A draft genome of the ascomycotal fungal species *Pseudopithomyces maydicus* (family *Didymosphaeriaceae*)

Krithika Arumugam[1], Sherilyn Ho[2], Irina Bessarab[3], Falicia Q. Y. Goh[2], Mindia A. S. Haryono[3], Ezequiel Santillan[1], Stefan Wuertz[1,4], Yvonne Chow[2], Rohan B. H. Williams[3,]*

[1]Singapore Centre for Environmental Life Sciences Engineering (SCELSE), Nanyang Technological University, Singapore

[2]Singapore Institute of Food and Biotechnology Innovation (SIFBI), Agency for Science, Technology and Research (A*STAR), Singapore.

[3]Singapore Centre for Environmental Life Sciences Engineering (SCELSE), National University of Singapore, Singapore

[4]School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

*Corresponding author: Rohan Williams (lsirbhw@nus.edu.sg)

For submission to *Microbology Resource Announcements*

Abstract

We report a draft genome of the ascomycotal fungal species Pseudopithomyces maydicus (isolate name SBW1) obtained using a culture isolate from brewery wastewater. From a 22 contig assembly, we predict 13502 protein coding gene models, of which 4389 (32.5%) were annotated to KEGG Orthology and identify 39 biosynthetic gene clusters.

Announcement

Pseudopithomyces maydicus is a fungal species within phylum *Ascomycota* (order *Pleosporales*), previously named *Pithomyces maydicus* and recently renamed with the introduction of genus *Pseudopithomyces* into family *Didymosphaeriaceae* (Ariyawansa *et al.*, 2015). Members of this species have been identified as potential human pathogens, based on identification from several clinical specimens (da Cunha *et al.*, 2014), and natural products from this species have recently been characterized, some of which hold antimicrobial activity (Ningsih *et al.*, 2021). Neither *Pseudopithomyces maydicus* nor any member of genus *Pseudopithomyces* have a reference or draft genome available, with the most relevant related genome sequence data being 11 draft genomes collected from other genera in family *Didymosphaeriaceae* (NCBI Assembly, accessed 2022/03/31).

We report a draft genome of *Pseudopithomyces maydicus* (isolate name SBW1) obtained from a culture isolate from brewery wastewater in Singapore. Genomic characterization of microbes isolated from food-processing wastewater can provide foundational data for biotechnological applications in the circular economy, such as the production of microbial protein (Vethathirri *et al.* 2021).

We obtained an isolate from brewery wastewater by culturing on solid Yeast Extract–Peptone–Dextrose (YPD) agar media at 30^{o}C for 2 days. The colony was isolated and streaked out on a new YPD plate. Taxonomic classification was made via Sanger sequencing of the D1/D2 domain of the large-subunit (28S) ribosomal DNA (NCBI BLASTN webserver against the NCBI nr/nt database in megablast mode with top hit annotated to a *P. maydicus* partial 28S sequence, MF919633.1, at 99% identity; see **Supplementary Figure 1** for alignment; https://blast.ncbi.nlm.nih.gov/Blast.cgi, executed on 06/31/2022). Genomic DNA was extracted using liquid nitrogen and mechanical grinding of the fungal culture followed by application of the Qiagen DNeasy PowerSoil Pro Kit. 1.5µg of input DNA was subjected to shearing (Megaruptor3, Diagenode Inc, Denville, NJ, USA; operated for 20kb target at speed 35) and then 800ng sheared DNA was used to construct a sequencing library using the SQK-LSK109 ligation sequencing kit (Oxord Nanopore Technologies Ltd, Oxford, UK), barcoded using the EXP-NDB104 native barcoding kit (Oxford Nanopore Technologies; barcode 10). Following construction, 200 ng of the library was sequenced on a GridION instrument (Oxford Nanopore Technologies; release 21.05.20) for 72 hours. Basecalling was performed using Guppy 5.0.13 (Oxford Nanopore Technologies) in SUP mode.

