 2002), as well as the β-oxidation genes (*PXA1/2*, *FAA2*, *POT1*, *FOX2*, *POX1)*.

The close taxonomic relationship to the plant pathogen *U. maydis* (Fig. 1) also prompted us to examine whether *P. hubeiensis* BOT-O contains homologs to *U. maydis* genes for infection and tumor formation on plant tissue. *U. maydis* and related smut fungi infect their plant hosts by means of secretion of so-called effector proteins (Lanver et al., 2017) and it has been shown that genes are evolutionary conserved among related plant-infecting smut fungi, including *Pseudozyma* species (Benevenuto et al., 2018; Schirawski et al., 2010; Sharma et al., 2019). Indeed, 42 of 55 assessed *U. maydis* virulence-related proteins had homologs in BOT-O, including 18 secreted hydrolases for breakdown of plant cell-walls (Supplemental File S3). 33 of the BOT-O homologs had a secretion signal peptide (Supplemental File S3), which is a typical trait for *U. maydis* plant infection-related proteins (Lanver et al., 2017). As a control experiment, we also compared the *U. maydis* proteins to the non-plant pathogenic yeast *S. cerevisiae*. There were only very few potential homologs in *S. cerevisiae* (10 of 55), and these mostly included glucosidases and sucrose transporters, but no virulence factors (Supplemental File S3).

### 3.5. Differential gene expression

Expression levels for each predicted ORF were calculated using the RNAseq reads from the four conditions: exponential growth on glucose or xylose ($gt_1$ and $xt_1$), and nitrogen-starvation on glucose or xylose ($gt_2$ and $xt_2$). A total of 690 mRNA genes and 122 tRNA genes with low read mapping, i.e. genes that did not fulfil the cut-off criteria of $> 5$ reads in $\geq 3$ of the 12 samples, were omitted from the analysis, which left 5880 genes to be analysed. Notably, 616 of the 812 omitted genes (76%) had zero read counts in all samples (Supplemental File S2), suggesting that they were either not expressed in our samples, or that they were incorrectly predicted ORFs that lack biological activity. The reproducibility of the biological replicates was assessed with a principal component analysis (PCA), which confirmed the quality of the experimental design (cultivation, RNA extraction and sequencing) since the biological triplicates of the four conditions clearly clustered separately from each other (Figure S7 in Supplemental File S1).

The differential gene expression analysis was performed in four different comparisons: nitrogen-starvation versus exponential growth on glucose (g2g1) and on xylose (x2x1); exponential growth on xylose compared to glucose (x1g1); and nitrogen-starvation on xylose compared to glucose (x2g2). A total 1207 different genes were found to change their expression at the designated cut-off used in this study (adjusted p-value $< 10^{-6}$ and $|\log_2$ fold change$| \geq 2$) across all four conditions (Fig. 4). One single gene (PHBOTO_000828), annotated as inorganic phosphate transporter *PHO84,* fulfilled the differential expression cut-off in all four comparisons (Fig. 4), being downregulated during exponential growth and upregulated during nitrogen-starvation on both sugars. While many genes were differentially expressed during nitrogen-starvation compared to exponential growth on either of the sugars (g2g1 and x2x1; Fig. 5A-B), only few differentially expressed genes were detected when comparing glucose and xylose at either time point (x1g1 and x2g2; Fig. 5C-D). In addition to analysing the differentially expressed ORFs at a single-gene level, Gene Ontology (GO) term enrichments in the four comparisons (g2g1, x2x1, x1g1, x2g2) were assessed using gene set analysis (Figs. 6-7).

### 3.6. Transcriptional changes between glucose and xylose during exponential growth and nitrogen-starvation

Since BOT-O grew equally well on glucose and xylose with regard to growth rate (Fig. 2), we hypothesised that only minor transcriptional changes would be seen between the sugars, since such changes are often linked to changes in growth rate (Regenberg et al., 2006). Indeed, only 73 genes fulfilled the designated differential expression cut-off when comparing xylose and glucose during exponential growth (x1g1;

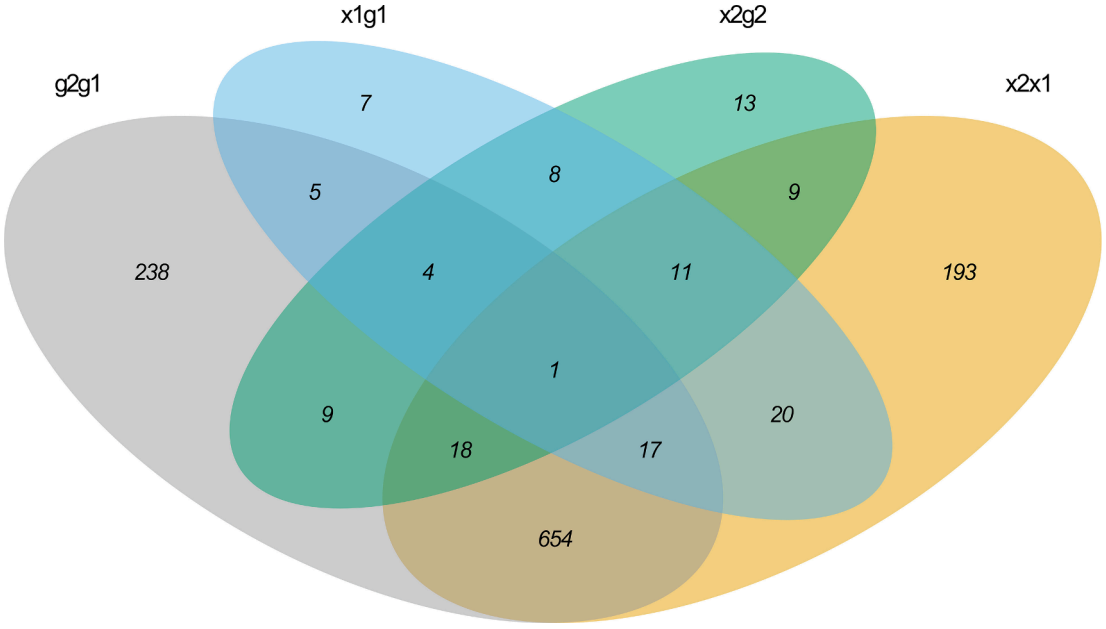



**Fig. 4.** Venn diagram illustrating the 1207 genes found to be differentially expressed at the designated cut-off across all four comparisons. g2g1 - differential gene expression during nitrogen starvation on glucose; x2x1 - differential gene expression during nitrogen starvation on xylose; x1g1 - differential gene expression during exponential growth on xylose compared to glucose; x2g2 - differential gene expression during nitrogen-starvation on xylose compared to glucose. The single gene found to be differentially expressed in all conditions was *PHO84* (PHBOTO_000828).

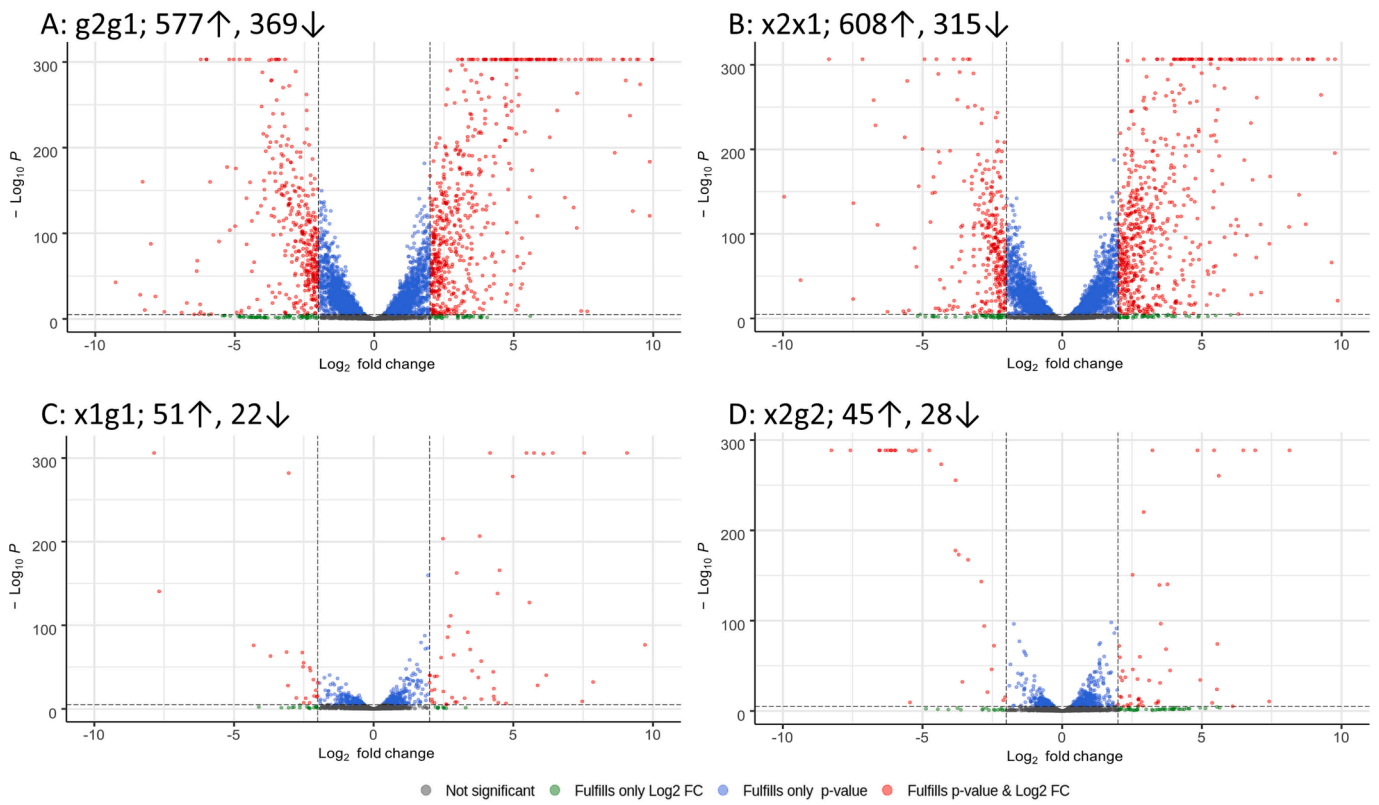



**Fig. 5.** Volcano plots displaying all differentially expressed genes for all 4 conditions: A: differential gene expression during nitrogen-starvation on glucose (g2g1); B: differential gene expression during nitrogen-starvation on xylose (x2x1); C: differential gene expression during exponential growth on xylose compared to glucose (x1g1); D: differential gene expression during nitrogen-starvation on xylose compared to glucose (x2g2). The number of up- and downregulated genes in the different conditions are listed with arrows representing the direction of regulation. The following differential expression significance cut-offs were used throughout the study: p-value $< 10^{-6}$ (plotted as the dashed horizontal line as $-\log_{10}(p) > 5$) and $|\log_2 \text{fold change}| \geq 2$ (vertical dashed lines in the figure).

Supplemental File S2). Most of these, 51 of 73, were upregulated on xylose, and included xylose pathway genes such as *XKS1* and *XYL2*. Thirteen of the 73 genes encoded hypothetical proteins, some of which were successfully annotated with GO terms (Sheet "5. DE x1_vs_g1" in Supplemental file S2). When comparing glucose and xylose at time point 2 (nitrogen-starvation; x2g2), the same number of genes (73) fulfilled the designated differential expression cut-off when comparing the two sugars, but not necessarily the same genes as at time point 1 (x1g1). When compiling the x1g1 and x2g2 datasets, a total of 122 different genes displayed changes in expression, out of which 24 genes were common to both comparisons (Table 3).

We hypothesize that these 24 genes represent a core set of *P. hubeiensis* genes that are highly differentially expressed between growth on xylose and glucose. Most of these 24 genes (87.5%) were upregulated on xylose during both exponential growth (x1g1) and nitrogen-starvation (x2g2) and several were related to sugar uptake and degradation. These included two highly upregulated putative β-xylanases (PHBOTO_003675 and PHBOTO_006039), which implies that xylan-degradation in BOT-O is co-regulated with other xylose metabolic genes (*XKS1* and *XYL2*). Several *Pseudozyma* species have been reported to secrete xylanases (Adsul et al., 2009; Borges et al., 2014; Faria et al., 2015) and we therefore assessed the capacity of BOT-O to grow on plates containing xylan as the sole carbon source. It was found that xylans derived from various types of plants indeed were able to support growth (Figure S8 in Supplemental File S1).

