Egyptian Journal of Medical
Human Genetics

RESEARCH **Open Access**

In silico identification of prospective virulence factors associated with candidiasis in *Meyerozyma guilliermondii* strain SO from genome dataset

Robiatul Azilah Zainudin[1,2], Suriana Sabri[1,3], Abu Bakar Salleh[1], Arpah Abu[4], Raja Farhana Raja Khairuddin[5] and Siti Nurbaya Oslan[1,2*]

Abstract

Background *Meyerozyma guilliermondii* is a prospective yeast that has extensively contributed to the biotechnology sector. In 2015, *M. guilliermondii* strain SO which was isolated from spoiled orange has successfully been developed as an inducer-free expression system and attained a significant impact in producing industrially important recombinant proteins. The species possesses high similarity to *Candida albicans* which may cause candidiasis. The industrial-benefiting *M. guilliermondii* strain SO has been underexplored for its virulence status. Thus, this study aimed to document the potential virulence factors through the comprehensive in silico analysis of *M. guilliermondii* strain SO genome. This analysis demonstrated the molecular characterization which could distinguish the pathogenicity status of *M. guilliermondii*.

Results The genome data were generated from Illumina HiSeq 4000 sequencing platform and assembled into 51 scaffolds successfully accumulating a genome size of 10.63 Mbp. These enclosed 5,335 CDS genes and 5,349 protein sequences with 43.72% GC content. About 99.29% of them were annotated to public databases. *Komagataella phaffii*, *Saccharomyces cerevisiae* and the reference strain of *M. guilliermondii* (ATCC 6260) were used as the controls. They were compared with our in-house strain SO to identify the consensus domain or subdomain which could putatively be considered as virulence factors. *Candida albicans* was used as the pathogenic model. Hence, hidden Markov model against strain SO proteome had identified secreted aspartic proteases (*SAP*), phospholipase C (*PLC*) and phospholipase D (*PLD*) with an E-value of 2.4e^{-107}, 9.5e^{-200} and 0.0e^{+00}, respectively, in resemblance of *C. albicans*. The topology of the phylogenetic analysis indicated that these virulence factors in *M. guilliermondii* strain SO and *C. albicans* branched from the same node and clustered together as a clade, signifying their molecular relatedness and congeneric among these species, subsequently proposing the virulence status of *M. guilliermondii*.

Conclusion The *SAP*, *PLC* and *PLD* genes' features that were significant in expressing determinants of pathogenicity were successfully identified in *M. guilliermondii* strain SO genome dataset, thus concluding the virulency of this species. On account of this finding, the strategy of gene knockout through CRISPR-Cas9 or homologous recombination strategies is needed to engineer the feasible novel expression host system. Over and above, the genetically modified strain of *M. guilliermondii* allegedly may eradicate the risk of candidiasis infection.

*Correspondence:
Siti Nurbaya Oslan
snurbayaoslan@upm.edu.my
Full list of author information is available at the end of the article

Keywords Virulence factors, Whole genome sequencing, Yeast expression system, *Meyerozyma guilliermondii*

Background

The universal inventory of microbial diversity by a congruence of the dominance scaling law was estimated upward of 1 trillion species [1]. Meanwhile to date, the assessment of global fungal diversity was predicted to range from 2.2 to 3.8 million species based on the improved statistical and molecular phylogenetic approach [2]. Of those fractions, about 650 species were reported as pathogens to cause human diseases [3] and relatively 150 species were discovered as opportunistic fungi to immunocompromised individuals [4].

Invasive candidiasis is a life-threatening fungal infection that is commonly caused by *Candida* spp. [5]. *Candida albicans* is the most predominant species that has immense potential of spreading and infecting human from superficial mucosal to systemic candidiasis [6, 7], indisputably, progressively demonstrating morbidity and mortality worldwide [8, 9]. The main pathogenic feature in *Candida* that obligates persistency in the human host is the mechanism of biofilm formation [10]. The yeast-to-hypha morphology transition manifests adherence ability to attach and invade human epithelial cells henceforth triggering pathogenesis [11]. The proteolytic activity from the proteinase enzyme is involved in pathogenicity [12] and possibly carries a virulence factor including secreted aspartyl proteinase (*SAP*) and phospholipases [13].

Candida guilliermondii (teleomorphic: *Pichia guilliermondii*) is currently renowned as *Meyerozyma guilliermondii* [14]. This fungi species is associated with Ascomycota phylum under *Saccharomycotina* CTG clade, where the genetic CUG codon is reassigned from leucine to serine [15, 16]. Natural habitat of this species is highly diverse in clinical and environmental samples, but most frequently found in oil-containing soil [17]. A strain of *M. guilliermondii* was isolated from a spoiled orange and identified as strain SO (GenBank JN084128) [18]. It had been developed as a prospective system for heterologous protein expression, providing an alternative to the intensively used species, *Komagataella phaffii* [19] (also referred to as *Pichia pastoris* [20]).

Veritably, this novel strain is capable of producing recombinant enzymes such as lipase [18], protease, diamine oxidase [21] and α-amylase [22]. Moreover, the competency of this microbial organism to facilitate expression mediated by alcohol oxidase (*AOX*) and formaldehyde dehydrogenase (*FLD*) promoters successfully demonstrates the independent commencement of mRNA transcription within a shorter time in the absence of any inducer such as methanol or methylamine [22]. The achievement may obliquely reduce the production cost, minimize methanol toxicity effects and would further innovate the technology of enzyme research. However, this potential yeast has been reported as a causative pathogen and associated with candidiasis [23]. An emergence incidence of candidiasis recently shifts toward non-*albicans Candida* spp. that demonstrates low intrinsic susceptibility to antimycotic drugs [24], subsequently denotes this opportunistic yeast [25–27] and probably associated with virulence determinants in *C. albicans*.

In fact, *M. guilliermondii* aligns the phylogenetic branch closely to *C. albicans*, signifying an affiliation among these two sensu stricto species [28]. True hyphae are absent in *M. guilliermondii* [29], and the haploid features of chromosome may exhibit less virulence in systemic infection model [30]. The composition of cell wall is also dissimilar, therefore exaggerating human innate immune cells to boost the production of cytokine and phagocytosis in dectin-1-dependent pathway, where *C. albicans* is incompetent to be recognized by the host due to its macromolecule structure [31]. The thickness of *M. guilliermondii* cell wall is about 160–170 nm [21] in comparison to the cell wall of *C. albicans* (approximately 400 nm) which is two times greater [32], emphasizing the establishment of defense mechanisms by the latter species [33]. Virulence factor proteins (enzymes) that can substantially be found in yeast are secreted aspartic proteases (*SAP*), phospholipases, lipases, agglutinin-like sequences (Als), 70-kDa heat shock protein, enolases and phytases [34]. In-depth microbiological characterization of isolates with *M. guilliermondii* candidiasis has produced biofilms, but with low metabolic activity and moderate biomass, supported by clinical presentation in human with the predisposing condition is less severe and lower mortality as compared to *C. albicans* [35].

Nevertheless, the production of virulence properties in *M. guilliermondii* may be detected by enzymatic activities of esterase (phospholipase C and D) and aspartic protease [36]. A comparative study of toxicity between *Saccharomyces cerevisiae* and *M. guilliermondii* on zebrafish embryos had shown 0% survival rate within 120 h post-exposure to the latter strain and the breakdown of lipid membrane revealed the activity of phospholipase enzyme in *M. guilliermondii*, meanwhile, the absence of scoliosis and 20% mortality rate of embryos treated with *S. cerevisiae* strengthened the status of *S. cerevisiae* as a non-toxic yeast [37]. Hence, it is critical to recognize the source of factors that may invoke the safety standards of the yeast expression system, especially yeasts that may possess a high similarity of pathogenic properties to *C. albicans*. To date, there has been limited published research documented on the virulence properties in *M. guilliermondii* as per commonly inspected in pathogenic fungal species like *Candida* spp. [38]. Thus, in this work, the presumed *Candida*-like virulence factors were identified in *M. guilliermondii* strain SO genome dataset and subsequently analyzed using in silico approaches according to the presence of domain SAP, phospholipase C (*PLC*) and phospholipase D (*PLD*), similar to those in *C. albicans*.