The run generated 233,209 raw reads from the cognate barcode (232,644 reads following the application of Porechop version 0.2.4 using default parameters except --discard_middle, -t 20) (Porechop, 2018), comprising a total of 1.53 Gbp of sequence. Genome assembly was performed using Flye version 2.9 (using parameters --nano-hq, -t 44) (Kolmogorov *et al.*, 2019). A total of 36 contigs were obtained with a total sequence length of 39,781,613 bp.

Based on visualization of per-contig GC content, mean coverage and length, we determined a working draft of the genome to be comprised of 22 contigs (mean length 1,792,464 bp, range: 81,297- 3,886,452 bp), with an N50 of 2,331,148 bp and a total sequence length of 39,434,212 bp. The mean GC content was 0.5 (range: 0.48-0.51) and mean coverage was 35 (range: 33-36). (**Supplementary Figure 2** and **Supplementary Data File 1**). One high coverage circular contig (37,662 bp; contig 40) was aligned to the mitochondrial genome of the closely related species *Pseudopithomyces chartarum* (97% nucleotide identity with 80% query coverage to Genbank KY792993.1; annotated to *Pithomyces chartarum*, but refer Ariyawansa *et al.*, 2015, for discussion of reassignment to genus *Pseudopithomyces*). The remaining 13 contigs held substantially lower mean GC content values than those observed from the draft genome. These 13 contigs accounted for 309,739 bp of sequence (mean: 23,826 bp, median: 8,979 bp, range: 2897-126,848 bp), and were considered to arise from potentially mis-assembled telomeric or repeat sequences and/or sequences from intra-plate contaminants (**Supplementary Figure 2)**.

The quality of the draft genome was examined using gene-level analysis with BUSCO package (v5.3.2; exeuted in genome mode using the lineage dataset for *Pleosporales*; pleosporales_odb10) (Manni *et al.*, 2021). BUSCO identified 5637 complete marker genes (of 6641 searched), of which 5610 were complete and single copy, 27 complete and duplicated, 231 were fragmented and 733 missing (BUSCO notation: C:84.9% [S:84.5%, D:0.4%], F:3.5%, M:11.6%, *n*:6641).

From the draft genome, five 18S SSU-rRNA genes were predicted with RNAmmer (Lagesen *et al.*, 2007) all of which annotated to order *Pleosporales* using the SILVA Alignment, Classification and Tree Service (ACT) (Pruesse *et al.*, 2012) (**Supplementary Data File 2**). Three 28S LSU-rRNA genes were recovered from the assembly, to which the partial 28S sequence obtained above aligned with 99% identity (BLASTN, run with default settings; Camacho *et al.*, 2009; alignments provided in **Supplementary Data File 3**). We note these ribosomal gene numbers are less than expected based on recent estimates (Lofgren *et al.*, 2019) made within ensembles of complete fungal genomes and may be related to limited reconstructability of closely related DNA fragments harbouring ribosomal operons. Further taxonomic analysis was undertaken using sourmash (Pierce *et al.*, 2019), comparing the draft genome against all 9563 fungal genomes downloadable from the NCBI (2022/03/29), with the 7 most similar genomes observed to be members of family *Didymosphaeriaceae* (**Supplementary Data File 4**). Collectively these results are consistent with the original taxonomic assignment, within the limits of fungal genome availability for closely related fungal groups.

To gain some insight into possible chromosomal structures, all assembled contig sequences were searched for more than three or more repeat units of exact matches to CCCTAA and TTAGGG motifs (Rahnama et al., 2021) using the find function in a text editor to identify possible telomeric regions. Of the 22 contigs in the draft genome, 4 contigs have multiple repeat units of CCCTAA and 8 contigs have multiple repeat units of TTAGGG at the 5' and 3' end respectively. 3 contigs have both CCCTAA and TTAGGG at the 5' and 3' end (**Supplementary Data File 1)** suggesting these sequences may represent distinct chromosomes.