Other genes in this xylose core set were related to uptake and utilization of other carbon sources than xylose, including homologs related to galactose transport and metabolism (*GAL2* and *GAL10*), glycerol catabolism (*GCY1*), hexose transporters (*HXT8/13*), a mannose-6-phosphate isomerase (PHBOTO_006511), a cutinase (PHBOTO_005903) and a carboxylic ester hydrolase (PHBOTO_004937). The upregulation of these genes during xylose cultivation (x1g1 and x2g2) suggests that growth on xylose as a carbon source is perceived by the cell as being less optimal than growth on glucose, in the sense that the cell seems to be expressing genes related to uptake and metabolism of a range of non-glucose carbon sources. The putative *MAL31* maltose permease gene (PHBOTO_003640) was downregulated on xylose during exponential growth (x1g1) but upregulated during nitrogen starvation (x2g2), and the putative mitochondrial D-lactate dehydrogenase gene *DLD1* (PHBOTO_001232) was upregulated during exponential growth and downregulated during nitrogen starvation (Table 3), which implies that their function during xylose assimilation is dependent on exponential growth and nutrient-starvation respectively. Only one gene, the putative drug/proton antiporter gene *YHK8* (PHBOTO_004167), was downregulated in both comparisons (exponential growth and nitrogen starvation; Table 3), suggesting that it has a role in assimilation of glucose but not xylose.

The indications from the proposed xylose core gene set (Table 3) that growth on glucose might be the preferred carbon source prompted us to perform an additional round of physiological characterization using mixed sugars, with 10 g/L each of glucose and xylose (Fig. 8). While the maximum growth rate (0.24 $h^{-1}$), final OD and nitrogen-depletion time (25 h) were comparable to the previous experiment on single sugars (Fig. 2), the strain clearly preferred glucose over xylose (Fig. 8), with sugar consumption rates (0–72 h) of 0.14 $gL^{-1}h^{-1}$ for glucose and 0.08 $gL^{-1}h^{-1}$ for xylose. However, both sugars were used simultaneously, despite the difference in consumption rate (Fig. 8B), and the total sugar consumption rate (0.22 $gL^{-1}h^{-1}$) was similar to the single sugar cultivations (Fig. 2).

The differentially expressed sugar comparison gene sets (x1g1 and

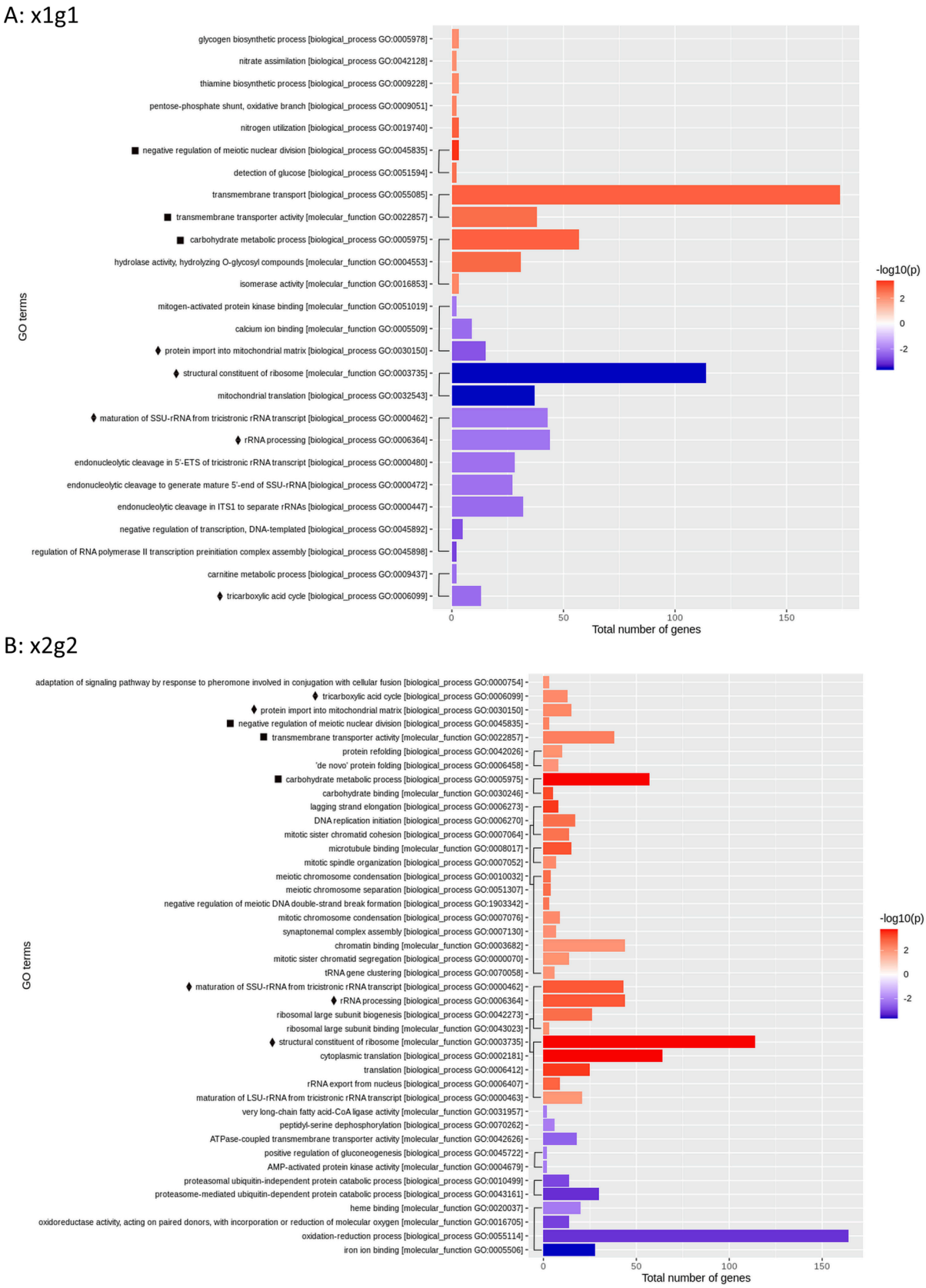



**Fig. 6.** Gene set analysis for the sugar comparison during exponential growth or nitrogen starvation. A: differential gene expression during exponential growth on xylose compared to glucose (x1g1), and B: differential gene expression during nitrogen-starvation on xylose compared to glucose (x2g2). Squares show GO terms shared between both conditions and the same direction of enrichment (up-regulation or down-regulation); diamonds show GO terms shared between both conditions with directional changes in enrichment. A p-value cut-off of $< 0.009$ was used to select the final data.

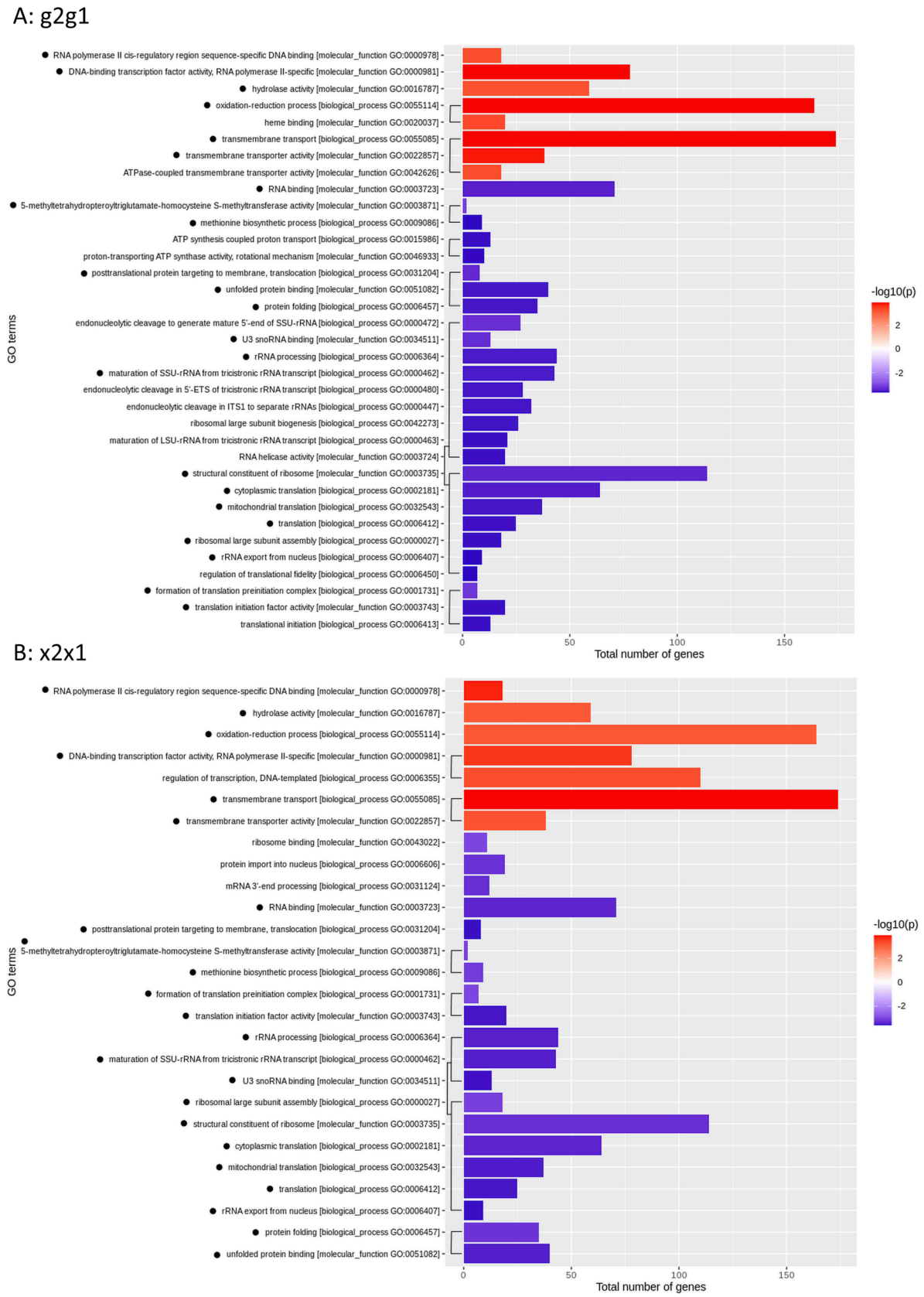



**Fig. 7.** Gene set analysis for the nitrogen comparison on glucose or xylose. A: differential gene expression during nitrogen-starvation on glucose (g2g1), and B: differential gene expression during nitrogen-starvation on xylose (x2x1). Circles show GO terms shared between both conditions and the same direction of enrichment (up-regulation or down-regulation). A p-value cut-off of $< 0.001$ was used to select the final data.

**Table 3**
Core set of genes found to fulfill the designated differential expression cut-off (adjusted p-value $< 10^{-6}$ and $|\log_2$ fold change$| \geq 2$) in both sugar comparisons: exponential growth on either glucose or xylose (x1g1) and nitrogen-starved growth on either glucose or xylose (x2g2). A more detailed version of this list can be found in Sheet "8. Sugars core set" in Supplemental File S2.

| Locus name | Suggested annotation | Gene name | $\log_2$ fold change x1g1 | $\log_2$ fold change x2g2 |
|---|---|---|---|---|
| PHBOTO_000828 | Inorganic phosphate transporter | *PHO84* | 2.18 | 2.53 |
| PHBOTO_001232 | D-Lactate dehydrogenase [cytochrome] 1, mitochondrial | *DLD1* | 2.69 | −2.44 |
| PHBOTO_001564 | Xylulose kinase | *XKS1* | 4.51 | 2.78 |
| PHBOTO_001594 | Hexose transporter | *HXT8* | 4.17 | 2.52 |
| PHBOTO_001681 | Hypothetical protein | – | 4.98 | 3.77 |
| PHBOTO_002497 | Zn(2)-C6 fungal-type domain-containing protein | – | 4.30 | 2.84 |
| PHBOTO_002755 | Polygalacturonase | *PGU1* | 3.37 | 5.61 |
| PHBOTO_003056 | Glycerol 2-dehydrogenase (NADP(+)) | *GCY1* | 5.47 | 4.84 |
| PHBOTO_003491 | D-Xylulose reductase | *XYL2* | 2.48 | 3.23 |
| PHBOTO_003550 | Galactose transporter | *GAL2* | 3.85 | 3.72 |
| PHBOTO_003630 | Carboxylic ester hydrolase | – | 3.85 | 5.54 |
| PHBOTO_003640 | Maltose permease | *MAL31* | −3.69 | 2.74 |
| PHBOTO_003675 | β-Xylanase | – | 5.87 | 10.01 |
| PHBOTO_004167 | Probable drug/proton antiporter | *YHK8* | −7.86 | −2.68 |
| PHBOTO_004810 | Sugar transporter | *STL1* | 5.58 | 2.70 |
| PHBOTO_004937 | Carboxylic ester hydrolase | – | 2.76 | 5.44 |
| PHBOTO_004942 | Non-reducing end α-L-arabinofuranosidase | – | 2.41 | 6.48 |
| PHBOTO_005858 | Bifunctional protein | *GAL10* | 2.86 | 5.56 |
| PHBOTO_005903 | Cutinase | – | 2.18 | 4.95 |
| PHBOTO_005904 | Hypothetical protein | – | 2.49 | 3.87 |
| PHBOTO_005943 | Glycoside hydrolase | – | 6.08 | 3.48 |
| PHBOTO_005944 | Low glucose sensor | *SNF3* | 9.07 | 6.91 |
| PHBOTO_006039 | Endo-1,4-β-xylanase | – | 7.54 | 8.14 |
| PHBOTO_006511 | Mannose-6-phosphate isomerase | – | 9.71 | 3.44 |

x2g2) were also analysed for their enrichments of GO terms (Fig. 6). Overall, the x1g1 gene set analysis revealed many GO terms related to nutrient transport and metabolism (Fig. 6A). Eight out of 60 total significant GO terms were shared between exponential growth (x1g1; Fig. 6A) and nitrogen-starvation (x1g1; Fig. 6B). Five of the eight terms changed their direction from being enriched for downregulated genes during exponential growth (x1g1) to being enriched for upregulated genes during nitrogen-starvation (x2g2): TCA cycle (GO:0006099), and GO terms related to protein import and rRNA (GO:0030150, GO:0003735, GO:0000462, GO:0006364).

