IDENTIFICATION OF CELL CYCLE-REGULATED, PUTATIVE HYPHAL GENES IN *CANDIDA ALBICANS*

RALUCA GORDÂN†

Division of Genetics, Department of Medicine,
Brigham & Women's Hospital and Harvard Medical School,
Boston, MA 02115, USA
Current address: Department of Biostatistics and Bioinformatics,
Institute for Genome Sciences and Policy,
Duke University
Email: raluca.gordan@duke.edu

SAUMYADIPTA PYNE†

Department of Medical Oncology,
Dana-Farber Cancer Institute, Harvard Medical School,
Boston, MA 02115, USA
Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
Email: spyne@broad.mit.org

MARTHA L. BULYK†

Division of Genetics, Department of Medicine, Department of Pathology,
Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
Harvard-MIT Division of Health Sciences & Technology (HST)
Harvard Medical School, Boston, MA 02115, USA
Email: mlbulyk@receptor.med.harvard.edu

†Correspondence can be addressed to R.G., S.P. or M.L.B.

Candida albicans, a major fungal pathogen in human, can grow in a variety of morphological forms ranging from budding yeast to pseudohyphae and hyphae, and its ability to transition to true hyphae is critical for virulence in various types of *C. albicans* infections. Here, we identify 17 putative hyphal genes whose expression peaks during the S/G2 transition of the cell cycle in *C. albicans*. These genes are *Candida*-specific (*i.e.*, they do not have orthologs in *S. cerevisiae*, a related fungal species that does not exhibit hyphal growth and is primarily non-pathogenic), and their promoters are enriched for the DNA binding site motifs of Tec1 and Rfg1, two transcription factors (TFs) known to play important roles in hyphal growth and virulence. For 5 of the 17 genes we found strong evidence in the literature that confirms our hypothesis that these genes are involved in hyphal growth and/or virulence, for 5 additional genes we found suggestive (albeit weak) evidence, while the other genes remain to be tested. It will be interesting to determine in future studies whether these 17 putative hyphal genes, whose expression peaks during the S/G2 transition, are part of a mechanism for this pathogenic fungus to 'turn on' hyphal growth late during the cell cycle, or if these genes are used to sustain hyphal growth and ensure that the cell does not transition back to yeast growth. In either case, the involvement of these genes in hyphal growth makes them putative targets for new antifungal drugs aimed at inhibiting hyphae formation in *C. albicans*.

1. Introduction

Candida albicans is a major human pathogen and the number one cause of fungal infection in human. Unlike other pathogens, it can be found in skin and the gastrointestinal tract as a harmless commensal organism, producing serious disease in people with weakened immune systems [1]. *C. albicans* is a truly polymorphic organism: it has the ability to undergo morphological changes between the yeast form (with rounded cells and daughter buds that physically separate from the mother cell), the pseudohyphal form (which consists of chains of cells with various degrees of elongation that still show constrictions between adjacent cells), and the true hyphal form (which consists of long tubes with parallel sides and no constrictions) [2]. Yeast cells disseminate more easily in the bloodstream, while hyphae are invasive and can penetrate host tissues during the early stages of infections [2]. Furthermore, switching of *C. albicans* to the hyphal form in the host has long been considered to be important for pathogenesis, since mutants defective in hyphal growth are known to be less virulent [3]. Thus, identification of genes involved in the yeast-to-hyphae transition is important for the development of new antifungal agents.

Here, we identify 17 putative hyphal genes that have a particular characteristic: they are periodically expressed during the *C. albicans* cell cycle, with their expression peaking during the S/G2 transition. We analyzed the gene expression data of Côte *et al.* [4], who examined the periodic expression of genes through the cell cycle in cultures of *C. albicans* synchronized by mating pheromone treatment. Côte *et al.* reported a set of 494 genes that are periodically expressed during the cell cycle, 100 of which do not have homologs in *S. cerevisiae*, a related fungal species that does not exhibit hyphal growth and is primarily non-pathogenic [5]. We henceforth refer to these 100 genes as "*Candida*-specific", and we anticipate that at least some of these genes may be necessary for hyphal growth or pathogenicity.

We investigated the transcriptional regulation of the *Candida*-specific genes, in an attempt to find possible clues about *C. albicans* hyphal growth and its connection to the cell cycle. We analyzed the promoter regions of periodically expressed genes that peak during different cell cycle transitions: G1/S, S/G2, G2/M, and M/G1. For each cell cycle transition we performed two motif enrichment analyses to identify: 1) DNA motifs enriched upstream of genes that peak at that particular transition, and 2) DNA motifs enriched upstream of *Candida*-specific genes that peak at that particular transition. Since high-resolution DNA binding site motifs, such as motifs derived from protein binding microarray (PBM) [6], SELEX-seq [7], or MITOMI [8] data, are not available for *C. albicans* TFs, we used as a proxy 139 high-resolution motifs of *S. cerevisiae* TFs [9-11] (see Section 2.2). To find significantly enriched motifs in the promoters of *C. albicans* cell cycle-regulated genes, we use a method that we developed recently to compute the enrichment of TF DNA binding motifs in genome-scale chromatin immunoprecipitation data on *in vivo* TF occupancy (ChIP-chip) [11]. Previously, we used this method successfully to distinguish between direct and indirect TF-DNA interactions in the yeast *S. cerevisiae*. Here, we apply a similar enrichment analysis to the sets of promoters of *C. albicans* cell cycle-regulated genes.