An initial catalogue of gene models, predicted using GeneMark-ES (run with --fungus flag set) (Ter-Hovhannisyan *et al.*, 2008), was comprised of 13502 protein coding genes, of which 4389 (32.5%) were annotated to one or more KEGG Orthology identifiers using BlastKOALA (Kanehisa *et al.*, 2016; **Supplementary Data File 5**). A total of 162 tRNA encoding genes were predicted using EuFindtRNA search algorithm in tRNAscan-SE (version 2.0, running default parameters) (Lowe and Eddy, 1997).

Recently Ningsih *et al.* (2021) isolated and characterized seven natural product compounds from an isolate of *P. maydicus* isolated from marine bryozoan (genus *Schizoporella*). To identify potentially-related biosynthetic gene clusters, we analysed the recovered genome sequence using the biosynthetic gene cluster (BGC) finder antiSMASH6 (Blin *et al.*, 2021). In total we identified 39 BGCs, comprised of 19 Type 1 polyketide synthase (T1PKS) clusters and 14 non-ribosomal peptide synthetase (NRPS) or NRPS-like clusters, 4 terpene encoding clusters and two indole encoding clusters (**Supplementary Results**). Further examination of the relationships between these detected BGCs and the compounds defined by Ningsih *et al.* (2021) may provide insight into the relevant biosynthesis pathways for these specalised metabolites, as recently highlighted by Louwen and van der Hooft (2021).

References

da Cunha KC, Sutton DA, Gené J, Cano J, Capilla J, Madrid H, Decock C, Wiederhold NP, Guarro J (2014). Pithomyces species (Montagnulaceae) from clinical specimens: identification and antifungal susceptibility profiles. *Medical Mycology* **52** (7): 748-57. https://doi.org/10.1093/mmy/myu044

Ningsih BNS, Rukachaisirikul V, Pansrinun S, Phongpaichit S, Preedanon S, Sakayaroj J (2021). New aromatic polyketides from the marine-derived fungus *Pseudopithomyces maydicus* PSU-AMF350 and their antimicrobial activity. *Natural Product Research*. **27**: 1-8. https://doi.org/10.1080/14786419.2021.1915309

Vethathirri RS, Santillan E, Wuertz S (2021). Microbial community-based protein production from wastewater for animal feed applications. *Bioresoure Technology* **341**:125723. https://doi.org/10.1016/j.biortech.2021.125723

Porechop: https://github.com/rrwick/Porechop

Ariyawansa HA, Hyde KD, Jayasiri SC *et al.* (2015). Fungal diversity notes 111–252—taxonomic and phylogenetic contributions to fungal taxa. *Fungal Diversity* **75**, 27–274 (2015). https://doi.org/10.1007/s13225-015-0346-5.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**(5): 540-546. https://doi.org/10.1038/s41587-019-0072-8

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38** (10): 4647-4654. https://doi.org/10.1093/molbev/msab199

Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**(9):3100-8. https://doi.org/10.1093/nar/gkm160

Lofgren LA, Uehling JK, Branco S, Bruns TD, Martin F, Kennedy PG (2019). Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. *Molecular Ecology* **28**(4): 721-730. doi: 10.1111/mec.14995.

Pruesse E, Peplies J, Glöckne FO (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, **28**, 1823-1829. https://doi.org/10.1093/bioinformatics/bts252

Camacho, C., Coulouris, G., Avagyan, V. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). https://doi.org/10.1186/1471-2105-10-421

Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. (2019). Large-scale sequence comparisons with sourmash. *F1000Res.* **8**: 1006. https://doi.org/10.12688/f1000research.19675.1

Rahnama M, Wang B, Dostart J, Novikova O, Yackzan D, Yackzan A, Bruss H, Baker M, Jacob H, Zhang XF, Lamb A, Stewart A, Heist M, Hoover J, Calie P, Chen L, Liu J, Farman ML (2021). Telomere roles in fungal genome evolution and adaptation. *Frontiers in Genetics* **12**: 676751. https://doi.org/10.3389/fgene.2021.676751