### 3.7. *Transcriptional changes during nitrogen-starvation on glucose or xylose*

A total of 946 genes were found to be differentially expressed during nitrogen-starvation on glucose (g2g1; Figs. 4 & 5A), with 577 upregulated and 369 downregulated genes. It was found that 206 of the 946 differentially expressed genes encoded hypothetical proteins, but only 30 of these were annotated with GO terms. Very similar counts were found for the xylose nitrogen-starvation case (x2x1; Figs. 4 & 5B), with a total of 923 genes fulfilling the designated differential expression cut-off (608 upregulated, 315 downregulated; 195 of 923 encoded hypothetical proteins, 27 with at least one associated GO term). The fatty acid biosynthesis pathway was clearly differentially expressed during both nitrogen-starvation conditions (g2g1 and x2x1), as visualized using the KEGG Mapper overlay function (Figure S9 in Supplemental File S1).

In all, there was a much more drastic change in transcriptional profile upon nitrogen starvation than when comparing the assimilation of xylose and glucose (Fig. 5). A total of 1178 different genes were found to be differentially expressed during the nitrogen-starvation comparisons (g2g1 and x2x1), and 690 of these were differentially expressed on both sugars, which likely indicates a core set of genes related to the nitrogen-starvation response of *P. hubeiensis* BOT-O (Sheet "7. N-starvation core set" in Supplemental File S2). While the 690 genes in the set are too many to discuss on a single-gene level, some physiologically relevant highlights included putative genes related to fatty acid biosynthesis (*FASB, POX1, FOX2*), MEL pathway genes (*EMT1, MAC1/2),* ammonium transport (*MEP2),* mitochondrial oxaloacetate transport (*OAC1),* pyruvate decarboxylase *(PDC1),* aldehyde dehydrogenase (*ALD2)*, fructose-1,6-bisphosphatase *(FBP1),* as well as xylose reductase *XR1* (PHBOTO_004254).

About 80 ribosomal protein genes and genes related to ribosomal activity were detected among the downregulated genes in the nitrogen-starvation core set (Sheet "7. N-starvation core set" in Supplemental File S2). Signs of ribosomal protein down-regulation were also observed in the Bioanalyzer electropherograms of the extracted RNA: the six nitrogen-starvation time point samples had consistently smaller peaks and lower RNA Integrity Number (RIN) values compared to the

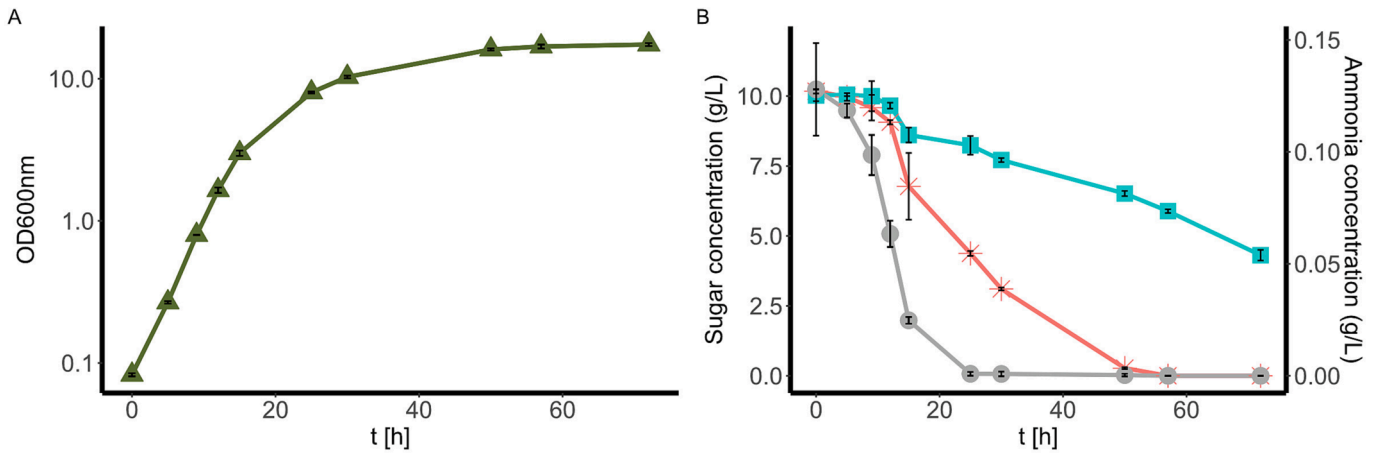



**Fig. 8.** Simultaneous sugar consumption cultivation experiment. The experiment was run for 72 h, with equal concentrations of glucose and xylose (10 g/L each) in shake flasks with a YNB base and ammonium sulphate as nitrogen source. A: Growth measured as $OD_{600nm}$. B: Sugar concentration for glucose (∗, red) and xylose (■, turquoise), as well as nitrogen concentration (●, grey). All experiments were performed in biological triplicates and error bars represent the standard deviations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exponential growth samples (Figure S10 in Supplemental File S1).

As for the gene set analysis, the general trend for both nitrogen-starvation on glucose (g2g1) and on xylose (x2x1) was that a majority of the GO terms were enriched for downregulated genes (Fig. 7), which makes sense from a physiological point-of-view since the cellular state changed from exponential growth to carbon storage. Furthermore, 23 out of a total of 39 GO terms were found to be shared between nitrogen-starvation on glucose and nitrogen-starvation on xylose (Fig. 7). Six shared GO terms were enriched for upregulated genes: two trans-membrane transport terms, hydrolase activity, oxidation–reduction process and two RNA polymerase II DNA binding GO terms (Fig. 7A and B). 15 shared GO terms were enriched for downregulated genes and were related to rRNA, translation and protein folding. The two remaining shared GO terms were related to methionine biosynthetic process and methyltransferase activity.

## 4. Discussion

### 4.1. A high-quality P. hubeiensis genome assembly was obtained using long-read genome sequencing and transcriptome-guided genome annotation

To date, only two strains of *P. hubeiensis* have been subjected to whole-genome sequencing: SY62 (Konishi et al., 2013) and NBRC 105055 (Genbank: GCA_001736105.1). The sequence of the latter does however lack annotation and has a very high degree of fragmentation (21.32 Mb distributed in 9,793 contigs, compared to the 18.44 Mb in 74 contigs for SY62). The low contig count in the final BOT-O assembly can be directly attributed to the long-read data generated by the MinION, and to how five different long-read assemblers were benchmarked for their performance on the reads (Table S1 in Supplemental File S1). While some long-read sequencing studies with Ascomycota yeasts have had good results with e.g. Canu and SMARTdenovo (Istace et al., 2017; Ola et al., 2020), these assemblers were vastly outperformed on the BOT-O data by miniasm and Flye (Table S1 in Supplemental File S1). SMARTdenovo did not perform well in our initial benchmarking and was therefore omitted from the assembler screening, but we were at a later stage able to improve its performance by careful tweaking of each of the assembler's sub-algorithms (Table S1 in Supplemental File S1). This shows that SMARTdenovo is a very promising assembler for long-read data, but that it might require careful parameter adjustment to run well. These results contribute to the increasing number of studies showing how different assemblers have varying performance on genomic material of different origin (Fournier et al., 2017; Jünemann et al., 2014; Molina-Mora et al., 2020; Sun et al., 2021; Wick and Holt, 2019), which highlights the importance of performing assembler benchmarking prior to deciding on the final assembly algorithm. There are numerous factors that affect the quality of an assembly and it has for instance been shown that computational complexity increases with increasing ploidy (Kyriakidou et al., 2018) or repeat-rich regions (Howe et al., 2021), and some assemblers may be better tailored to deal with these complexities than others.

Prior to the differential gene expression analysis, a transcriptome was *de novo* assembled from the 12 RNAseq samples and used as biological evidence for the gene prediction pipeline. Experimental mRNA evidence is known to improve prediction of intron–exon boundaries (Wang et al., 2009), and while *P. hubeiensis* is considered an intron-poor fungus (Lim et al., 2021), the final BOT-O genome assembly nevertheless contained introns. Our gene model resulted in a lower number of identified ORFs than the SY62 annotation (6540 and 7472, respectively), which can probably be attributed to the longer average gene length in BOT-O (2188 bp, compared to 1668 bp in SY62; Table 3). This was also reflected in the increase in complete BUSCOs and the decrease in fragmented BUSCOs in BOT-O compared to SY62 (Table 3). These results suggests that the lower number of genes in our mRNA-guided gene model might represent a better prediction of splice events and, consequently, gene sizes. The number of predicted ORFs in the final BOT-O annotation was indeed closer to another RNAseq-supported annotation: *M. aphidis* DSM 70725 (previously *Pseudozyma aphidis*) that identified 6011 genes (Gunther et al., 2015).

Despite being a distant yeast relative, we prioritized finding *S. cerevisiae* homolog names and applying their annotations and gene names to the predicted BOT-O genes in order to facilitate comparisons with other yeast annotations also using *S. cerevisiae* nomenclature. In this manner, 3657 BOT-O ORFs (56%) were annotated by *S. cerevisiae* homologs. The remaining ORFs were annotated based on basidiomycetes proteins, which for instance resulted in identification of genotypes typical for oleaginous yeasts (*ACL1*) and biosurfactant producers (the MEL pathway). This strategy generally worked well for the central metabolic pathways, as evidenced by the small-scale metabolic reconstruction (Fig. 3), but also resulted in several different BOT-O genes being annotated with the same *S. cerevisiae* homolog. To address this, we manually curated 1138 "duplicate" annotations: the best scoring BLASTp hit for each of the duplicate *S. cerevisiae* proteins was assigned the functional annotation, and the other duplicates were annotated as encoding protein of putatively similar activity. Another challenge with using *S. cerevisiae* as a benchmark to annotate *P. hubeiensis* is that the former has undergone a whole-genome duplication event during its evolutionary history, and thus many *S. cerevisiae* genes have multiple paralogous copies (Kellis et al., 2004). Indeed, during the reconstruction of the central metabolic pathways (Fig. 3), several glycolysis and TCA cycle reactions that are associated with multiple paralogues in *S. cerevisiae* only had one candidate gene in BOT-O; e.g., glyceraldehyde-3-phosphate dehydrogenase that has three paralogs in *S. cerevisiae*, ScTdh1p/2p/3p, but only one in *P. hubeiensis,* PHBOTO_002420.

### 4.2. BOT-O consumes glucose or xylose at equal rates and only 73 genes changed expression levels between the sugars during exponential growth

Equally fast consumption of hexoses and pentoses is a desired trait for industrial applications of lignocellulose fermenting yeasts (Hahn-Hägerdal et al., 2007; Nogue and Karhumaa, 2015). Another *P. hubeiensis* strain, IPM1-10, has previously been reported to consume glucose, xylose or arabinose at similar rates when provided as sole carbon sources, and being capable of co-consumption of the three sugars with a preference for glucose (Tanimura et al., 2016). Likewise, BOT-O displayed an equal aerobic consumption rate of glucose or xylose (0.20 $g{\cdot}L^{-1}{\cdot}h^{-1}$, and 0.19 $g{\cdot}L^{-1}{\cdot}h^{-1}$; Fig. 2) when growing on single sugars. During co-cultivation on a glucose and xylose mixture, simultaneous consumption of glucose and xylose occurred, but just like for *P. hubeiensis* IPM1-10 (Tanimura et al., 2016), BOT-O consumed glucose faster (0.14 $g{\cdot}L^{-1}{\cdot}h^{-1}$ for glucose and 0.08 $g{\cdot}L^{-1}{\cdot}h^{-1}$ for xylose; Fig. 8). However, the total sugar consumption rate in the glucose and xylose mixture (0.22 $g{\cdot}L^{-1}h^{-1}$) was in the same range as the rates of the single sugar cultivations (0.20 $g{\cdot}L^{-1}{\cdot}h^{-1}$, and 0.19 $g{\cdot}L^{-1}{\cdot}h^{-1}$).