We find that the DNA motifs of Tec1 and Rfg1, two known regulators of hyphal growth and virulence [12], are significantly enriched upstream of *Candida*-specific genes that peak during the S/G2 transition, and are not enriched in general upstream of genes that peak at this stage. Since these 17 genes are regulated by hyphal TFs and do not have orthologs in *S. cerevisiae* (which is non-pathogenic and does not form true hyphae), we hypothesize that the 17 genes may be involved in hyphal growth and/or virulence. To test this hypothesis, we performed a literature search to see whether these genes are overexpressed during hyphal growth or whether there is any evidence of a role in virulence. Most of the 17 genes have unknown functions [13] and are not well represented in *C. albicans* gene expression data. Despite this fact, our literature search revealed strong evidence that 5 of the 17 genes are indeed involved in hyphal growth. For 5 additional genes we found suggestive (albeit weak) evidence, while the other 7 genes remain to be tested (see Table 1).

Since the expression of these 17 *Candida*-specific genes peaks at late stages of the cell cycle (more precisely, the S/G2 transition), and since it has been shown previously that *C. albicans* can be induced to start hyphal growth not only in G1 but also in later stages [14], our results suggest that these genes may be part of a mechanism for this pathogenic fungus to ‘turn on’ hyphal growth late during the cell cycle. Alternatively, the genes may be important for sustaining hyphal development throughout the cell cycle and preventing the cell from transitioning back to yeast growth.

2. Data and Methods

2.1. *Candida albicans gene expression data*

We used the cell cycle gene expression data of Côte *et al.* [3] to build sets of periodically expressed genes that are potentially regulated by a shared set of TFs. Côte *et al.* examined the periodic expression of genes through the cell cycle in cultures of *C. albicans* synchronized by mating pheromone treatment, and found 494 genes that are periodically expressed and peak at different cell cycle transitions: G1/S, S/G2, G2/M, or M/G1. The samples were collected at 0, 30, 60, 90, 120, 150, and 180 minutes after pheromone treatment, and each time point was manually assigned to a cell cycle transition [3]. For each cell cycle transition we built a “foreground” set of sequences that contains the 1-kb regions upstream of the transcription start sites of all *C. albicans* genes whose expression peaks at that particular transition, as shown in Figure 1 for the S/G2 transition. Next, for each foreground set (e.g., $F_{S/G2}$) we built a corresponding “background” set (e.g., $B_{S/G2}$) that contains the 1-kb promoter regions of the remaining *C. albicans* genes. Having constructed the foreground and background sequence sets, we next computed the enrichment of each query DNA motif in the foreground sequences as compared to the background sequences, for each cell cycle transition.

Of the 494 genes that are periodically expressed during the *C. albicans* cell cycle, 100 genes do not have orthologs in the non-pathogenic yeast *S. cerevisiae* [4]. We used the promoters of these *Candida*-specific genes to construct new sets of foreground (e.g., $FC_{S/G2}$) and background sequences for each cell cycle transition, to search for significantly enriched motifs upstream of the cell cycle-regulated *Candida*-specific genes. We were particularly interested to see whether there are TF motifs significantly enriched upstream of *Candida*-specific genes but not enriched upstream of all genes that peak at a particular time during the cell cycle.