Results and discussion

Genomic features of *M. guilliermondii* strain SO

The draft genome of *M. guilliermondii* strain SO was generated using Illumina HiSeq 4000 system with a total throughput of 10.6 Mbp. Across the whole genome, the structure of 142 tRNA and 3 rRNA (8S; 18S; 28S) were detected by HMM model. The dataset comprising of 69 contigs which 52% of it were larger than 1 kbp and the largest contigs racked up to 1.34 Mbp. The final assembled contigs were merged into 51 scaffolds with GC content of 43.72%. About 5,372 CDS were predicted according to fungal-specific intron organization in GeneMark-ES algorithm (Table 1), covering 78.1% of the entire genome.

Table 1 Assembly statistics of *M. guilliermondii* strain SO genome features demonstrated from de novo Velvet program

Features	Value
Raw reads	20,587,778
Clean reads	19,277,985
Scaffolds	51
Contigs	69
Genome size (bp)	10,634,970
Scaffold N50 (bp)	1,679,168
Scaffold N75 (bp)	999,727
GC content (%)	43.72
tRNA	142
rRNA	3
Total number of genes (≥ 99 bp)	5,349
Hits to NCBI	5,297 (99.03%)
Hits to Swiss-Prot	4,512 (84.35%)
Hits to GO	4,734 (88.50%)
Hit to EC	1,613 (30.16%)
Hit to KEGG	966 (18.06%)

The annotation of genes was identified up to 99% homolog in public databases, and 88.5% of the sequences was classified in GO categories (Additional file 1: Fig. 1). Concurrently, 3,495 domains in protein families (pfam) database locally aligned to 4,597 (85.9%) protein sequences of *M. guilliermondii* strain SO (Additional file 1: Fig. 2). In addition, a total of 966 predicted proteins were assigned to 851 EC number and subsequently mapped to 134 KEGG pathway maps as per updated on April 2021.

Approximately, 1,918 (35.8%) of protein queries in *M. guilliermondii* strain SO dataset were assigned to the KEGG network-based system. Furthermore, KofamKOALA was performed to complement the analysis and identified 64.4% genes subset to 502 modules that further linked to 6 major classes (metabolism pathway, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases) and secondarily, associated with 53 KEGG networks (Fig. 1).

The initial verification of the genome set examined using BUSCO Assessment succeeded to inspect 97% of the assembly completeness subjected to the catalog of conserved single-copy orthologs from Ascomycota dataset in OrthoDB. From the assessment of predicted gene models, all 1,315 BUSCO orthologs are accounted within the predicted gene sets, albeit a few of those orthologs are fragmented. Gene prediction analysis was carried out adequately as most predicted single-copy orthologs were presented in the genome (Additional file 1: Table 1).

Prospective virulence factor in putative aspartic proteases encoding gene

Proteolytic enzymes or also defined as proteases, peptidases and proteinases catalyze the breakdown of peptide bonds in proteins through hydrolysis reaction. Aspartic proteases (EC 3.4.23.24) classified under the group of enzymes candidapepsin possess virulence determinants through the degradation of subendothelial extracellular matrix, albumin, immunoglobulin G, hemoglobin, fibronectin and laminin in the host components [39]. Specifically named as eukaryotic aspartyl protease, this domain architecture can be found in 11,794 sequences in the Ensembl public database (verified in April 2021). These catalytic residues are α/β monomer that is composed of bilobed shape, positioned within the hallmark motif Asp-Thr/Ser-Gly to the active site based on the InterPro HMM database. The eight genes of aspartic proteases were identified in *M. guilliermondii* strain SO with encoded ID's; 1971_t, 1972_t, 4064_t, 4978_t, 4979_t, 4980_t, 70_t and 999_t (Additional file 1: Table 2). From the eukaryotic protease homolog, the closest identity to *M. guilliermondii* strain SO profile besides the other

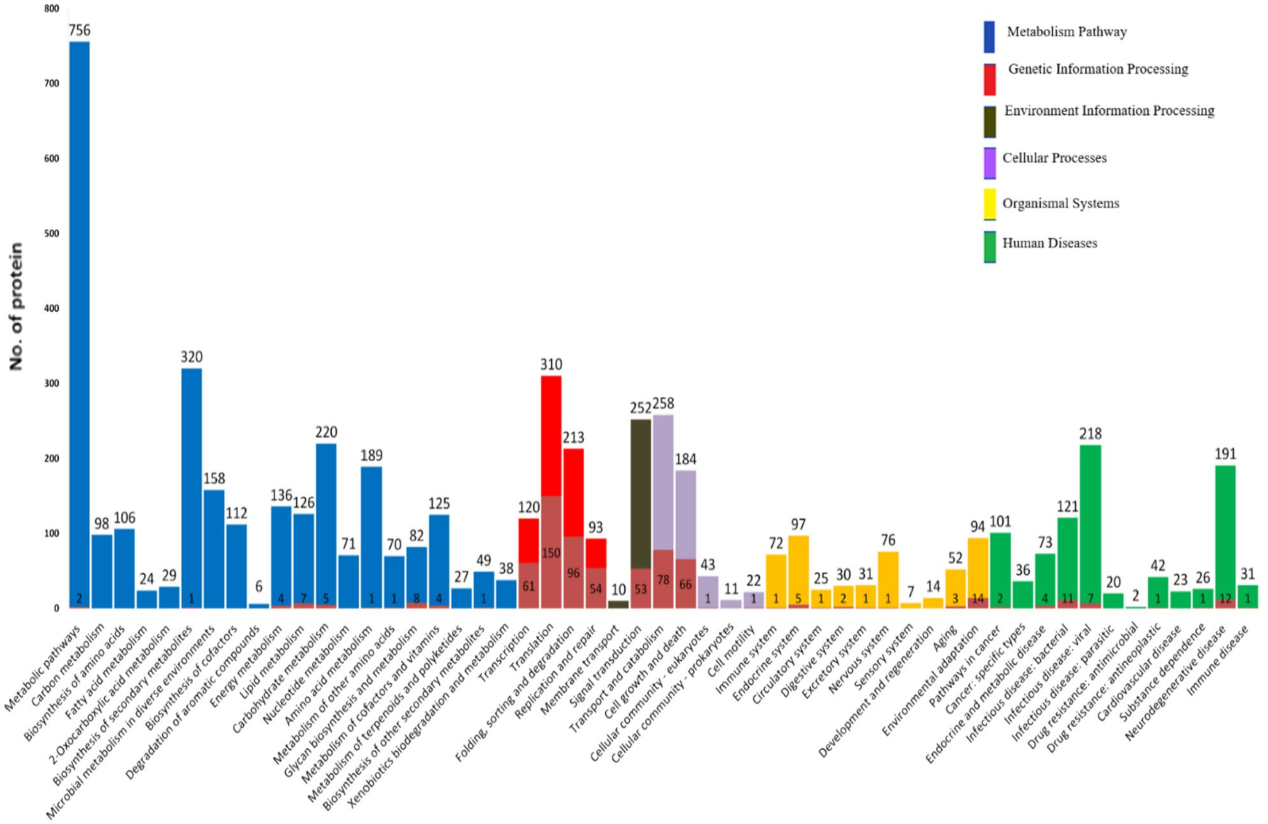

Fig. 1 Tabulation of *M. guilliermondii* strain SO proteins into 53 interconnected KEGG pathway. The stacked on the column base indicates solely hits of annotation to the respected pathway maps

Table 2 HMM homologs hits on *M. guilliermondii* strain SO eukaryotic aspartyl protease profiles

Target	Species	*E*-value
PGUG_03957	*Meyerozyma guilliermondii* (strain ATCC 6260 / CBS 566 / DSM 6381 / JCM 1539 / NBRC 10279 / NRRL Y-324)	3.2e^{-168}
HYPBUDRAFT_158240	*Hyphopichia burtonii* NRRL Y-1933	4.9e^{-98}
CANTADRAFT_52809	*Suhomyces tanzawaensis* NRRL Y-17324	2.9e^{-99}
PICST_63754	*Scheffersomyces stipitis* (strain ATCC 58785 / CBS 6054 / NBRC 10063 / NRRL Y-11545)	1.0e^{-106}
CANTEDRAFT_115143	*Candida tenuis* (strain ATCC 10573 / BCRC 21748 / CBS 615 / JCM 9827 / NBRC 10315 / NRRL Y-1498 / VKM Y-70)	1.3e^{-99}
METBIDRAFT_31852	*Metschnikowia bicuspidata var. bicuspidata* NRRL YB-4993	1.4e^{-70}
MG3_02996	*Candida albicans* P78048 (GCA_000773725)	2.4e^{-107}
BON23_0850	*Saccharomyces cerevisiae* str. 131 (GCA_001983315)	7.5e^{-120}
AT250_GQ6804937	*Komagataella phaffii* (strain GS115 / ATCC 20864)	6.1e^{-65}

strains of *M. guilliermondii* are the proteases from *K. phafii, S. cerevisiae* and *C. albicans* with the E-value 6.1e^{-65}, 7.5e^{-120} and 2.4e^{-107}, respectively (Table 2).