In many other yeasts that naturally utilize xylose, or have been engineered to do so, xylose consumption does not fully start in glucose-xylose mixed sugar cultivations until glucose has been depleted (Agbogbo et al., 2006; Dos Santos et al., 2013; Gong et al., 2012; Hou et al., 2017; Kumar and Gummadi, 2011; Ribeiro et al., 2021; Ryu et al., 2016; Shin et al., 2015; Yamada et al., 2017; Yu et al., 2014; Zhao et al., 2008). Compared to these species, BOT-O displayed much slower sugar consumption rates, but the fact that it was able to simultaneously consume both glucose and xylose makes it stand out among xylose-utilizing yeasts. In total, we were only able to find records of two yeast species other than *P. hubeiensis* reported to be capable of simultaneous glucose-xylose consumption without delayed xylose-assimilation: *Cystobasidium iriomotense* (Tanimura et al., 2018) and *Cutaneotrichosporon cutaneum* (Guerfali et al., 2018; Hu et al., 2011). These three are thus highly relevant species for future studies on the mechanisms yeast glucose-xylose co-utilization.

With the similar growth and sugar consumption rates during

cultivation of glucose or xylose in mind (Fig. 2), we hypothesized that the differential gene expression patterns would be similar on either sugar. Indeed, only 73 genes displayed significant expression difference on xylose as compared to glucose during exponential growth (Fig. 4; Sheet "5. DE x1_vs_g1" in Supplemental File S2). The 73 genes included ORFs like *XKS1*, *XYL2, ALD4, PHO84, PCK1* and *MAE1*. In general, genes related to central carbon metabolism did not change their expression levels much between sugars (Fig. 3), which suggest that there is a high degree of co-expression of these pathways in the presence of glucose or xylose. Another possible implication of these results is that the similar sugar utilization rates might be caused not so much by the genes that were differently expressed, but rather the high number of genes that had a baseline expression in both cases (see e.g. normalized counts per gene in Sheet "2. Pathway assignment" in Supplemental File S2). In future studies, it would therefore be highly relevant to investigate this co-expression pattern further, and for instance see if genes related to hexose and xylose metabolism are regulated by the same transcription factors in *P. hubeiensis*.

As for the key genes encoding enzymes for catabolism of xylose and its resulting metabolites, it was found that the XR candidate genes (*XR1*: PHBOTO_004254 and *XR3*: PHBOTO_005965) did not increase their fold change above the threshold value of $\log_2$ 2. *XR1* had consistently high normalized read counts (17325 ± 1601 and 21974 ± 918 on glucose and xylose respectively; Supplemental File S2), meaning that it was expressed during growth on both sugars. *XYL2* (PHBOTO_003491) was only upregulated during growth on xylose, with $\log_2$ fold changes of 2.5 and 3.2 during exponential growth and nitrogen starvation respectively. The *U. bevomyces XR3* protein used to identify PHBOTO_005965 has been shown to catalyse the first step of arabinose assimilation (Lee et al., 2016) and *P. hubeiensis* BOT-O and other *P. hubeiensis* strains have indeed been shown to grow on this sugar (Tanimura et al., 2016). Xylitol accumulation is a commonly reported symptom of inefficient XR/XDH pathways due to unbalanced NAD(P)H requirements of the two enzymes, as often observed in engineered *S. cerevisiae* strains (Matsushika et al., 2009). The absence of xylitol in the HPLC chromatograms of the xylose cultivation samples thus indicates an efficient redox balance over the BOT-O XR/XDH pathway. The number of putative XRs and XDHs identified in BOT-O might indicate that multiple paralogs work together to improve the mass flow through the pathway.

Several putative hydrolases were co-expressed with the xylose pathway and sugar transporter genes when grown on xylose compared to glucose (Table 3). These included a β-xylanase (PHBOTO_003675), a cutinase (PHBOTO_005903), a polygalacturonase (PHBOTO_002755), and two carboxylic ester hydrolases (PHBOTO_004937, PHBOTO_004937). Their upregulation during xylose conditions indicates that the presence of xylose prepares the cell for plant matter degradation and that a panel of different hydrolases, such as xylanases and cutinases, are needed to infect and degrade the plant. Several *Pseudozyma* and *Moesziomyces* species have been reported to secrete extracellular xylanases (Adsul et al., 2009; Borges et al., 2014; Faria et al., 2019; Faria et al., 2015), and we were able to confirm that xylan can be used as a sole carbon source by BOT-O (Figure S8 in Supplemental File S1). Cutin is a biopolymer consisting of ester-linked C16 and C18 fatty acids and a main constituent of the plant cuticle layer, which covers the above-ground parts of plants and serves as a protective barrier (Chen et al., 2013). Xylose has previously been found to strongly induce the expression of a cutinase capable of biopolymer degradation in *M. antarcticus*, but induction was not observed in *M. aphidis*, *P. tsukubasensis* and *P. rugulosa* (Watanabe et al., 2014). The putative cutinase (PHBOTO_005903) found in the BOT-O xylose core set suggests that *P. hubeiensis* has this phenotype, but it remains to be verified *in vivo*. The possibility to cultivate *P. hubeiensis* directly on solid plant matter, and on biopolymers such as xylan and cutin for production of its native lipid and biosurfactant products makes this species a promising host for bioprocessing.

The gene set analysis complemented the single gene analysis by analysing GO term enrichments across all genes that were differentially expressed across the different conditions. When comparing the expression on sugars during exponential growth (x1g1), GO terms enriched for upregulated genes on xylose included PPP (GO:0009051) and detection of glucose (GO:0051594). A representative of the latter term was the glucose-related membrane protein gene *SNF3* that was upregulated on xylose with a $\log_2$ fold change of 9.0 and 6.9 at time point 1 and time point 2, respectively, despite growing well on xylose (Fig. 2). In some yeasts, Snf3p acts as a glucose transporter, but in *S. cerevisiae* it has evolved into a low glucose concentration sensor (Özcan et al., 1996), and the "detection of glucose" GO term might therefore be defined based on the type of function found in *S. cerevisiae*. Transmembrane transport terms (GO:0055085, GO:0022857) were enriched for upregulated genes on xylose at both time points (x1g1 and x2g2) and included sugar transporters such as *HXT8*, *HXT17*, and *GAL2*. In *S. cerevisiae,* the galactose transporter Gal2p is also capable of transporting xylose (Hamacher et al., 2002), but the substrate affinities of the putative *GAL2* (PHBOTO_003550) remains to be verified.

The number of transcriptional changes at the designated differential expression cut-off in the BOT-O glucose and xylose comparisons (x1g1 and x2g2) seems to be low compared to other transcriptomics studies on yeasts naturally capable of xylose assimilation, such as *Scheffersomyces stipitis* CBS 6054 (Yuan et al., 2011) and *Candida intermedia* CBS 141442 (Geijer et al., 2020). While these studies have used slightly different approaches and cut-offs and are therefore not directly comparable, we speculate that the similar consumption rate of the two sugars in BOT-O is related to the low amount of differentially expressed genes between the conditions. The *P. hubeiensis* glucose-xylose co-consumption trait (Fig. 8) either implies that the transporter affinity is lower to xylose than to glucose, or that glucose catabolite repression is activated to some degree during presence of multiple sugar types, despite the equal rates on the single sugars. Glucose catabolite repression have been shown to vary between yeasts. For instance, *SNF3* is not related to glucose repression in the naturally xylose utilizing yeast *K. marxianus,* and instead *SNF1* and *HXK1* seem to fill this role (Hua et al., 2019). In *S. stipitis*, deletion of *HXK1* has been found to derepress xylose utilization (Dashtban et al., 2015). Neither *SNF1* nor *HXK2*, a paralog to *HXK1*, fulfilled the designated differential expression cut-off in BOT-O. However, sugar signalling pathways in yeasts are regulated by post-translational modifications such as phosphorylations (Santangelo, 2006), and not necessarily by expression levels. Yeast sugar signalling pathways have primarily been studied with hexose sugars such as glucose and galactose using the model yeast *S. cerevisiae*, which cannot naturally grow on xylose, and thus the current knowledge of the effect of xylose on the sugar signalling network is limited (Brink et al., 2021). The difference in sugar consumption rate during co-consumption – but not during cultivation on single sugars – indicates that some degree of repression is active, but it is difficult to predict which genes cause this. Future studies dedicated to *P. hubeiensis* sugar sensing and signalling will likely be needed to increase the understanding of the underlying mechanisms of this phenotype. Attempts at transforming BOT-O have not yet been successful, and therefore efficient genetic engineering protocols have to be developed before genetic validation of putative mechanisms in this new, natural isolate can be achieved. Assessment of the transcriptional profile during glucose-xylose mixed sugar cultivation would also be of high relevance.

### 4.3. Nitrogen-starvation resulted in an oleaginous phenotype and a considerable change in gene expression in BOT-O

The nitrogen-starvation sampling points were chosen based on a time point where the ammonium in the medium was depleted and the strain was clearly accumulating lipids, as observed by microscopy and quantified by GC–MS. The lipid accumulation reached ~ 30% of cell dry weight on either sugar at the nitrogen-starvation time points, which clearly qualifies BOT-O as an oleaginous species, but there is likely room for increasing the total accumulated lipids by optimization of C/N ratios,

bioprocess design and lipid turnover minimisation.

The nitrogen-starvation conditions resulted in a drastic change in the transcriptional profile, on both glucose and xylose media (1178 unique genes were differentially expressed in g2g1 and x2x1). This is expected from a cell transitioning from nutrient-unlimited exponential growth to starvation and has also been observed in other oleaginous yeasts (Zhu et al., 2012). In BOT-O, genes in the central metabolic pathways displayed generally higher $\log_2$ fold changes on nitrogen-starvation than during the sugar comparison conditions (Fig. 3). Genes for production of cytosolic acetyl-CoA and its subsequent conversion to storage lipids were expressed during nitrogen-starvation conditions, but the $\log_2$ fold changes and normalized read counts varied between genes and not all of them fulfilled the designated differential expression cut-off. In BOT-O, we discovered a single-gene FAS complex, just like in its relative *U. maydis*. However, we found not one, but two putative genes that each encode a single FAS complex: *FASA* and *FASB*. These putative *FAS* genes in BOT-O had vast differences in $\log_2$ fold changes, suggesting that they share the same function but are induced under different conditions. The *FASA* candidate (PHBOTO_002928) had a consistently high normalized read count (15 k-25 k reads) in all samples and time points (see e.g. Sheet "2. Pathway assignment" in Supplemental File S2). However, *FASB* (PHBOTO_006517) was highly up-regulated in the nitrogen-starved glucose and xylose cultures ($\log_2$ 9.5 and 3.4 times, respectively), and was, in terms of normalized read count, one of the most highly expressed genes in the glucose time point 2 samples (222784 ± 9734 normalized reads; Supplemental File S2). We hypothesize that BOT-O *FASA* and *FASB* have the same function but are induced by different conditions: *FASA* might be constitively expressed for production of basal amounts of lipids, while *FASB* might be triggered by nitrogen starvation.