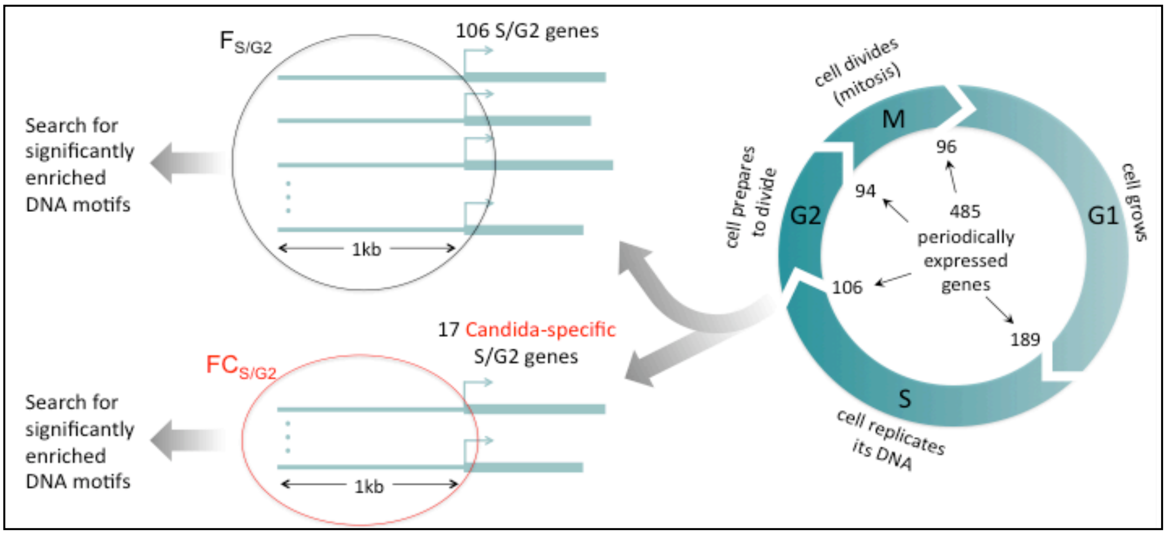

$F_{S/G2}$
106 S/G2 genes
Search for significantly enriched DNA motifs
1kb
$FC_{S/G2}$
17 Candida-specific S/G2 genes
Search for significantly enriched DNA motifs
1kb
cell divides (mitosis)
M
96
94
G2
cell prepares to divide
485 periodically expressed genes
G1
cell grows
106
189
S
cell replicates its DNA

Figure 1. Building the sets of foreground sequences. We use the promoter regions of genes that peak at each cell cycle transition to construct sets of foreground sequences that are then searched for significantly enriched motifs.

2.2. *TF binding site motif data*

Ideally we would perform the motif enrichment analysis using *C. albicans* high-resolution TF DNA binding site motifs, such as motifs derived from PBMs [6], SELEX-Seq [7], or MITOMI [8] data. However, since such motifs are not available for *C. albicans* TFs, in our analyses we used *S. cerevisiae* TF motifs as a proxy. We tested a previously assembled collection of 139 high-resolution motifs derived from universal PBM data [9-11]. For the TF binding site motifs significantly enriched in particular *C. albicans* data sets (including Tec1 and Rox1/Rfg1 – see Section 3.1), we compared the DNA binding domain (DBD) of the *S. cerevisiae* TF against the DBD of the *C. albicans* ortholog to ensure that the DBDs of the two proteins are similar, implying that the *C. albicans* TF likely has highly similar, if not essentially the same, DNA-binding specificity as its *S. cerevisiae* ortholog. To verify the similarity between an *S. cerevisiae* TF and its *C. albicans* ortholog, we performed a BLASTP search using the DBD of the *S. cerevisiae* TF (as defined in UniProt) or of the entire protein if the DBD was not well defined, and required that the *C. albicans* ortholog be recovered at an E-value < 1e-10. The DNA binding specificity of each TF is represented as a position weight matrix (PWM) [15].

2.3. *Method for computing enrichment of DNA motifs*

We recently developed a novel method for computing the enrichment of a TF DNA binding site motif in a set of foreground DNA sequences compared to a set of background sequences (derived from ChIP-chip data), and we used this method to successfully distinguish between direct and indirect TF-DNA interactions in *S. cerevisiae* [11]. Here, we apply a similar enrichment method to the sets of promoter regions of *C. albicans* cell cycle-regulated genes [4]. Previously [11], we used our method to compare the enrichment of several TFs in each set of foreground sequences; in this work, we compare the enrichment of each TF across several foreground sets (see below). Formally, the method can be described as follows.

Let ***F*** and ***B*** denote the sets of foreground and background sequences, respectively (e.g., $F_{S/G2}$ and $B_{S/G2}$, as described in Section 2.1). Let ***T*** denote a TF, and ***M*** denote the PWM describing the DNA binding specificity motif of ***T***: $\boldsymbol{M}(b, j)$ = the probability of finding base *b* at location *j* within the binding site (*b* = A, C, G, or T, and $1 \leq j \leq k$, where *k* is the motif width). Let ***Q*** denote the background nucleotide frequencies. Given a DNA site $\boldsymbol{S} = S_1S_2\ldots S_k$, we score it according to the PWM and background models, and use the ratio of the two scores to approximate the dissociation constant:

$$K_d(\boldsymbol{T},\boldsymbol{S}) = \prod_{j=1}^{k} \left(\boldsymbol{Q}(S_j)/\boldsymbol{M}(S_j, j)\right) \tag{1}$$