Overall, aspartic proteases have limited homology except for the active site region approximately at residue 78 to 271. This profile is involved in aspartic-type endopeptidase for proteolysis activity. Figure 2 shows eight genes of strain SO slightly diversed between one another and dispersed to the same branch of *M. guilliermondii* ATCC 6260, *C. albicans, K. phaffii* and *S. cerevisiae* separately. It could be defined that the gene 70_t, 1971_t and 1972_t were prospective to be the virulence-associated protease based on 37.45% to 43.19% identical to the pathogenic yeast *C. albicans*

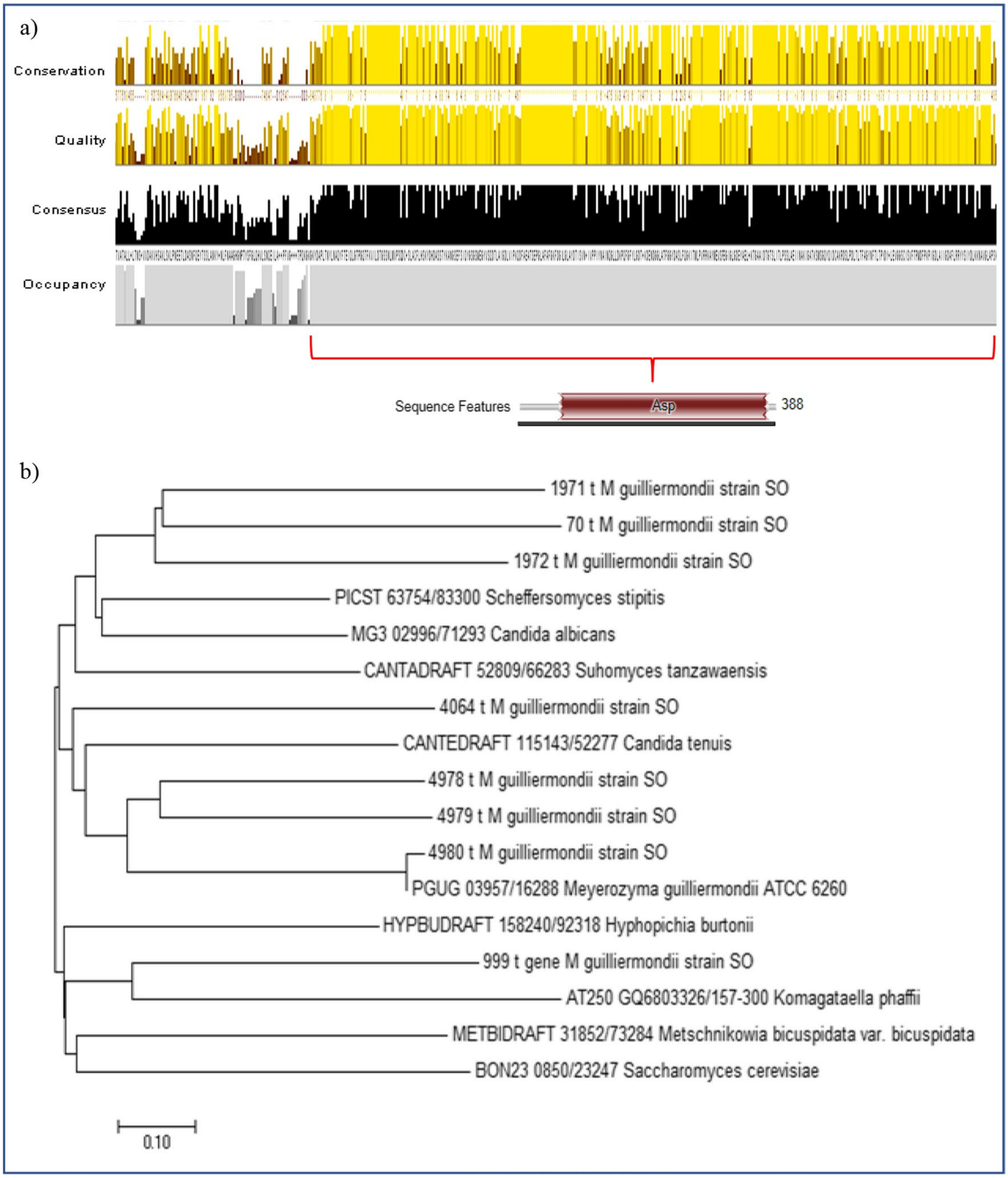

Fig. 2 Profiling of putative aspartic proteases encoding gene according to public databases to signify virulence factors as indicated pathogenic in *C. albicans*. **a** The consensus region of aspartic proteases from HMM hits coordinate represented using BioEdit and **b** the phylogenetic tree of *M. guilliermondii* strain SO from HMM protease model. Three genes of *M. guilliermondii* strain SO (1971_t, 1972_t and 70_t) shared connected elements with *C. albicans*

with 99% query cover in BLAST analysis. But, as compared to *Scheffersomyces stipitis* which converged from the same taxon of *C. albicans*, it can be classified as non-pathogenic species [40]. Whereas, the enzymatic activity from 999_t gene which was similar to the non-pathogenic species, *S. cerevisiae* and *K. phaffii* might exert less toxicity. Proteases genes denoted as 4064_t, 4978_t, 4979_t, and 4980_t were in parallel to the protease motif of *Candida tenuis* and its reference genome, *M. guilliermondii* ATCC 6260. The quality of consensus and conservation region generated by BioEdit also exhibited the data proficiency.

Prospective virulence factor in putative phospholipases encoding gene

The construction of HMM analyses was based on the curated architecture of the phospholipase domain which was annotated to the domain of lysophospholipase catalytic (family PLA2_B; PF01735.19), phosphoinositide phospholipase C (family PI-PLC-X; PF00388.20), phosphatidylinositol-specific phospholipase C, Y domain (family PI-PLC-Y; PF00387.20), phospholipase D (family PLDc; PF00614.23) and PLD-like domain (family PLDc_2; PF13091.7). Pfam database had discovered six genes of *M. guilliermondii* strain SO denoted to the enzyme belonging to this superfamily (Additional file 1: Table 3). The family of hydrolases consisting of lysophospholipase (phospholipase B) and cytosolic phospholipase A2 (included the domain type PLA2_B) were found in the *M. guilliermondii* strain SO proteome at the gene ID 1966_t and 913_t. However, based on prior study, the production of esterase enzymes carried virulence properties in *M. guilliermondii* [36]. Thus, the family of phospholipase B, which is not classified as a group of esterase enzyme, was withdrawn from further investigation. Nevertheless, both *PLC* and *PLD* are rationally considered to play roles in pathogenicity because the mechanism is contributed to Candida virulence factors [40]. These putative lytic genes possibly become a specific target site for downstream analysis in facilitating gene modification.

Phospholipase C

Phospholipase C (*PLC*) mediates in assigning signal transduction intracellularly and intercellularly by cleaving phospholipids in eukaryotic organisms. The family member of this enzyme has been characterized into thirteen different isoform and contributes to specific cellular function [41]. The gene that encodes phosphatidylinositol-specific phospholipase C (*PI-PLC*) in *S. cerevisiae* is PLC1 which is also an ortholog to the gene *CAPLC1* in *C. albicans* delta-form isoenzyme and consists of the associated of catalytic domain X and Y [42]. Nevertheless, the phospholipase C identified in *C. albicans* does not contain N-terminal signal peptides which probably secrete intracellular phospholipases and are unidentical to the other classes of phospholipases [43].