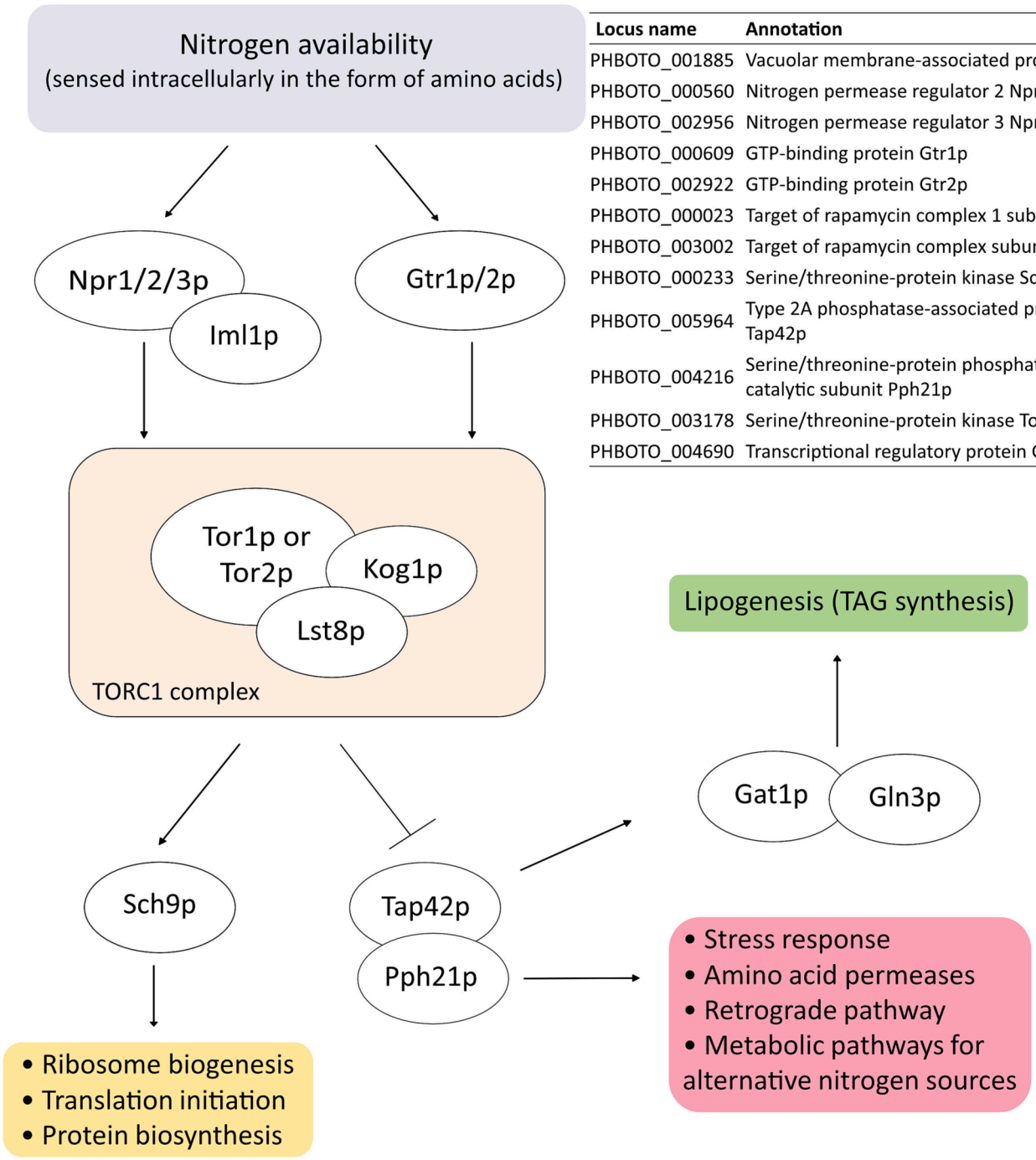
The oleagenicity-related *ACL1* (PHBOTO_001285) gene candidate also had consistently high normalized read counts in all conditions (Supplemental File S2). Compared to the read levels of many other genes, *FASA* and *ACL1* were highly expressed throughout all conditions, suggesting that they have physiologically important roles also during exponential growth and cell proliferation. Basal levels of lipids (~5% of CDW), likely related to cell structure, were indeed observed in the exponential growth samples (Fig. 2D). The other predicted fatty acid pathway genes had normalized counts of > 1000 reads in all samples (Supplemental File S2). Similar trends of basal transcription levels for fatty acid pathway genes before and after nitrogen-starvation have also been observed in *Y. lipolytica* (Morin et al., 2011) and *R. toruloides* (Zhu et al., 2012). Many of the fatty acid genes, as well as genes involved in TAG biosynthesis, did display a fold-change increase during nitrogen-starvation, but most were below the $\log_2$ 2-fold-cut-off applied in this study.

The mannosylerythritol lipids (MEL) pathway for production of secreted glycolipid biosurfactants was also strongly upregulated during nitrogen-starvation (Fig. 3 and Supplemental File S2), with three putative genes (*EMT1*, *MAC1*, *MAC2*) reaching $\log_2$ fold changes of 2.8–9.2 on either sugar; *MAT1,* the last gene of the pathway (Fig. 3), was upregulated 2.3 times on glucose and 1.5 times on xylose. The fold change was slightly higher for the MEL pathway when grown on glucose than on xylose, but how this affects MEL accumulation levels is not possible to extrapolate from the current data. Since the current study was focused on intracellular storage lipid accumulation in *P. hubeiensis*, measurements of MELs were outside of the scope of the project; however, nitrogen-starvation has in fact been shown to increase MEL production in other *Pseudozyma* species (Faria et al., 2014). The upregulation of the MEL genes during nitrogen-starvation strongly indicates that MELs were produced by BOT-O in parallel to the lipid accumulation phase, and while previous studies mainly produced MELs from vegetable oils, the metabolic pathway for production of MELs from glucose has been described (Morita et al., 2014). It is therefore likely that the measurement of only the intracellular storage lipids might have underestimated the total BOT-O lipid production as it did not include the extracellularly stored MELs.

Two of the major GO terms enriched with up-regulated genes after nitrogen-starvation (Fig. 7) were transmembrane transport (GO:0055085) and transmembrane transport activity (GO:0022857) with over 150 genes that were differentially expressed in either glucose or xylose. During nutrient starvation, yeast cells will attempt to scavenge for the limited nutrients in various ways, one of which is increased transporter expression (Conrad et al., 2014). In BOT-O, the *MEP2* ammonium transporter candidate gene (PHBOTO_002135) increased its expression during nitrogen-starvation on glucose and on xylose by $\log_2$ 3.7 and 2.6 fold respectively, which could indicate that the cell is trying to scavenge any available nitrogen from the environment; *MEP2* was also highly upregulated during nitrogen-starvation in the oleaginous yeast *R. toruloides* (Zhu et al., 2012), and has been related to pseudohyphal growth for nutrient scavenging in *S. cerevisiae* upon nitrogen-starvation (Lorenz and Heitman, 1998). Likewise, transcripts of two BOT-O putative oligopeptide transporter genes (PHBOTO_004291 and PHBOTO_004987) increased by $\log_2$ 3.3 and 4.3-fold on glucose or xylose, which indicates that the cell is trying to satisfy its nitrogen demand by scavenging for short peptides that can be broken down to their respective amino acids and used as nitrogen supply. *GAP1* (PHBOTO_000497), putatively encoding a general amino acid permease, also fulfilled the designated differential expression cut-off on glucose during nitrogen-starvation (g2g1), but not in any of the other conditions. Several genes with homology to *S. cerevisiae PHO84*, which encodes an inorganic phosphate transporter, changed expression after nitrogen-starvation: PHBOTO_006079 and PHBOTO_000024 increased more than $\log_2$ 2-fold (Supplemental File S1) whereas PHBOTO_000828 decreased by over $\log_2$ 4.6-fold on both sugars. This implies that BOT-O was not only limited in nitrogen at the time point 2 samples, but possibly also in phosphate, which was neither measured nor controlled in the cultivation.

The gene set analysis for the nitrogen-starvation differential expression conditions also revealed that GO-terms related to ribosomal processes and translational initiation were enriched with downregulated genes (Fig. 7). Down-regulation of expression of genes coding for ribosomal proteins has also been observed in nitrogen-starvation RNAseq studies in yeasts (Aliyu et al., 2021; Duncan and Mata, 2017; Pomraning et al., 2016; Tesniere et al., 2018; Zhu et al., 2012), as well as in a filamentous fungus (Twumasi-Boateng et al., 2009). The overall lower RIN values observed in the six nitrogen-starved RNA samples are likely also related to the ribosomal down-regulation pattern. The RIN is calculated from multiple parameters, but two of the main contributors are the 28S:18S ribosomal peak area ratio and the 28S peak height; the latter being one of the first peaks to decrease during total RNA degradation (Schroeder et al., 2006). While high RNA integrity in the samples will always be desirable due to its correlation to increased RNAseq data quality (Romero et al., 2014), our results suggest that the RIN value can be misleading for assessment of RNA sample quality from non-growth conditions, e.g. during nitrogen-starvation.

A link between nitrogen-starvation and ribosomal downregulation can be found in the Target of Rapamycin (TOR) signalling pathway, which is found in yeasts, plants and mammals and induces genes related to cell growth upon sensing of available nitrogen sources (Conrad et al., 2014; Inoki et al., 2005; Zhang et al., 2018). Several key TOR pathway genes in *S. cerevisiae* such as *NPR2/3, IML1, GTR1/2, KOG1, TOR1, LST8, SCH9, TAP42* and *PPH21* (Hughes Hallet et al., 2014; Zhang et al., 2018) had homologs in BOT-O (Sheet "2. Pathway assignment" in Supplemental File S2), and we thus hypothesize that the TOR signalling pathway functions in a similar manner in *P. hubeiensis* (Fig. 9). The key signal element of the TOR pathway is the TORC1 complex (Fig. 9), which in yeasts has been shown to be able to be formed with either Tor1p or Tor2p as one of the subunits (Beauchamp and Platanias, 2013). The BOT-O functional annotation gave the same single significant hit for Tor1p and Tor2p: PHBOTO_003178. Sensing of utilizable nitrogen sources by TOR pathway proteins in *S. cerevisiae* results in a signal




| Locus name | Annotation |
|---|---|
| PHBOTO_001885 | Vacuolar membrane-associated protein Iml1p |
| PHBOTO_000560 | Nitrogen permease regulator 2 Npr2p |
| PHBOTO_002956 | Nitrogen permease regulator 3 Npr3p |
| PHBOTO_000609 | GTP-binding protein Gtr1p |
| PHBOTO_002922 | GTP-binding protein Gtr2p |
| PHBOTO_000023 | Target of rapamycin complex 1 subunit Kog1p |
| PHBOTO_003002 | Target of rapamycin complex subunit Lst8p |
| PHBOTO_000233 | Serine/threonine-protein kinase Sch9p |
| PHBOTO_005964 | Type 2A phosphatase-associated protein Tap42p |
| PHBOTO_004216 | Serine/threonine-protein phosphatase PP2A-1 catalytic subunit Pph21p |
| PHBOTO_003178 | Serine/threonine-protein kinase Tor2p |
| PHBOTO_004690 | Transcriptional regulatory protein Gat1p |

**Fig. 9.** Overview of key elements of the *S. cerevisiae* TOR signalling pathway found to have homologs in *P. hubeiensis* BOT-O. The TOR pathway senses nitrogen availability and induces the ribosomal translational machinery upon presence of nitrogen and represses it during absence of nitrogen via the Sch9p kinase. The so-called *PP2A branch* (here represented by Tap42p/Pph21p) regulates genes that control stress response, utilization of alternative nitrogen-sources and lipogenesis. Arrows with arrowheads: induction signals; arrows with hammerheads: repression signals.
Adapted from Bracharz et al. (2017); Conrad et al. (2014); Hughes Hallet et al. (2014); Madeira et al. (2015); Zhang et al. (2018)

cascade that induces the Sch9p kinase (homologous to PHBOTO_000233), which in turn induces expression of genes related to protein and ribosomal biogenesis by direct induction of the genes (Hughes Hallet et al., 2014; Zhang et al., 2018), see Fig. 9. *S. cerevisiae* also has an additional parallel control over the same ribosomal biogenesis genes in form of the Dot6p/Tod6p transcriptional repressors, that are inactivated upon sensing of available nitrogen by TOR (Lippman and Broach, 2009). No significant homologs to *S. cerevisiae* Dot6p/Tod6p were however found in BOT-O. All proteins annotated as Dot6p/Tod6p in Uniprot at the time of writing were in fact only found in ascomycete yeasts, which further implies that basidiomycetes either do not possess this additional layer of control over the ribosomal biogenesis genes, or have an alternative mechanism using proteins that do not share homology to those of *S. cerevisiae*. In addition to gene expression level regulation, the TOR pathway has also been implicated in induction of selective ribosome degradation by autophagy during nitrogen starvation in *S. cerevisiae* (Kraft et al., 2008). Absence of nitrogen sources thus seem to send a clear signal to down-regulate the rate of translation, by

simultaneously limiting expression of new ribosomal proteins and increasing the degradation rate of certain ribosomal proteins.

Sensing of absence of nitrogen by the TOR pathway has been linked to lipid production via Gat1p and Gln3p in *S. cerevisiae* (Madeira et al., 2015) and in the oleaginous yeast *Trichosporon oleaginosus* (Bracharz et al., 2017). We believe that this might also be the case for *P. hubeiensis* since a majority of the TOR pathway elements were conserved in BOT-O (see Sheet "2. Pathway assignment" in Supplemental File S2). All BOT-O TOR pathway homologs were basally expressed during all four comparisons with the sole exception of *KOG1*, which was downregulated on exponential growth. However, as was discussed above for the sugar signalling case, the fact that signalling pathways function mainly by protein–protein interactions and phosphorylations (Hyduke and Palsson, 2010) means that transcriptomics cannot capture these signalling events. Nevertheless, the genes regulated by the TOR pathway, such as ribosome-related genes (Huber et al., 2011), were clearly downregulated in BOT-O.