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008). Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Research* **18** (12): 1979-90. https://doi.org/10.1101/gr.081612.108

Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology* **428**, 726-731. https://doi.org/10.1016/j.jmb.2015.11.006

Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**: 955-964. https://doi.org/10.1093/nar/25.5.955

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49** (W1): W29-W35. https://doi.org/10.1093/nar/gkab335

Louwen JJR, van der Hooft JJJ (2021). Comprehensive large-scale integrative analysis of omics data to accelerate specialized metabolite discovery. *mSystems* **6** (4): e0072621. https://doi.org/10.1128/msystems.00726-21

Chan PP, Lin BY, Mak AJ, Lowe TM (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes, *Nucleic Acids Research* **49**(16): 9077-9096. https://doi.org/10.1093/nar/gkab688.

Disclosure

R.B.H.W is a scientific cofounder and equity holder at BluMaiden Biosciences Pte Ltd, a Singaporean biotech platform company engaged in drug discovery from human microbiomes.

Acknowledgements

We thank colleagues at the Integrated Genomics Platform at the Genome Institute of Singapore for performing DNA sequencing and basecalling. This research was supported by the Singapore National Research Foundation (NRF) and Ministry of Education under the Research Centre of Excellence Program and the NRF Competitive Research Programme (NRF-CRP21-2018-0006) "*Recovery and microbial synthesis of high-value aquaculture feed additives from food-processing wastewater*". We thank Dr Chang Ying (National University of Singapore) for her critical reading of the manuscript.

Author contributions

Y.C, S.W and R.B.H.W each conceptualized separate components of this work as part of a broader collaborative research programme. E.S. coordinated and project managed the broader project, and obtained and characterized source wastewater samples. S.H, F.Q.Y.G and Y.C isolated and cultured the microbe, performed taxonomic classification and extracted genomic DNA. I.B. performed further characterization of genomic DNA and coordinated sequencing. K.A., M.A.S.H and R.B.H.W performed data analysis. All authors were involved in data interpretation. The manuscript was written by R.B.H.W with inputs from other authors.

Data Availability Statement

The draft genome sequence has been deposited at DDBJ/ENA/GenBank under the accession JANTUC000000000. The version described in this paper is version JANTUC010000000. Raw sequence data are pending release at the time of writing.

List of Supplementary Material

Supplementary Figure 1

Sequence alignment between Sanger-sequenced partial 28S LSU-rRNA sequence and the top ranked BLASTN hit from NCBI nr/nt database.

Supplementary Figure 2

Pairs plot for contig GC-content, contig coverage and contig length from the *P. maydicus assembly*.

Supplementary Data File 1

Table listing properties of contigs from the *P. maydicus* assembly.

Supplementary Data File 2

Summary of taxonomic classification analysis of recovered 18S SSU-rRNA sequences to the SILVA 138 database.

Supplementary Data File 3

Alignment of Sanger-sequenced partial 28S LSU-rRNA sequence against three 28S LSU-rRNA gene sequences recovered from the *P. maydicus* long read genome assembly and a set of 62 28S LSU-rRNA sequences from members of genus *Psuedopithomyces* (NCBI Nucleotide searched for “Pseudopithomyces AND 28S" on 30th May 2022).

Supplementary Data File 4

MASH similarity statistics obtained by comparing the *P. maydicus* long read genome assembly sequence to 9563 fungal genomes obtained from NCBI. The reference genomes from NCBI were downloaded using the NCBI ‘dataset’ (version 13.6.0) command line tool (datasets_13.6.0 download genome taxon 4751 --filename fungi.zip --assembly-level complete_genome,chromosome,scaffold,contig --exclude-gff3 --exclude-protein --exclude-rna).

Supplementary Data File 5

BlastKOALA annotation data for all proteins predicted from *P. maydicus* long read assembly.

Supplementary Results

Complete output from the antiSMASH6 analysis of the *P. maydicus* long read assembly.