PLC breaks the glycerophosphate bond at position before phosphate in glycerophospholipid molecule to produce diacylglycerol and a phosphate-containing head group. Thirteen isoenzymes are categorized into six subfamilies (β, γ, δ, ε, ζ and η) according to their structures [44]. The secretion of phosphoinositide phospholipase C (EC 3.1.4.11) constitutes an inositol signaling pathway and is involved in the invasion and penetration of host barrier cells [45]. These phosphatidylinositol-specific phospholipase C proteins contain highly conserved X and Y catalytic domains coordinated at residue 499 to 644 and 703 to 818, respectively. The order of these domains is (–NH2–X–Y–COOH–), nevertheless, the spacing may fluctuate according to species. This domain architecture is identified in 1,140 sequences in the Ensembl database (as per analyzed in April 2021) and is responsible for lipid metabolic process, signal transduction and intracellular signal transduction.

The genes denoted as 3887_t in *M. guilliermondii* strain SO proteome was verified by hmmscan to annotate PLC-1 enzyme. The specific hits for gene 3887_t (length size: 2.8 kbp) showed a high confident association to be inferred as PLC domain X. The MSA of this gene to the PI-PLC gene of *C. albicans* P37005 (Genbank: KGQ83690.1) had shown 42% identity that covered 88% at $0.0e^{+00}$ E-value. The ORF of this gene mapped to the conserved domain comprised of PI-PLC, EF-hand and C2 superfamilies region (Additional file 1: Fig. 3). The alignment results also showed high identity of the gene to the pathogenic *C. albicans* and contained complete structure of domain X and Y in phosphatidylinositol-specific phospholipase C.

Phmmer and hmmscan were used as homolog finder to identify the motif pattern of PLC in 3887_t gene against reference proteome database and pfam database, respectively, with default E-value cutoff, 0.01 and BLOSUM62 scoring matrix substitution. From the HMM searching algorithm, *PLC* that targeted on *Hyphopichia burtonii, Scheffersomyces stipites* and *Suhomyces tanzawaensis* were the most significant to *M. guilliermondii* strain SO according to the E-value. Notably, the coverage hits of these two domains (PI-PLC-X and PI-PLC-Y) consensus motif was also not established consistently in the conserved region, supported clearly by incongruent quality, conservation and consensus (Fig. 3). The phylogenetic tree specifying the relatedness of *PLC* enzyme in *M. guilliermondii* strain SO in comparison with *S. cerevisiae* had shown two lineages that split taxon from ancestral node, as well as the latter species grouped together to the other non-pathogenic yeast, *K. phaffi*. The constitution of *PLC* properties in *M. guilliermondii* strain SO and *C. albicans* might need to be considered as toxic to the host as attributed in the topography scale that corresponded to share similar bases which might impute to have virulence factors (Fig. 3).

Phospholipase D

Phospholipase D (*PLD*) hydrolyzes the phosphodiester bond of glycerolipid phosphatidylcholine, yielding phosphatidic acid and free choline. The enzymes are

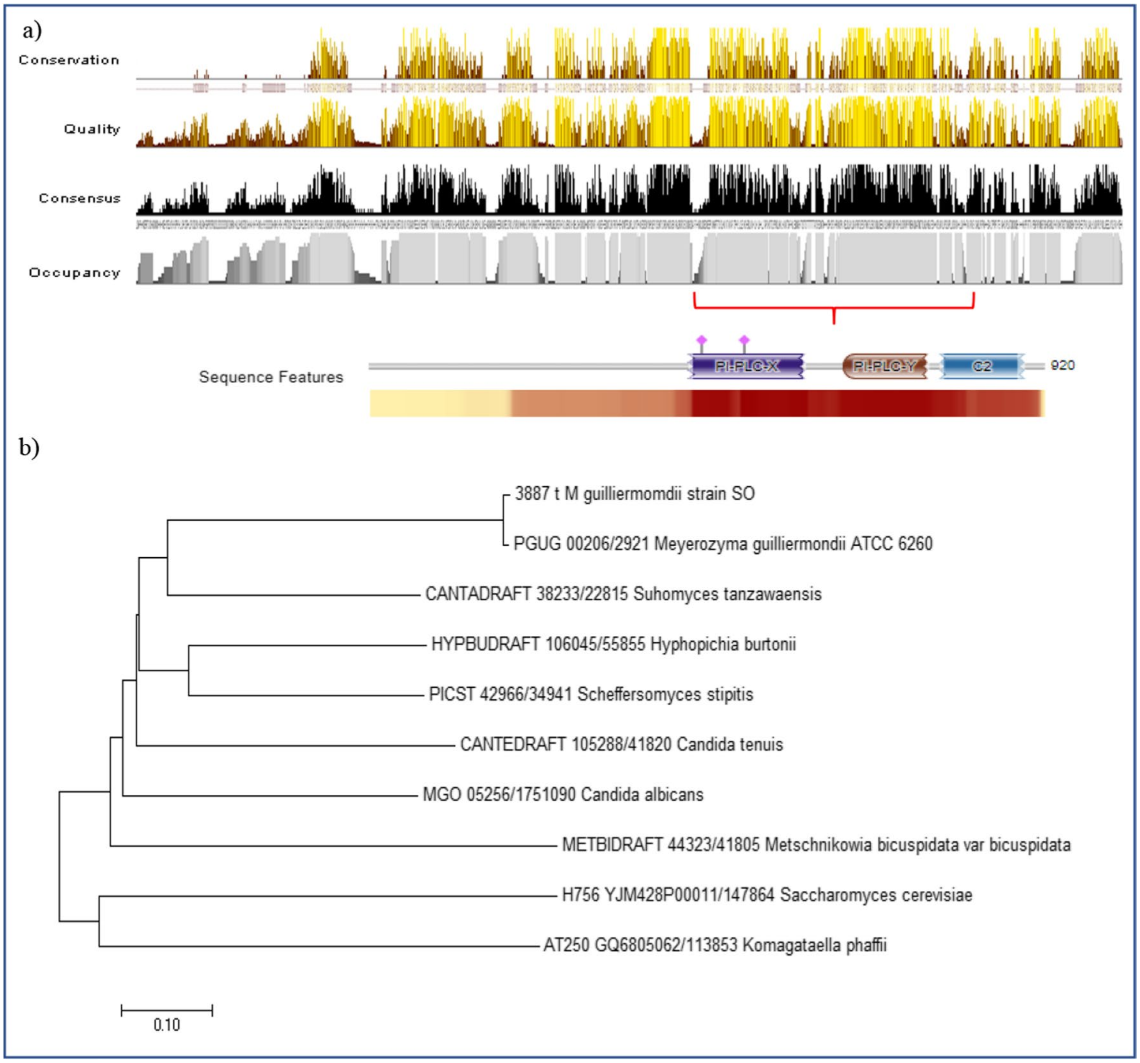

Fig. 3 Profiling of putative phospholipase C encoding gene according to public databases to signify virulence factors as indicated pathogenic in *C. albicans*. **a** The consensus region of phospholipase C from HMM hits coordinate represented using BioEdit and **b** the phylogenetic tree of *M. guilliermondii* strain SO from HMM phospholipase C model

important in modifying membrane lipids in eukaryotes and the production of phosphatidic acid signifies the signal of gene activity [46]. Two isoforms of *PLD*; *PLD1* and *PLD2* have been reported [47]. *PLD1* (EC 3.1.4.4) stimulates the morphogenic transition from yeast to hyphal growth and plays the role as a modifier of membrane structure and function [48]. These phosphatidylcholine-hydrolyzing PLD isoforms possess a homolog of phosphatidylserine synthase and cardiolipin synthase, which apparently displayed the occurrence of two motifs containing well-conserved lysine, histidine and/or asparagine residues, and this may subsequently contribute to the active site aspartic acid as in sequence feature (Additional file 1: Fig. 4).

The proteomic data of *M. guilliermondii* strain SO had distinguished the gene with sequence ID 5149_t annotated as *PLD1* supported by the output analysis from hmmscan. The following verification was using the same platform as *PLC*, and recognized the molecular structure of *PLD* identified by the motif of catalytic sequence; HxKxxxxD (HKD), represented by the abbreviations of amino acids histidine (H), lysine (K), aspartic acid (D), while x represents non-conservative amino acids [49].