## 5. Conclusions

While *P. hubeiensis* is among the lesser studied yeasts in literature – partly due to its first isolation only 16 years ago at the time of writing (Wang et al., 2006) – it possesses several biotechnologically relevant traits such as accumulation of MELs and intracellular storage lipids, and secretion of hydrolases for breaking down biomass, such as glycosidases and xylanases. It furthermore grows almost equally well on glucose or xylose. This is an uncommon and desirable trait among yeasts that makes it a promising candidate for biological valorisation of lignocellulose-derived sugar mixtures. All these traits give BOT-O a great potential to become an industrial strain. The current study has increased the knowledge on glucose and xylose utilization in *P. hubeiensis,* as well as its oleaginous phenotype, by analysing the differential gene expression response during exponential growth and nitrogen-starvation. By using the acquired mRNA data as evidence in the gene prediction pipeline, we have generated the first transcriptome-supported *P. hubeiensis* genome annotation. While more studies are needed to elucidate the mechanisms of glucose repression and transporter affinities during mixed-sugar cultivations, we believe that the data presented in the current study will be a useful asset not only for future studies on *P. hubeiensis,* but also for studies on other xylose-assimilating yeasts and oleaginous yeasts.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### *Acknowledgements*

The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory (SciLifeLab), the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

The computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Computations were performed in project snic2020-15-218, and project snic2020-16-242 was used for extended space for data storage.

### *Author contributions*

FM performed the wet experiments (cultivations, DNA and RNA extraction, MinION genome sequencing, lipid quantifications). DB and FM designed the bioinformatics analysis workflow, and together performed the genome assembly, differential gene expression, gene set analysis and data analysis. DB performed the genome annotation and *de novo* transcriptome assembly. DB and FM drafted the initial paper. TA conceived the study. VS, JN and TA revised the manuscript. All authors read and approved the final manuscript.

### *Funding*

This work was financed by the Swedish research council Formas through grant contract 2018–01875.

### *Availability of data*

The final annotated genome was uploaded to NCBI with accession number JAJJYS000000000. The raw genome reads generated by MinION and Illumina were uploaded to the Sequence Read Archive (SRA) with accession numbers SRR16964327 and SRR16964328. The RNAseq data was uploaded to the Gene Expression Omnibus (GEO) database with accession number GSE193998. The transcriptome assembly was uploaded to the Transcriptome Shotgun Assembly (TSA) database under the accession GJWK00000000. The version described in this paper is the first version, GJWK01000000. The BOT-O strain is available upon request from Chalmers University of Technology, Sweden, and is currently being deposited at the German Collection of Microorganisms and Cell Cultures (DSMZ).

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fgb.2023.103783.

## References

Adsul, M.G., Bastawde, K.B., Gokhale, D.V., 2009. Biochemical characterization of two xylanases from yeast *Pseudozyma hubeiensis* producing only xylooligosaccharides. Bioresour. Technol. 100, 6488–6495.

Agbogbo, F.K., Coward-Kelly, G., Torry-Smith, M., Wenger, K.S., 2006. Fermentation of glucose/xylose mixtures using *Pichia stipitis*. Process Biochem. 41, 2333–2336.

Aliyu, H., Gorte, O., Neumann, A., Ochsenreither, K., 2021. Global transcriptome profile of the oleaginous yeast *Saitozyma podzolica* DSM 27192 cultivated in glucose and xylose. J. Fungi 7, 758.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Andlid, T., Larsson, C., Liljenberg, C., Marison, I., Gustafsson, L., 1995. Enthalpy content as a function of lipid accumulation in *Rhodotorula glutinis*. Appl. Microbiol. Biotechnol. 42, 818–825.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Armenteros, J.J.A., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H., 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 37, 420–423.

Balat, M., Balat, H., 2009. Recent trends in global production and utilization of bio-ethanol fuel. Appl. Energy 86, 2273–2282.

Bao, W.D., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6.

Beauchamp, E.M., Platanias, L.C., 2013. The evolution of the TOR pathway and its role in cancer. Oncogene 32, 3923–3932.

Belgacem, M.N., Gandini, A., 2008. Materials from vegetable oils: major sources, properties and applications. Monomers, Polymers and Composites from Renewable Resources. Elsevier, pp. 39–66.

Benevenuto, J., Teixeira-Silva, N.S., Kuramae, E.E., Croll, D., Monteiro-Vitorello, C.B., 2018. Comparative genomics of smut pathogens: insights from orphans and positively selected genes into host specialization. Front. Microbiol. 9, 660.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., van Nimwegen, E., 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol. Biol. Evol. 31, 1077–1088.

Blighe, K., Rana, S., Lewis, M., 2019. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version.

Blitzblau, H.G., Consiglio, A.L., Teixeira, P., Crabtree, D.V., Chen, S.Y., Konzock, O., Chifamba, G., Su, A., Kamineni, A., MacEwen, K., Hamilton, M., Tsakraklides, V., Nielsen, J., Siewers, V., Shaw, A.J., 2021. Production of 10-methyl branched fatty acids in yeast. Biotechnol. Biofuels 14.

Boekhout, T., 1995. *Pseudozyma* Bandoni emend Boekhout, a genus for yeast-like anamorphs of Ustilaginales. J. Gen. Appl. Microbiol. 41, 359–366.
Borges, T.A., De Souza, A.T., Squina, F.M., Riaño-Pachón, D.M., dos Santos, R.A.C., Machado, E., de Castro Oliveira, J.V., Damásio, A.R., Goldman, G.H., 2014. Biochemical characterization of an endoxylanase from *Pseudozyma brasiliensis* sp. nov. strain GHG001 isolated from the intestinal tract of *Chrysomelidae larvae* associated to sugarcane roots. Process Biochem. 49, 77–83.
Borneman, A.R., Desany, B.A., Riches, D., Affourtit, J.P., Forgan, A.H., Pretorius, I.S., Egholm, M., Chambers, P.J., 2011. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. PLoS Genet. 7, e1001287.
Boulton, C.A., Ratledge, C., 1981. Correlation of lipid-accumulation in yeasts with possession of ATP-citrate lyase. J. Gen. Microbiol. 127, 169–176.
Bracharz, F., Redai, V., Bach, K., Qoura, F., Brück, T., 2017. The effects of TORC signal interference on lipogenesis in the oleaginous yeast *Trichosporon oleaginosus*. BMC Biotech. 17, 1–9.
Braunwald, T., Schwemmlein, L., Graeff-Honninger, S., French, W.T., Hernandez, R., Holmes, W.E., Claupein, W., 2013. Effect of different C/N ratios on carotenoid and lipid production by *Rhodotorula glutinis*. Appl. Microbiol. Biotechnol. 97, 6581–6588.
Brink, D.P., Borgström, C., Persson, V.C., Ofuji Osiro, K., Gorwa-Grauslund, M.F., 2021. D-xylose sensing in *Saccharomyces cerevisiae*: Insights from D-glucose signaling and native D-xylose utilizers. Int. J. Mol. Sci. 22, 12410.
Bruna, T., Hoff, K.J., Lomsadze, A., Stanke, M., Borodovsky, M., 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP plus and AUGUSTUS supported by a protein database. NAR Genom. Bioinform. 3.
Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60.
Campbell, M.S., Holt, C., Moore, B., Yandell, M., 2014. Genome annotation and curation using MAKER and MAKER-P. Curr. Protocols Bioinformat. 48, 4.11.1–4.11.39.
Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.
Chan, P.P., Lin, B.Y., Mak, A.J., Lowe, T.M., 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucl. Acids Res. 49, 9077–9096.
Chen, S., Su, L., Chen, J., Wu, J., 2013. Cutinase: characteristics, preparation, and application. Biotechnol. Adv. 31, 1754–1767.
Conrad, M., Schothorst, J., Kankipati, H.N., Van Zeebroeck, G., Rubio-Texeira, M., Thevelein, J.M., 2014. Nutrient sensing and signaling in the yeast *Saccharomyces cerevisiae*. FEMS Microbiol. Rev. 38, 254–299.
Dainat, J., 2021. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. Zenodo. https://doi.org/10.5281/zenodo.3552717.
Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008.
Dashtban, M., Wen, X., Bajwa, P.K., Ho, C.-Y., Lee, H., 2015. Deletion of *hxk1* gene results in derepression of xylose utilization in *Scheffersomyces stipitis*. J. Ind. Microbiol. Biotechnol. 42, 889–896.
De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., Van Broeckhoven, C., 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34, 2666–2669.
Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., Osorio, S., Tohge, T., Fernie, A.R., Feussner, I., Feussner, K., Meinicke, P., Stierhof, Y.D., Schwarz, H., Macek, B., Mann, M., Kahmann, R., 2011. Metabolic priming by a secreted fungal effector. Nature 478, 395-.
Doehlemann, G., van der Linde, K., Amann, D., Schwammbach, D., Hof, A., Mohanty, A., Jackson, D., Kahmann, R., 2009. Pep1, a secreted effector protein of *Ustilago maydis*, is required for successful invasion of plant cells. PLoS Pathog. 5.
Doehlemann, G., Reissmann, S., Assmann, D., Fleckenstein, M., Kahmann, R., 2011. Two linked genes encoding a secreted effector and a membrane protein are essential for *Ustilago maydis*-induced tumour formation. Mol. Microbiol. 81, 751–766.
Dos Santos, V.C., Bragança, C.R.S., Passos, F.J.V., Passos, F.M.L., 2013. Kinetics of growth and ethanol formation from a mix of glucose/xylose substrate by *Kluyveromyces marxianus* UFV-3. Antonie Van Leeuwenhoek 103, 153–161.
Duncan, C.D.S., Mata, J., 2017. Effects of cycloheximide on the interpretation of ribosome profiling experiments in *Schizosaccharomyces pombe*. Sci. Rep. 7, 10331.
Durrett, T.P., Benning, C., Ohlrogge, J., 2008. Plant triacylglycerols as feedstocks for the production of biofuels. Plant J. 54, 593–607.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.
Eilbeck, K., Moore, B., Holt, C., Yandell, M., 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinf. 10.
Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20.
Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Dwight, S.S., Hitz, B.C., Karra, K., Nash, R.S., Weng, S., Wong, E.D., Lloyd, P., Skrzypek, M.S., Miyasato, S.R., Simison, M., Cherry, J.M., 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3-Genes Genomes. Genetics 4, 389–398.
Ewels, P., Magnusson, M., Lundin, S., Kaller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048.
Faria, N.T., Santos, M.V., Fernandes, P., Fonseca, L.L., Fonseca, C., Ferreira, F.C., 2014. Production of glycolipid biosurfactants, mannosylerythritol lipids, from pentoses and D-glucose/D-xylose mixtures by *Pseudozyma* yeast strains. Process Biochem. 49, 1790–1799.
Faria, N.T., Marques, S., Fonseca, C., Ferreira, F.C., 2015. Direct xylan conversion into glycolipid biosurfactants, mannosylerythritol lipids, by *Pseudozyma antarctica* PYCC 5048T. Enzyme Microb. Technol. 71, 58–65.
Faria, N.T., Marques, S., Ferreira, F.C., Fonseca, C., 2019. Production of xylanolytic enzymes by *Moesziomyces* spp. using xylose, xylan and brewery's spent grain as substrates. N. Biotechnol. 49, 137–143.
Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. USA 117, 9451–9457.
Fournier, T., Gounot, J.S., Freel, K., Cruaud, C., Lemainque, A., Aury, J.M., Wincker, P., Schacherer, J., Friedrich, A., 2017. High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using Nanopore MinION sequencing. G3-Genes Genomes. Genetics 7, 3243–3250.
Foyle, T., Jennings, L., Mulcahy, P., 2007. Compositional analysis of lignocellulosic materials: evaluation of methods used for sugar analysis of waste paper and straw. Bioresour. Technol. 98, 3026–3036.
Fukuoka, T., Morita, T., Konishi, M., Imura, T., Sakai, H., Kitamoto, D., 2007. Structural characterization and surface-active properties of a new glycolipid biosurfactant, mono-acylated mannosylerythritol lipid, produced from glucose by *Pseudozyma antarctica*. Appl. Microbiol. Biotechnol. 76, 801–810.
Geijer, C., Faria-Oliveira, F., Moreno, A.D., Stenberg, S., Mazurkewich, S., Olsson, L., 2020. Genomic and transcriptomic analysis of *Candida intermedia* reveals the genetic determinants for its xylose-converting capacity. Biotechnol. Biofuels 13.
Gong, Z., Wang, Q., Shen, H., Hu, C., Jin, G., Zhao, Z.K., 2012. Co-fermentation of cellobiose and xylose by *Lipomyces starkeyi* for lipid production. Bioresour. Technol. 117, 20–24.
Guerfali, M., Ayadi, I., Belhassen, A., Gargouri, A., Belghith, H., 2018. Single cell oil production by *Trichosporon cutaneum* and lignocellulosic residues bioconversion for biodiesel synthesis. Process Saf. Environ. Prot. 113, 292–304.
Gunther, M., Grumaz, C., Lorenz, S., Stevens, P., Lindemann, E., Hirth, T., Sohn, K., Zibek, S., Rupp, S., 2015. The transcriptomic profile of *Pseudozyma aphidis* during production of mannosylerythritol lipids. Appl. Microbiol. Biotechnol. 99, 1375–1388.
Guzman, P.A., Sanchez, J.G., 1994. Characterization of telomeric regions from *Ustilago maydis*. Microbiology-UK 140, 551–557.
Hahn-Hägerdal, B., Karhumaa, K., Fonseca, C., Spencer-Martins, I., Gorwa-Grauslund, M. F., 2007. Towards industrial pentose-fermenting yeast strains. Appl. Microbiol. Biotechnol. 74, 937–953.
Hamacher, T., Becker, J., Gárdonyi, M., Hahn-Hägerdal, B., Boles, E., 2002. Characterization of the xylose-transporting properties of yeast hexose transporters and their influence on xylose utilization. Microbiology 148, 2783–2788.
Hamelinck, C.N., van Hooijdonk, G., Faaij, A.P.C., 2005. Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle- and long-term. Biomass Bioenergy 28, 384–410.
Hewald, S., Linne, U., Scherer, M., Marahiel, M.A., Kamper, J., Bolker, M., 2006. Identification of a gene cluster for biosynthesis of mannosylerythritol lipids in the basidiomycetous fungus *Ustilago maydis*. Appl. Environ. Microbiol. 72, 5469–5477.
Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182–192.
Hoff, K., Lomsadze, A., Borodovsky, M., Stanke, M., 2019. Whole-genome annotation with BRAKER. In: Kollmar, M. (Ed.), Gene Prediction. Springer, New York, pp. 65–95.
Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 12.
Hong, K.K., Nielsen, J., 2012. Metabolic engineering of *Saccharomyces cerevisiae*: a key cell factory platform for future biorefineries. Cell. Mol. Life Sci. 69, 2671–2690.
Hou, J., Qiu, C., Shen, Y., Li, H., Bao, X., 2017. Engineering of *Saccharomyces cerevisiae* for the efficient co-utilization of glucose and xylose. FEMS Yeast Res. 17.
Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., Torrance, J., Tracey, A., Wood, J., 2021. Significantly improving the quality of genome assemblies through curation. GigaScience 10, giaa153.
Hu, C., Wu, S., Wang, Q., Jin, G., Shen, H., Zhao, Z.K., 2011. Simultaneous utilization of glucose and xylose for lipid production by *Trichosporon cutaneum*. Biotechnol. Biofuels 4, 25.
Hua, Y., Wang, J., Zhu, Y., Zhang, B., Kong, X., Li, W., Wang, D., Hong, J., 2019. Release of glucose repression on xylose utilization in *Kluyveromyces marxianus* to enhance glucose-xylose co-utilization and xylitol production from corncob hydrolysate. Microb. Cell Fact. 18, 1–18.
Huber, A., French, S.L., Tekotte, H., Yerlikaya, S., Stahl, M., Perepelkina, M.P., Tyers, M., Rougemont, J., Beyer, A.L., Loewith, R., 2011. Sch9 regulates ribosome biogenesis via Stb3, Dot6 and Tod6 and the histone deacetylase complex RPD3L. EMBO J. 30, 3052–3064.
Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W.D., Smit, A.F.A., Wheelers, T.J., 2016. The Dfam database of repetitive DNA families. Nucl. Acids Res. 44, D81–D89.
Hughes Hallet, J.E., Luo, X.X., Capaldi, A.P., 2014. State transitions in the TORC1 signaling pathway and information processing in *Saccharomyces cerevisiae*. Genetics 198, 773–786.
Huson, D.H., Scornavacca, C., 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61, 1061–1067.
Hyduke, D.R., Palsson, B.Ø., 2010. Towards genome-scale signalling-network reconstructions. Nat. Rev. Genet. 11, 297–307.
Inoki, K., Ouyang, H., Li, Y., Guan, K.-L., 2005. Signaling by target of rapamycin proteins in cell growth control. Microbiol. Mol. Biol. Rev. 69, 79–100.
Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., Lemainque, A., Engelen, S., Wincker, P.,