Verification on public databases using BLASTX had found the gene of 5149_t, length size 5.2 kbp nucleotide, hit specifically on *PLD* in *C. albicans* P34048 (GenBank: KGU31558.1) that covered 69% of its gene with 55% identity and zero E-value. This gene was also mapped into conserved protein domain family PLN02866, phospholipase D, that would eventually construct the tertiary structure of this protein and perform the enzyme function. According to the constructed ORF, 5149_t gene fits the specification for toxicity analysis as it scores 1590 bits to *C. albicans*. The catalytic *PLD* motif was identified as "HEKLCVID" in *M. guilliermondii* strain SO and this is parallel with the subject strain of *C. albicans*, "HEKLCIID." (Table 3).

The analysis of phmmer showed that the homolog of PLD genes from subject species hit perfectly to the target gene, 5149_t *M. guilliermondii* strain SO with E-value $0.0e^{+00}$, concluding high identity motif for this domain family (Table 4) where 550 catalytically significant residues of this domain architecture were detected to conserve at reading frame 5' to 3' nucleotide position 757 to 1094 in HMM execution against Ensembl database. Similar to the phylogenetic tree constructed for *PLC*, *PLD* enzyme also clustered in one clade indicating that homologous structure of these proteins was inherited with minimum evolution (Fig. 4). From the analysis, it is fair to conclude that in silico findings showed close relation of virulence factor from *PLD* enzyme in *M. guilliermondii* and *C. albicans.* Hence, the suppression of this gene in *M. guilliermondii* strain SO may alleviate or revert the pathogenic effects for further development as a host.

Regardless of how the identification of virulence factors was conveyed from this analysis, there is still a requirement to evaluate *M. guilliermondii* strain SO through in vitro analysis. Assessment on quantifying the level of expression is recommended to certify the level of pathogenicity or toxicity corresponding to the enzymatic activity of those genes. In order to accelerate the research, one of the advancements in quantifying the transcription of those above-mentioned virulence factors is using NanoString technology which detects the target genes digitally in the samples of input material besides microarray analysis. Nevertheless, for the development of strain SO to host a heterologous expression system, the corresponding virulence factor genes are proposed to be knocked out to repress or inhibit the operational function of these enzymes through CRISPR-Cas9 advancement technology. The method is currently having a wide

Table 3 HMM homologs hits on *M. guilliermondii* strain SO phospholipase C

Target	Species	*E*-value
PGUG_00206	*Meyerozyma guilliermondii* (strain ATCC 6260 / CBS 566 / DSM 6381 / JCM 1539 / NBRC 10279 / NRRL Y-324)	$0.0e^{+00}$
HYPBUDRAFT_106045	*Hyphopichia burtonii* NRRL Y-1933	$6.6e^{-238}$
CANTADRAFT_38233	*Suhomyces tanzawaensis* NRRL Y-17324	$4.2e^{-233}$
PICST_42966	*Scheffersomyces stipitis* (strain ATCC 58785 / CBS 6054 / NBRC 10063 / NRRL Y-11545)	$1.1e^{-237}$
CANTEDRAFT_105288	*Candida tenuis* (strain ATCC 10573 / BCRC 21748 / CBS 615 / JCM 9827 / NBRC 10315 / NRRL Y-1498 / VKM Y-70)	$4.3e^{-218}$
METBIDRAFT _44323	*Metschnikowia bicuspidata var. bicuspidata* NRRL YB-4993	$6.4e^{-175}$
MGO_ 05,256	*Candida albicans* P76055 (GCA_000784505)	$9.5e^{-200}$
H756_YJM428P00011	*Saccharomyces cerevisiae* YJM428 (GCA_000975945)	$1.4e^{-104}$
AT250_GQ6805062	*Komagataella phaffii* (strain GS115 / ATCC 20864)	$4.1e^{-129}$

Table 4 HMM homologs hits on *M. guilliermondii* strain SO phospholipase D

Target	Species	E-value
PGUG_05838	*Meyerozyma guilliermondii* (strain ATCC 6260 / CBS 566 / DSM 6381 / JCM 1539 / NBRC 10,279 / NRRL Y-324)	$0.0e^{+00}$
HYPBUDRAFT_152158	*Hyphopichia burtonii* NRRL Y-1933	$0.0e^{+00}$
CANTADRAFT_38489	*Suhomyces tanzawaensis* NRRL Y-17324	$0.0e^{+00}$
PICST_81277	*Scheffersomyces stipitis* (strain ATCC 58785 / CBS 6054 / NBRC 10063 / NRRL Y-11545)	$0.0e^{+00}$
CANTEDRAFT_92504	*Candida tenuis* (strain ATCC 10573 / BCRC 21748 / CBS 615 / JCM 9827 / NBRC 10315 / NRRL Y-1498 / VKM Y-70)	$0.0e^{+00}$
METBIDRAFT_73296	*Metschnikowia bicuspidata var. bicuspidata* NRRL YB-4993	$0.0e^{+00}$
MGQ_01118	*Candida albicans* P76067 (GCA_000784495)	$0.0e^{+00}$
H763_YJM554K00257	*Saccharomyces cerevisiae* YJM554	$0.0e^{+00}$
PP7435_CHR3-0554	*Komagataella phaffii* (strain ATCC 76273 / CBS 7435 / CECT 11047 / NRRL Y-11430 / Wegner 21-1)	$0.0e^{+00}$

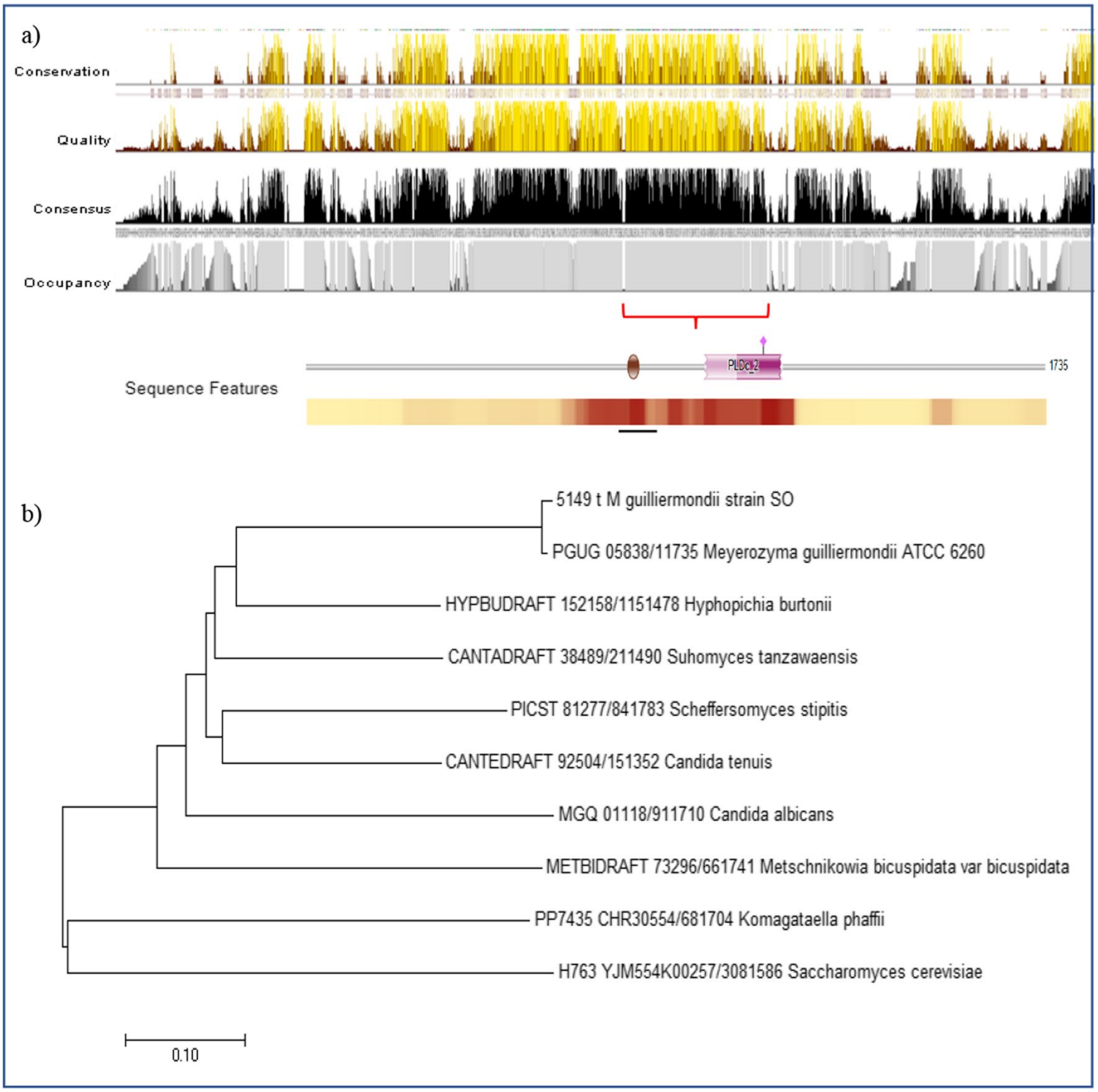

Fig. 4 Profiling of putative phospholipase D encoding gene according to public databases to signify virulence factors as indicated pathogenic in *C. albicans*. **a** The consensus region of phospholipase D from HMM hits coordinate represented using BioEdit and **b** the phylogenetic tree of *M. guilliermondii* strain SO from HMM phospholipase D model

range of potential, yet precise and versatile in genomic editing by removing, adding or altering the virulence domain section of DNA sequence, subsequently, qualifying GRAS standard.

Conclusion

Complete information on the genome *M. guilliermondii* strain SO may be used optimally to complement this recently developed expression system and supports the future genetic manipulation for the recombinant protein production industry. Prior to the commercialization of this methanol-free expression host, the previously available draft genome of *M. guilliermondii* strain ATCC 6260 should not be solely relied. The proteomic data of *M. guilliermondii* strain SO were deployed to comprehend the fundamental knowledge regarding virulence factors in this species. In silico toxicology analysis in identifying the presence of potential virulence factors; phospholipases and proteases protein would specify the potential traits according to the genome annotations, subsequently predicting the adverse effects implicated by the fungal pathogenicity. Resemblance molecular features of virulence factors in the most clinically isolated *C. albicans* were identified in *M. guilliermondii* strain SO proteome. The recognition algorithm programs delivered by the hidden Markov model had shown the properties of those selected virulence determinants with low E-value to *C. albicans*, indicating that these factors might elicit virulence for candidiasis infection and toxicity as a yeast expression host. The finding supported by the construction of neighbor-joining (NJ) phylogenetic tree was confirmed with 1000 bootstrapping.

Materials and methods

Yeast isolation and DNA extraction

The sample of wild-type yeast cell, *M. guilliermondii* strain SO was obtained from a previous study [18]. The cryopreserved cells in 30% (v/v) glycerol were streaked on YPD agar (Millipore Sigma, USA), prepared from (in w/v): 1% yeast extract, 2% peptone, 2% dextrose and 2% bacteriological agar, then incubated at 30 °C for 3 days. A single colony of the yeast cell revived on the plate was cultivated in YPD medium broth containing (in w/v): 1% yeast extract, 2% peptone and 2% dextrose, and was grown overnight in an incubator shaker at 30 °C, 250 rpm. The genomic DNA extraction was carried out using Wizard Genomic DNA Purification Kit (Promega, USA). A 600 µl of 0.5 M phosphate buffer, pH 7.5 (dipotassium phosphate and monopotassium phosphate) was added to the pellet cells from 10 mL culture and mixed gently with 50 U/g of Zymolyase$^{®}$-100 T enzymes (Nacalai Tesque, Japan) and 1% 14 mM β-mercaptoethanol (v/v) before incubating for 2 h at 37 °C. The next procedure followed the protocol supplied by the manufacturer. The final step is rehydration of DNA; the samples, however, were eluted with 10 mM Tris–Cl pH 8.5 without EDTA to avoid under-tagmentation by enzymatic inhibitors. Quantification using NanoDrop One Microvolume UV–Vis Spectrophotometers (Thermo Fisher Scientific, USA) determined the concentration of starting material contains 93.6 ng/µl in 24.52 µg total mass of DNA with purity $A_{260/280}$ at 1.9. The specificity assessment of genomic DNA was precisely measured using Qubit 4 Fluorometer (Thermo Fisher Scientific, USA) at 93.0 ng/µl and visualized the nucleic acid fragments in agarose gel electrophoresis (Additional file 1: Fig. 5).

Whole genome sequencing (WGS) analysis

The purified genomic DNA of *M. guilliermondii* strain SO was sent offsite to Codon Genomics Sdn. Bhd. in Selangor, Malaysia, for second-generation sequencing using Illumina instruments. The sequencing of *M. guilliermondii* strain SO genome was performed approximately at 146 × coverage accumulating 20.588 million raw reads containing 3.088 Gb of data from both paired-end (short insert; 300 bp) and mate-pair (long insert; 5 Kbp) library construction. The quality of data throughput was verified using FASTQC [50] and FASTX-Toolkit [51] before downstream analysis with 91% reads longer than 140 bp retained after filtering quality (Additional file 1: Fig. 6, 7). The de novo assembly was computed by Velvet assembler [52] using 99 k-*mer* length (Additional file 1: Fig. 8). The scaffolding and gap-filling assemblies were carried out by SSPACE and Gapfiller, respectively, for further improvement in the genome arrangement. Structural prediction of tRNA and rRNA were identified by tRNAscan-SE v1.3.1 [53] and rRNAmmer v1.2 [54], respectively. HMM spotter model was applied to detect rRNA (8S; 18S; 28S) genes across the whole genome. While covariance model search was applied to differentiate pseudogenes from true RNA where the score of covels exceeds 20.0 bit. Homology search algorithm, BLASTX [55] annotated the gene function from public protein databases, RefSeq National Center for Biotechnology Information (NCBI) reference sequence and Swiss-Prot (SP) curated database at default cutoff E-value (database downloaded at 05th Jan 2018). A gene prediction program, GeneMark-ES [56] employed hidden state Markov model (HSMM) approaches to predict the region of noncoding and coding (CDS) genes. The domains of retrieved protein were identified from protein family (pfam) database and subsequently interconnected to Gene Ontology (GO), KEGG Orthology (KO) and Enzyme Commission (EC) knowledge bases (Additional file 1: Fig. 9). The initial affirmation of the genome set was examined using BUSCO Assessment against the catalog of 1,315 conserved single-copy orthologs from Ascomycota dataset in OrthoDB. Concomitantly, in order to synchronize the annotation, GhostKOALA was generated to infer its molecular functional cluster according to KEGG ORTHOLOGY (KO), collaterally, reconstruct KEGG maps, BRITE hierarchies and KEGG modules. KofamKOALA was assigned to annotate based on hmmsearch and provided the identification of strain SO genes. The completeness of genome assembly was quantified using Benchmarking Universal Single-Copy Orthologs (BUSCO) software [57]. Assessment of genome assembly with a technical measure of N50 only inspected the quality of the assembly; however, did not assess assembly completeness in terms of gene content. On the other hand, BUSCO provides measures for quantitative assessment of genome assembly completeness based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB. The assembled draft genome of strain SO was subjected to completeness assessment in BUSCO using the Ascomycota dataset which comprised 1,315 conserved single-copy orthologs.

Furthermore, the full repertoire of peptide sequences (≥ 33 aa) from predicted genes in FASTA file format was verified by BUSCO to ensure the completeness of gene set.

The probabilities of candida-like virulence factors to be identified in *M. guilliermondii* strain SO genome dataset were analyzed according to the presence of aspartic proteases, phospholipase C and phospholipase D. The reference genome data of *M. guilliermondii*, strain ATCC6260 were included for interspecies comparison. *P. pastoris* or its binomial name *K. phaffi* and *S. cerevisiae* were used as a GRAS control model [58] while *Candida albicans* was used as a reference for pathogenic model for yeast [59].

Identification of virulence factors in *M. guilliermondii* strain SO

The proteomic profiles of *M. guilliermondii* strain SO were conferred initially to alternative yeast nuclear code (known as translation table 12) [60] and the regions that mapped significantly to virulence factor enzymes of aspartic proteases and phospholipases, then retrieved it in FASTA format. Subsequently, the annotated profiles were proceeded to multiple sequence alignments (MSA) using MUSCLE build-in provided by MEGAX platform to generate highly conserved polymorphic sequences. Meanwhile for the genes without consensus profiles, reverse PSI_BLAST approach was applied to BLASTX with a default cutoff value lower than 0.01 to ensure those determinant genes were accurately selected for toxicity analysis. The output with hits to the domain of selected virulence factors was pooled and interpreted the adequacy by implying the bit-score, the query coverage, E-value and percentage of identity. The alignment of *M. guilliermondii* strain SO genes that matched the query enzyme was visually checked according to the score colors. Furthermore, the architecture of the hit subject was verified whether the interval hits performed within specific or non-specific boundaries of superfamily domain, which defined the structural homology of the gene.