Schacherer, J., Aury, J.M., 2017. *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. GigaScience 6.
Jenkins, R.W., Sargeant, L.A., Whiffin, F.M., Santomauro, F., Kaloudis, D., Mozzanega, P., Bannister, C.D., Baena, S., Chuck, C.J., 2015. Cross-metathesis of microbial oils for the production of advanced biofuels and chemicals. ACS Sustain. Chem. Eng. 3, 1526–1535.
Jia, N., Wang, J., Shi, W., Du, L., Sun, Y., Zhan, W., Jiang, J.-F., Wang, Q., Zhang, B., Ji, P., 2020. Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities. Cell 182 (1328–1340), e13.
Jin, M.J., Slininger, P.J., Dien, B.S., Waghmode, S., Moser, B.R., Orjuela, A., Sousal, L.D., Balan, V., 2015. Microbial lipid-based lignocellulosic biorefinery: feasibility and challenges. Trends Biotechnol. 33, 43–54.
Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W.Z., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.
Jünemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., Harmsen, D., 2014. GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. PLoS One 9, e107014.
Kämper, J., Kahmann, R., Bolker, M., Ma, L.J., Brefort, T., Saville, B.J., Banuett, F., Kronstad, J.W., Gold, S.E., Muller, O., Perlin, M.H., Wosten, H.A.B., de Vries, R., Ruiz-Herrera, J., Reynaga-Pena, C.G., Snetselaar, K., McCann, M., Perez-Martin, J., Feldbrugge, M., Basse, C.W., Steinberg, G., Ibeas, J.I., Holloman, W., Guzman, P., Farman, M., Stajich, J.E., Sentandreu, R., Gonzalez-Prieto, J.M., Kennell, J.C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Munch, K., Rossel, N., Scherer, M., Vranes, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng, S., Ho, E.C.H., Cahill, M.J., Boyce, K.J., Klose, J., Klosterman, S. J., Deelstra, H.J., Ortiz-Castellanos, L., Li, W.X., Sanchez-Alonso, P., Schreier, P.H., Hauser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schluter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mannhaupt, G., Guldener, U., Munsterkotter, M., Haase, D., Oesterheld, M., Mewes, H.W., Mauceli, E.W., DeCaprio, D., Wade, C.M., Butler, J., Young, S., Jaffe, D.B., Calvo, S., Nusbaum, C., Galagan, J., Birren, B.W., 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444, 97–101.
Kanehisa, M., Sato, Y., 2020. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 29, 28–35.
Kanehisa, M., Sato, Y., Morishima, K., 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726–731.
Kellis, M., Birren, B.W., Lander, E.S., 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428, 617–624.
Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907-.
Kitamoto, D., Akiba, S., Hioki, C., Tabuchi, T., 1990. Extracellular accumulation of mannosylerythritol lipids by a strain of *Candida antarctica*. Agric. Biol. Chem. 54, 31–36.
Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. 37, 540-.
Konishi, M., Hatada, Y., Horiuchi, J.-I., 2013. Draft genome sequence of the basidiomycetous yeast-like fungus *Pseudozyma hubeiensis* SY62, which produces an abundant amount of the biosurfactant mannosylerythritol lipids. Genome Announcements. 1, e00409–13.
Konishi, M., Morita, T., Fukuoka, T., Imura, T., Kakugawa, K., Kitamoto, D., 2008. Efficient production of mannosylerythritol lipids with high hydrophilicity by *Pseudozyma hubeiensis* KM-59. Appl. Microbiol. Biotechnol. 78, 37–46.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736.
Korf, I., 2004. Gene finding in novel genomes. BMC Bioinf. 5.
Kraft, C., Deplazes, A., Sohrmann, M., Peter, M., 2008. Mature ribosomes are selectively degraded upon starvation by an autophagy pathway requiring the Ubp3p/Bre5p ubiquitin protease. Nat. Cell Biol. 10, 602–610.
Kumar, S., Gummadi, S.N., 2011. Metabolism of glucose and xylose as single and mixed feed in *Debaryomyces nepalensis* NCYC 3413: production of industrially important metabolites. Appl. Microbiol. Biotechnol. 89, 1405–1415.
Kurtzman, C., Fell, J.W., Boekhout, T., 2011. The yeasts: a taxonomic study. Elsevier.
Kyriakidou, M., Tai, H.H., Anglin, N.L., Ellis, D., Strömvik, M.V., 2018. Current strategies of polyploid plant genome sequence assembly. Front. Plant Sci. 9, 1660.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–U54.
Lanver, D., Tollot, M., Schweizer, G., Lo Presti, L., Reissmann, S., Ma, L.S., Schuster, M., Tanaka, S., Liang, L., Ludwig, N., Kahmann, R., 2017. *Ustilago maydis* effectors and their impact on virulence. Nat. Rev. Microbiol. 15, 409–421.
Larsson, A., 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30, 3276–3278.
Lee, S.M., Jellison, T., Alper, H.S., 2016. Bioprospecting and evolving alternative xylose and arabinose pathway enzymes for use in *Saccharomyces cerevisiae*. Appl. Microbiol. Biotechnol. 100, 2487–2498.
Li, H., 2016. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics 32, 2103–2110.
Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Proc, G.P.D., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.
Li, Y.-M., Shivas, R.G., Li, B.-J., Cai, L., 2019. Diversity of *Moesziomyces* (Ustilaginales, Ustilaginomycotina) on *Echinochloa* and *Leersia* (Poaceae). MycoKeys. 52, 1.
Liao, Y., Smyth, G.K., Shi, W., 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucl. Acids Res. 41.
Lim, C.S., Weinstein, B.N., Roy, S.W., Brown, C.M., 2021. Analysis of fungal genomes reveals commonalities of intron gain or loss and functions in intron-poor species. Mol. Biol. Evol.
Lippman, S.I., Broach, J.R., 2009. Protein kinase A and TORC1 activate genes for ribosomal biogenesis by inactivating repressors encoded by Dot6 and its homolog Tod6. PNAS 106, 19928–19933.
Liu, H., Wu, S., Li, A., Ruan, J., 2021. SMARTdenovo: a *de novo* assembler using long noisy reads. Gigabyte. 2021, 1–9.
Loman, N.J., Quick, J., Simpson, J.T., 2015. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat. Methods 12, 733–U51.
Lomsadze, A., Burns, P.D., Borodovsky, M., 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucl. Acids Res. 42.
Lorenz, M.C., Heitman, J., 1998. The MEP2 ammonium permease regulates pseudohyphal differentiation in *Saccharomyces cerevisiae*. EMBO J. 17, 1236–1247.
Love, M., Anders, S., Huber, W., 2014. Differential analysis of count data–the DESeq2 package. Genome Biol. 15, 10–1186.
Luque, R., Lovett, J.C., Datta, B., Clancy, J., Campelo, J.M., Romero, A.A., 2010. Biodiesel as feasible petrol fuel replacement: a multidisciplinary overview. Energ. Environ. Sci. 3, 1706–1721.
Ma, Y.Q., Gao, Z., Wang, Q.H., Liu, Y., 2018. Biodiesels from microbial oils: Opportunity and challenges. Bioresour. Technol. 263, 631–641.
Madeira, J.B., Masuda, C.A., Maya-Monteiro, C.M., Matos, G.S., Montero-Lomelí, M., Bozaquel-Morais, B.L., 2015. TORC1 inhibition induces lipid droplet replenishment in yeast. Mol. Cell. Biol. 35, 737–746.
Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. 14.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, 10–12.
Matsushika, A., Inoue, H., Kodaki, T., Sawayama, S., 2009. Ethanol production from xylose in engineered *Saccharomyces cerevisiae* strains: current state and perspectives. Appl. Microbiol. Biotechnol. 84, 37–53.
Mattanovich, D., Sauer, M., Gasser, B., 2014. Yeast biotechnology: teaching the old dog new tricks. Microb. Cell Fact. 13.
Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., Gurevich, A., 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34, 142–150.
Molina-Mora, J.A., Campos-Sanchez, R., Rodriguez, C., Shi, L.M., Garcia, F., 2020. High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 *Pseudomonas aeruginosa* genome: benchmark of hybrid and non-hybrid assemblers. Sci. Rep. 10.
Morin, N., Cescut, J., Beopoulos, A., Lelandais, G., Le Berre, V., Uribelarrea, J.L., Molina-Jouve, C., Nicaud, J.M., 2011. Transcriptomic analyses during the transition from biomass production to lipid accumulation in the oleaginous yeast *Yarrowia lipolytica*. PLoS One 6.
Morita, T., Konishi, M., Fukuoka, T., Imura, T., Kitamoto, H.K., Kitamoto, D., 2007. Characterization of the genus *Pseudozyma* by the formation of glycolipid biosurfactants, mannosylerythritol lipids. FEMS Yeast Res. 7, 286–292.
Morita, T., Koike, H., Hagiwara, H., Ito, E., Machida, M., Sato, S., Habe, H., Kitamoto, D., 2014. Genome and transcriptome analysis of the basidiomycetous yeast *Pseudozyma antarctica* producing extracellular glycolipids, mannosylerythritol lipids. PLoS One 9.
Nogue, V.S., Karhumaa, K., 2015. Xylose fermentation as a challenge for commercialization of lignocellulosic fuels and chemicals. Biotechnol. Lett. 37, 761–772.
Oelkers, P., Cromley, D., Padamsee, M., Billheimer, J.T., Sturley, S.L., 2002. The DGA1 gene determines a second triglyceride synthetic pathway in yeast. J. Biol. Chem. 277, 8877–8881.
Ola, M., O'Brien, C., Coughlan, A., Ma, Q.X., Donovan, P.D., Wolfe, K.H., Butler, G., 2020. Polymorphic centromere locations in the pathogenic yeast *Candida parapsilosis*. Genome Res. 30, 684–696.
Özcan, S., Leong, T., Johnston, M., 1996. Rgt1p of *Saccharomyces cerevisiae*, a key regulator of glucose-induced genes, is both an activator and a repressor of transcription. Mol. Cell. Biol. 16, 6419–6426.
Papanikolaou, S., Aggelis, G., 2011. Lipids of oleaginous yeasts. part I: biochemistry of single cell oil production. Eur. J. Lipid Sci. Technol. 113, 1031–1051.
Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290.
Pinzi, S., Garcia, I.L., Lopez-Gimenez, F.J., de Castro, M.D.L., Dorado, G., Dorado, M.P., 2009. The ideal vegetable oil-based biodiesel composition: a review of social, economical and technical implications. Energy Fuel 23, 2325–2341.
Pomraning, K.R., Kim, Y.-M., Nicora, C.D., Chu, R.K., Bredeweg, E.L., Purvine, S.O., Hu, D., Metz, T.O., Baker, S.E., 2016. Multi-omics analysis reveals regulators of the response to nitrogen limitation in *Yarrowia lipolytica*. BMC Genomics 17, 138.
Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., Finn, R.D., 2018. HMMER web server: 2018 update. Nucl. Acids Res. 46, W200–W204.
Qvirist, L., Mierke, F., Vazquez Juarez, R., Andlid, T., 2022. Screening of xylose utilizing and high lipid producing yeast strains as a potential candidate for industrial application. BMC Microbiol. 22, 173.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ramírez, L., Pérez, G., Castanera, R.l., Santoyo, F., Pisabarro, A.G., 2011. Basidiomycetes Telomeres-A Bioinformatics Approach. Bioinformat.-Trends Methodol.
Ratledge, C., Wynn, J.P., 2002. The biochemistry and molecular biology of lipid accumulation in oleaginous microorganisms. Adv. Appl. Microbiol. 51, 1–51.
Rau, U., Nguyen, L.A., Schulz, S., Wray, V., Nimtz, M., Roeper, H., Koch, H., Lang, S., 2005. Formation and analysis of mannosylerythritol lipids secreted by *Pseudozyma aphidis*. Appl. Microbiol. Biotechnol. 66, 551–559.
Redkar, A., Hoser, R., Schilling, L., Zechmann, B., Krzymowska, M., Walbot, V., Doehlemann, G., 2015. A secreted effector protein of *Ustilago maydis* guides maize leaf cells to form tumors. Plant Cell 27, 1332–1351.
Regenberg, B., Grotkjær, T., Winther, O., Fausbøll, A., Åkesson, M., Bro, C., Hansen, L.K., Brunak, S., Nielsen, J., 2006. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in *Saccharomyces cerevisiae*. Genome Biol. 7, 1–13.
Ribeiro, L.E., Albuini, F.M., Castro, A.G., Campos, V.J., de Souza, G.B., Mendonça, J.G., Rosa, C.A., Mendes, T.A., Santana, M.F., da Silveira, W.B., 2021. Influence of glucose on xylose metabolization by *Spathaspora passalidarum*. Fungal Genet. Biol. 157, 103624.
Romero, I.G., Pai, A.A., Tung, J., Gilad, Y., 2014. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 12.
Ryu, S., Hipp, J., Trinh, C.T., 2016. Activating and elucidating metabolism of complex sugars in *Yarrowia lipolytica*. Appl. Environ. Microbiol. 82, 1334–1345.
Saika, A., Koike, H., Fukuoka, T., Yamamoto, S., Kishimoto, T., Morita, T., 2016. A gene cluster for biosynthesis of mannosylerythritol lipids consisted of 4-O-β-D-mannopyranosyl-(2 R, 3 S)-erythritol as the sugar moiety in a basidiomycetous yeast *Pseudozyma tsukubaensis*. PLoS One 11, e0157858.
Santangelo, G.M., 2006. Glucose signaling in *Saccharomyces cerevisiae*. Microbiol. Mol. Biol. Rev. 70, 253–282.
Scarlat, N., Dallemand, J.F., Monforti-Ferrario, F., Nita, V., 2015. The role of biomass and bioenergy in a future bioeconomy: policies and facts. Environ. Develop. 15, 3–34.
Schirawski, J., Mannhaupt, G., Münch, K., Brefort, T., Schipper, K., Doehlemann, G., Di Stasio, M., Rössel, N., Mendoza-Mendoza, A., Pester, D., 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330, 1546–1548.
Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., Ragg, T., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7, 3.
Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: assessing genome assembly and annotation completeness. Gene prediction. Springer, pp. 227–245.
Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H.E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F.J., Marschall, T., Mayes, S., Costa, V., Zook, J.M., Liu, K.V.J., Kilburn, D., Sorensen, M., Munson, K.M., Vollger, M.R., Monlong, J., Garrison, E., Eichler, E.E., Salama, S., Haussler, D., Green, R.E., Akeson, M., Phillippy, A., Miga, K.H., Carnevali, P., Jain, M., Paten, B., 2020. Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. Nat. Biotechnol. 38, 1044.
Sharma, R., Okmen, B., Doehlemann, G., Thines, M., 2019. Saprotrophic yeasts formerly classified as *Pseudozyma* have retained a large effector arsenal, including functional Pep1 orthologs. Mycol. Prog. 18, 763–768.
Shin, H.Y., Nijland, J.G., de Waal, P.P., de Jong, R.M., Klaassen, P., Driessen, A.J.M., 2015. An engineered cryptic Hxt11 sugar transporter facilitates glucose-xylose co-consumption in *Saccharomyces cerevisiae*. Biotechnol. Biofuels 8.
Soetaert, W., Vandamme, E., 2006. The impact of industrial biotechnology. Biotechnol. J.: Healthcare Nutr. Technol. 1, 756–769.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucl. Acids Res. 34, W435–W439.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550.
Sun, J., Li, R.S., Chen, C., Sigwart, J.D., Kocot, K.M., 2021. Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. Philos. Trans. Royal Soc. B-Biol. Sci. 376.
Tanimura, A., Takashima, M., Sugita, T., Endoh, R., Ohkuma, M., Kishino, S., Ogawa, J., Shima, J., 2016. Lipid production through simultaneous utilization of glucose, xylose, and l-arabinose by *Pseudozyma hubeiensis*: a comparative screening study. AMB Express 6.
Tanimura, A., Sugita, T., Endoh, R., Ohkuma, M., Kishino, S., Ogawa, J., Shima, J., Takashima, M., 2018. Lipid production via simultaneous conversion of glucose and xylose by a novel yeast, *Cystobasidium iriomotense*. PloS One 13, e0202164.
Tao, B.Y., 2007. Industrial applications for plant oils and lipids. Bioprocessing for Value-added Products from Renewable Resources. Elsevier 611–627.
Teichmann, B., Linne, U., Hewald, S., Marahiel, M.A., Bölker, M., 2007. A biosynthetic gene cluster for a secreted cellobiose lipid with antifungal activity from Ustilago maydis. Mol. Microbiol. 66, 525–533.
Templeton, D.W., Scarlata, C.J., Sluiter, J.B., Wolfrum, E.J., 2010. Compositional analysis of lignocellulosic feedstocks. 2. method uncertainties. J. Agric. Food Chem. 58, 9054–9062.
Tesniere, C., Pradal, M., Bessiere, C., Sanchez, I., Blondin, B., Bigey, F., 2018. Relief from nitrogen starvation triggers transient destabilization of glycolytic mRNAs in *Saccharomyces cerevisiae* cells. Mol. Biol. Cell 29, 490–498.
Thorpe, R., Ratledge, C., 1972. Fatty acid distribution in triglycerides of yeasts grown on glucose or n-alkanes. Microbiology 72, 151–163.
Twumasi-Boateng, K., Yu, Y., Chen, D., Gravelat, F.N., Nierman, W.C., Sheppard, D.C., 2009. Transcriptional profiling identifies a role for BrlA in the response to nitrogen depletion and for StuA in the regulation of secondary metabolite clusters in *Aspergillus fumigatus*. Eukaryot. Cell 8, 104–115.
UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucl. Acids Res. 47, D506–D515.
Varemo, L., Nielsen, J., Nookaew, I., 2013. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucl. Acids Res. 41, 4378–4391.
Vaser, R., Sovic, I., Nagarajan, N., Sikic, M., 2017. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. 27, 737–746.
Vijay, V., Pimm, S.L., Jenkins, C.N., Smith, S.J., 2016. The impacts of oil palm on recent deforestation and biodiversity loss. PLoS One 11, e0159668.
Wada, K., Koike, H., Fujii, T., Morita, T., 2020. Targeted transcriptomic study of the implication of central metabolic pathways in mannosylerythritol lipids biosynthesis in *Pseudozyma antarctica* T-34. PLoS One 15.
Wahl, R., Wippel, K., Goos, S., Kämper, J., Sauer, N., 2010. A novel high-affinity sucrose transporter is required for virulence of the plant pathogen *Ustilago maydis*. PLoS Biol. 8, e1000303.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Wang, Q.-M., Jia, J.-H., Bai, F.-Y., 2006. *Pseudozyma hubeiensis* sp. nov. and *Pseudozyma shanxiensis* sp. nov., novel ustilaginomycetous anamorphic yeast species from plant leaves. Int. J. Syst. Evol. Microbiol. 56, 289–293.
Wang, L.G., Wang, S.Q., Li, W., 2012. RSeQC: quality control of RNA-seq experiments. Bioinformatics 28, 2184–2185.
Watanabe, T., Shinozaki, Y., Yoshida, S., Koitabashi, M., Sameshima-Yamashita, Y., Fujii, T., Fukuoka, T., Kitamoto, H.K., 2014. Xylose induces the phyllosphere yeast *Pseudozyma antarctica* to produce a cutinase-like enzyme which efficiently degrades biodegradable plastics. J. Biosci. Bioeng. 117, 325–329.
Wernig, F., Born, S., Boles, E., Grininger, M., Oreb, M., 2020. Fusing α and β subunits of the fungal fatty acid synthase leads to improved production of fatty acids. Sci. Rep. 10, 1–7.
Wick, R.R., Holt, K.E., 2019. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research 8.
Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E., 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial. Genomics 3.
Yamada, R., Yamauchi, A., Kashihara, T., Ogino, H., 2017. Evaluation of lipid production from xylose and glucose/xylose mixed sugar in various oleaginous yeasts and improvement of lipid production by UV mutagenesis. Biochem. Eng. J. 128, 76–82.
Yu, X., Zheng, Y., Xiong, X., Chen, S., 2014. Co-utilization of glucose, xylose and cellobiose by the oleaginous yeast *Cryptococcus curvatus*. Biomass Bioenergy 71, 340–349.
Yuan, T., Ren, Y., Meng, K., Feng, Y., Yang, P., Wang, S., Shi, P., Wang, L., Xie, D., Yao, B., 2011. RNA-Seq of the xylose-fermenting yeast *Scheffersomyces stipitis* cultivated in glucose or xylose. Appl. Microbiol. Biotechnol. 92, 1237–1249.
Zhang, W., Du, G., Zhou, J., Chen, J., 2018. Regulation of sensing, transportation, and catabolism of nitrogen sources in Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev. 82, e00040-17.
Zhao, L., Zhang, X., Tan, T., 2008. Influence of various glucose/xylose mixtures on ethanol production by *Pachysolen tannophilus*. Biomass Bioenergy 32, 1156–1161.
Zhu, Z.W., Zhang, S.F., Liu, H.W., Shen, H.W., Lin, X.P., Yang, F., Zhou, Y.J.J., Jin, G.J., Ye, M.L., Zou, H.F., Zhao, Z.B.K., 2012. A multi-omic map of the lipid-producing yeast *Rhodosporidium toruloides*. Nat. Commun. 3.
Zhu, Z.W., Zhou, Y.J.J., Krivoruchko, A., Grininger, M., Zhao, Z.B.K., Nielsen, J., 2017. Expanding the product portfolio of fungal type I fatty acid synthases. Nat. Chem. Biol. 13, 360.
Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics 29, 2669–2677.
Zimin, A.V., Salzberg, S.L., 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput. Biol. 16.