Hidden Markov model analysis

The identification of those selected genes in *M. guilliermondii* strain SO were analyzed in silico using hidden Markov model (HMM) approach based on the proteome database gathered from prior analysis. Profile HMMs implemented in HMMER was used to perform the interactive searches against the queries from *M. guilliermondii* strain SO proteome database and next, proposed the probabilistic models to predict the similar homologs on selected reference databases. In this study, hmmsearch was used to configure the similarity of sequences that matched a multiple sequence alignment of domain proteases. While phmmer and hmmscan were executed to a single query protein like phospholipase C and phospholipase D. For the identification of match features, hmmscan utilized the HMM database from pfam with default cutoff parameter. The sequence database ensembled genomes (v.44) was selected as a target to correspond to this analysis. Phylogenetic tree was constructed inferring the Neighbor-Joining (NJ) method [61] to compute the distance matrix of protease enzymes among these species. The predicted virulence factor genes in *M. guilliermondii* strain SO were executed using HMMER and phmmer algorithms to configure the pathogenic domain in other species. Henceforth, pfam database was used as a reference to annotate those domains and subsequently visualized the output analysis in BioEdit and MegaX.

Abbreviations

AOX	Alcohol oxidase
CDS	Coding region sequences
EC	Enzyme commission
E-value	Expected value
FLD	Formaldehyde dehydrogenase
GC content	Guanine–cytosine content
GO	Gene ontology
GRAS	Generally recognized as safe
HMM	Hidden Markov model
KEGG	Kyoto encyclopedia of genes and genomes
Kbp	Kilobase pair
Mbp	Megabase pair
mRNA	Messenger RNA
MSA	Multiple sequence alignment
nm	Nanometer
pfam	Protein families
PLC	Phospholipase C
PLD	Phospholipase D
rRNA	Ribosomal ribonucleic acid
SAP	Aspartic proteases
tRNA	Transfer ribonucleic acid

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43042-023-00384-3.

Additional file 1: Supplementary tables and figures.

Acknowledgements

This study was supported by Graduate Research Fellowship (GS46691) and Geran Inisiatif Putra Siswazah from Universiti Putra Malaysia (GP- IPS/2016/9513300).

Author contributions

The authors confirm their contribution to the paper as follows: RAZ, SS, ABS, AA, RFRK and SNO contributed to conception or design of the work; RAZ collected the data; RAZ and RFRK analyzed and interpreted the data; RAZ, RFRK and SNO drafted the article; RAZ, SS, ABS, RFRK and SNO critically revised the article; RAZ, SS, ABS, RFRK and SNO provided final approval of the version to be published. All authors read and approved the final manuscript.

Funding
This work, including the efforts of Siti Nurbaya Oslan, was funded by Geran Inisiatif Putra Siswazah from Universiti Putra Malaysia (GP- IPS/2016/9513300).

Availability of data and materials
A whole genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VMSI00000000 and Sequence Read Archive (SRA) under BioProject accession number PRJNA547962. The version described in this paper is the second version, VMSI01000000 (https://www.ncbi.nlm.nih.gov/nuccore/VMSI00000000).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Enzyme and Microbial Technology Research Centre (EMTech), Centre of Excellence, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia. [2]Department of Biochemistry, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia. [3]Department of Microbiology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia. [4]Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. [5]Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjung Malim, Perak, Malaysia.

Received: 11 September 2022 Accepted: 2 January 2023
Published online: 12 January 2023

References

1. Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. Proc Natl Acad Sci 113(21):5970–5975
2. Hawksworth DL, Lücking R (2017) Fungal diversity revisited: 2.2 to 3.8 million species. Microbiology Spectrum. 5(4):4–5
3. de Hoog GS, Ahmed SA, Danesi P, Guillot J, Gräser Y (2018) Distribution of pathogens and outbreak fungi in the fungal kingdom. In: Seyedmousavi S, de Hoog GS, Guillot J, Verweij P (eds) Emerging and epizootic fungal infections in animals. Springer, Dordrecht, pp 3–16
4. Hyde KD, Abdullah MSA, Andersen B, Boekhout T, Buzina W, Dawson TL Jr, Eastwood DC, Jones EBG, de Hoog S, Kang Y, Longcore JE, Richard-Forget F, Stadler M, Theelen B, Thongbai B, Tsui CKM (2018) The world's ten most feared fungi. Fungal Diversity 93:161–194
5. Atiencia-Carrera MB, Cabezas-Mera FS, Tejera E, Machado A (2022) Prevalence of biofilms in *Candida* spp bloodstream infections: a metaanalysis. PLoS ONE 17(2):e0263522
6. Lopes JP, Lionakis MS (2022) Pathogenesis and virulence of *Candida albicans*. Virulence 13(1):89–121
7. Bates S (2008) Pathogenic fungi: insights in molecular biology. Expert Rev Anti Infect Ther 6(5):591–592
8. Kainz K, Bauer MA, Madeo F, Carmona-Gutierrez D (2020) Fungal infections in humans: the silent crisis. Microbial Cell 7(6):143–145
9. Xiao Z, Wang Q, Zhu F, An Y (2019) Epidemiology, species distribution, antifungal susceptibility and mortality risk factors of candidemia among critically ill patients: a retrospective study from 2011 to 2017 in a teaching hospital in China. Antimicrob Resist Infect Control 8:89
10. Cavalheiro M, Teixeira MC (2018) *Candida* biofilms: threats, challenges, and promising strategies. Front Med (Lausanne) 5:28
11. Lo HJ, Kohler JR, DiDomenico B, Loebenberg L, Cacciapuoti A, Fink GR (1997) Nonfilamentous C albicans mutants are avirulent. Cell 90:939–949
12. Nunes CS, Kumar V (2018) Enzymes in human and animal nutrition: principles and perspectives. Academic Press 267–277
13. Calderone RA, Fonzi WA (2001) Virulence factors of *Candida albicans*. Trends Microbiol 9(7):327–335
14. Kurtzman CP, *Meyerozyma* Kurtzman & M, Suzuki (2010) In Kurtzman CP, Fell JW, Boekhout T, eds. The yeast. 5th ed. London, Elsevier; 2011; p 621–624.
15. Dujon B (2010) Yeast evolutionary genomics. Nat Rev Genet 11(7):512–524
16. Santos MA, Gomes AC, Santos MC, Carreto LC, Moura GR (2011) The genetic code of the fungal CTG clade. CR Biol 334(8):607–611
17. Sibirny AA, Boretsky YR (2009) Pichia guilliermondii. In: Satyanarayana T, Kunze G (eds) Yeast biotechnology: diversity and applications. Springer, Dordrecht, pp 113–134
18. Oslan SN, Salleh AB, Rahman RNZRA, Basri M, Chor ALT (2012) Locally isolated yeasts from Malaysia: identification phylogenetic study and characterization. Acta Biochim Pol 59(2):225–229
19. Oslan SN, Salleh AB, Rahman RNZRA, Leow TC, Sukamat H, Basri M (2015) A newly isolated yeast as an expression host for recombinant lipase. Cell Molecul Biol Lett 20(2):279–293
20. Valli M, Tatto NE, Peymann A, Gruber C, Landes N, Ekker H, Thallinger GG, Mattanovich D, Gasser B, Graf AB (2016) Curation of the genome annotation of Pichia pastoris (Komagataella phaffii) CBS7435 from gene level to protein function. FEMS Yeast Res 16(6):fow051
21. Mahyon NI (2017) Structural investigation of alcohol oxidase from *Meyerozyma guilliermondii* and the use of its promoter for recombinant protein expression. Master's thesis. Universiti Putra Malaysia
22. Nasir NSM, Leow CT, Oslan SNH, Salleh AB, Oslan SN (2020) Molecular expression of a recombinant thermostable bacterial amylase from *Geobacillus stearothermophilus* SR74 using methanol-free *Meyerozyma guilliermondii* strain SO yeast system. BioResources 15(2):3161–3172
23. Castillo-Bejarano JI, Tamez-Rivera O, Mirabal-García M, Luengas-Bautista M, Montes-Figueroa AG, Fortes-Gutiérrez S, González-Saldaña N (2020) Invasive candidiasis due to *Candida guilliermondii* complex: epidemiology and antifungal susceptibility testing from a third-level pediatric center in Mexico. J Pediatric Infectious Diseases Soc 9(3):404–406
24. Deorukhkar SC, Saini S, Mathew S (2014) Non-albicans *Candida* infection: an emerging threat. Interdisciplinary perspectives on infectious diseases
25. Girmenia C, Pizzarelli G, Cristini F, Barchiesi F, Spreghini E, Scalise G, Martino P (2006) *Candida guilliermondii* fungemia in patients with hematologic malignancies. J Clin Microbiol 44(7):2458–2464
26. Pfaller MA, Diekema DJ, Gibbs DL, Newell VA, Ellis D, Tullio V, Rodloff A, Fu W, Ling TA (2010) Results from the ARTEMIS DISK global antifungal surveillance study 1997 to 2007: a 10.5-year analysis of susceptibilities of *Candida* species to fluconazole and voriconazole determined by CLSI standardized disk diffusion. J Clinic Microbiol 48(4):1366–1377
27. Tseng TY, Chen TC, Ho CM, Lin PC, Chou CH, Tsai CT, Wang JH, Chi CY, Ho MW (2017) Clinical features, antifungal susceptibility and outcome of *Candida guilliermondii* fungemia: an experience in a tertiary hospital in mid-Taiwan. J Microbiol Immunol Infect 51:552–558
28. Fitzpatrick DA, Logue ME, Stajich JE, Butler G (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol 6:99
29. Castanheira M, Woosley LN, Diekema DJ, Jones RN, Pfaller MA (2013) *Candida guilliermondii* and other species of *Candida* misidentified as *Candida famata*: assessment by vitek 2, DNA sequencing analysis, and matrix-assisted laser desorption ionization-time of flight mass spectrometry in two global antifungal surveillance programs. J Clin Microbiol 51(1):117–124
30. Fan S, Li C, Bing J, Huang G, Du H (2020) Discovery of the diploid form of the emerging fungal pathogen *Candida auris*. ACS Infectious Diseases 6(10):2641–2646
31. Navarro-Arias MJ, Hernández-Chávez MJ, Garcia-Carnero LC, Amezcua-Hernandez DG, Lozoya-Perez NE, Estrada-Mata E, Martinez-Duncker I, Franco B, Mora-Montes HM (2019) Differential recognition of *Candida tropicalis, Candida guilliermondii, Candida krusei,* and *Candida auris* by human innate immune cells. Infection Drug Resistance 12:783–794
32. Mukherjee S, Mukherjee N, Saini P, Gayen P, Roy P, Babu SPS (2014) Molecular evidence on the occurrence of co-infection with *Pichia guilliermondii* and *Wuchereria bancrofti* in two filarial endemic districts of India. Infectious Disease Poverty 3(13):1–10

33. Ruiz-Herrera J, Elorza MV, Valentín E, Sentandreu R (2006) Molecular organization of the cell wall of *Candida albicans* and its relation to pathogenicity. FEMS Yeast Res 6(1):14–29
34. Lim SJ, Mohamad Ali MS, Sabri S, Muhd Noor ND, Salleh AB, Oslan SN (2021) Opportunistic yeast pathogen *Candida* spp.: secreted and membrane-bound virulence factors. Med Mycol 59(12):1127–1144
35. Marcos-Zambrano LJ, Puig-Asensio M, Pérez-García F, Escribano P, Sánchez-Carrillo C, Zaragoza O, Padilla B, Cuenca-Estrella M, Almirante B, Martín-Gómez MT, Muñoz P, Bouza E, Guinea J (2017) *Candida guilliermondii* complex is characterized by high antifungal resistance but low mortality in 22 cases of candidemia. Antimicrob Agents Chemother 61(7):e00099-e117
36. Chaves ALS, Trilles L, Alves GM, Figueiredi-Carvalho MHG, Brito-Santos F, Coelho RA, Martins IS, Almeida-Paes R (2020) A case-series of bloodstream infections caused by the *Meyerozyma guilliermondii* species complex at a reference center of oncology in Brazil. Med Mycol 59(3):235–243
37. Radzi SNF (2020) Toxicity studies of *Meyerozyma guilliermondii* strain SO using zebrafish as a model. Universiti Putra Malaysia, Malaysia
38. Santos ALS, Soares RM (2005) *Candida guilliermondii* isolated from HIV-infected human secretesa 50 kDa serine proteinase that cleaves a broad spectrum of proteinaceous substrates. FEMS Immunol Med Microbiol 43(1):13–20
39. Estevez SV, Armitage A, Bates HJ, Harrison RJ, Buscaino A (2021) The genome of the CTG (Ser1) yeast S*cheffersomyces stipitis* is plastic. Am Soc Microbiol J 12(5):e1817
40. Ghannoum MA (2000) Potential role of phospholipases in virulence and fungal pathogenesis. Clin Microbiol Rev 13(1):122–143
41. Bill CA, Vines CM (2020) Phospholipase C. Adv Exp Med Biol 1131:215–242
42. Bennett DE, McCreary CE, Coleman DC (1998) Genetic characterization of a phospholipase C gene from *Candida albicans*: presence of homologous sequences in Candida species other than *Candida albicans*. Microbiology 144:55–72
43. Kunze D, Melzer I, Bennett D, Sanglard D, MacCallum D, Norskau J, Coleman DC, Odds FC, Schafer W, Hube B (2005) Functional analysis of the phospholipase C gene CaPLC1 and two unusual phospholipase C genes, CaPLC2 and CaPLC3, of *Candida albicans*. Microbiology 151:3381–3394
44. Bandana K, Jashandeep K, Jagdeep K (2018) Phospholipases in bacterial virulence and pathogenesis. Adv Biotechnol Microbiol 10(5):106–113
45. Deepika D, Amarjeet S (2022) Plant phospholipase D: novel structure, regulatory mechanism, and multifaceted functions with biotechnological application. Crit Rev Biotechnol 42(1):106–124
46. Nakamura Y, Kanemaru K, Shoji M, Totoki K, Nakamura K, Nakaminami H, Nakase K, Noguchi N, Fukami K (2020) Phosphatidylinositol-specific phospholipase C enhances epidermal penetration by *Staphylococcus aureus*. Sci Rep 10:17845
47. Jenkins GM, Frohman MA (2005) Phospholipase D: a lipid centric review. Cell Mol Life Sci 62(19–20):2305–2316
48. Dolan JW, Bell AC, Hube B, Schaller M, Warner TF, Balish E (2004) *Candida albicans* PLD1 activity is required for full virulence. Med Mycol 42:439–447
49. Ponting CP, Kerr ID (1996) A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: identification of duplicated repeats and potential active site residues. Protein Sci Publ Protein Soc 5(5):914–922
50. Andrews S (2010) FastQC: A quality control tool for high throughput sequence data [Internet]. [cited 2018 Jun 1]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
51. Hannon GJ (2010) FASTX-Toolkit [Internet]. [cited 2018 Jun 1]. Available from: http://hannonlab.cshl.edu/fastx_toolkit
52. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5):821–829
53. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964
54. Lagesen K, Hallin PF, Roedland EA, Stærfeldt H, Rognes T, Ussery DW (2007) RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res 35(9):3100–3108
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
56. 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 Res 18:1979–1990
57. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212
58. Sewalt V, Shanahan D, Gregg L, La Marta J, Carrillo R (2016) The Generally Recognized as Safe (GRAS) process for industrial microbial enzymes. Ind Biotechnol 12(5):295–302
59. Singh DK, Tóth R, Gácser A (2020) Mechanisms of pathogenic *Candida* species to evade the host complement attack. Front Cell Infect Microbiol 10:94
60. Ohama T, Suzuku T, Mori M, Osawa S, Ueda T, Watanabe K, Nakase T (1993) Non-universal decoding of the leucine codon CUG in several Candida species. Nucleic Acids Res 21(17):4039–4045
61. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.