Next, we write the probability that ***T*** binds ***S*** as:

$$P(\boldsymbol{T}\text{ binds }\boldsymbol{S}) = \frac{[\boldsymbol{T}\cdot\boldsymbol{S}]}{[\boldsymbol{T}\cdot\boldsymbol{S}]+[\boldsymbol{S}]} = \frac{[\boldsymbol{T}]}{[\boldsymbol{T}]+K_d(\boldsymbol{T},\boldsymbol{S})} = 1\Big/\left(1+\frac{1}{[\boldsymbol{T}]}\times\prod_{j=1}^{k}\frac{\boldsymbol{Q}(S_j)}{\boldsymbol{M}(S_j, j)}\right) \tag{2}$$

where the concentration of free TF, [***T***], is set to the dissociation constant for the site with the optimal PWM score, as in the GOMER [16] model (this is equivalent to setting the TF occupancy of the optimal binding site to 50%). For a DNA sequence ***X*** longer than the motif width *k*, the probability that ***T*** binds ***X*** is:

$$P(\boldsymbol{T}\text{ binds }\boldsymbol{X}) = P(\boldsymbol{T}\text{ binds any }X_{i\ldots i+k-1}) = 1-\prod_{i=1}^{n-k+1}\left(1-1\Big/\left(1+\frac{1}{[\boldsymbol{T}]}\times\prod_{j=i}^{i+k-1}\frac{\boldsymbol{Q}(X_j)}{\boldsymbol{M}(X_j, j-i+1)}\right)\right) \tag{3}$$

Previously [11], we also considered the nucleosome occupancy when computing P(***T*** binds ***S***) (Eq. (3)). However, since here we are analyzing cell cycle expression data and nucleosome occupancy data for *C. albicans* cells in particular stages of the cell cycle are not yet available, here we do not incorporate DNA accessibility into the score.

After scoring all foreground and background sequences using Eq. (3), we use the computed scores to construct a receiver operating characteristic (ROC) curve and calculate the area under the ROC curve (AUC) as a measure of enrichment of the PWM in the foreground set as compared to the background set. An AUC value of 1 corresponds to perfect enrichment (*i.e.*, all foreground sequences have higher scores than the background sequences), while an AUC of 0.5 corresponds to lack of any enrichment (or of depletion), such as would be obtained using a random motif.

To compute the statistical significance of an AUC enrichment value (which corresponds to the results obtained for a particular PWM and particular sets of foreground and background sequences), we randomly permute the PWM 1,000 times as previously described [11], and for each random PWM we compute its AUC enrichment in the foreground versus the background sequences. We use the 1,000 AUC values of random motifs to compute an empirical p-value for the PWM of interest. We consider the PWM significantly enriched if it has an AUC >= 0.6 and an associated p-value <= 0.005 (i.e., for at most 2 out of 1,000 random motifs we obtain an AUC greater than or equal to the AUC of the real motif). The Perl scripts used to conduct the analyses are available online at http://thebrain.bwh.harvard.edu/psb2012.

3. Results

3.1. *DNA motifs significantly enriched upstream of periodically expressed genes*

When searching for DNA motifs significantly enriched upstream of genes that peak during specific cell cycle transitions in *C. albicans*, we recover known master regulators of the cell cycle: Mbp1, Swi4, Mcm1, Fkh2, etc. (Figure 2).

When we restrict the analysis to *Candida*-specific genes, we again recover the motifs of the master regulators at the various cell cycle transitions. In addition, we find that the DNA binding site motifs of *S. cerevisiae* TFs Tec1 and Rox1 are significantly enriched upstream of S/G2 genes unique to *C. albicans*, and are not significantly enriched in the set of all S/G2 genes. The orthologs of these two *S. cerevisiae* TFs in *C. albicans* are Tec1 and Rfg1, respectively, and they are both known regulators of hyphal growth and virulence in *C. albicans* [12]. Thus, we hypothesize that the 17 *Candida*-specific genes whose expression peaks at the S/G2 transition may also play a role in hyphal growth and/or virulence (see below).

Since we used *S. cerevisiae* motifs, and not *C. albicans* motifs, in our motif enrichment analysis, we checked to make sure that the proteins are well conserved over their DNA binding domains (TEA/ATTS domain for Tec1, and HMG box domain for Rox1/Rfg1), which implies that the *C. albicans* TFs likely have highly similar DNA binding specificities as their *S. cerevisiae* orthologs. Conservation over the DBDs was high, with BLASTP E-values as follows for the *C. albicans* orthologs recovered for the TFs shown in Figures 2 and 3: Fkh2-Fkh2 (1e-37), Mcm1-Mcm1 (1e-25), Hcm1-Hcm1 (1e-26), Swi4-Swi4 (1e-66), Mbp1-Mbp1 (1e-11), Azf1-Azf1 (1e-74), Ace2-Ace2 (1e-38), Tec1-Tec1 (1e-17), Rox1-Rfg1 (1e-21).

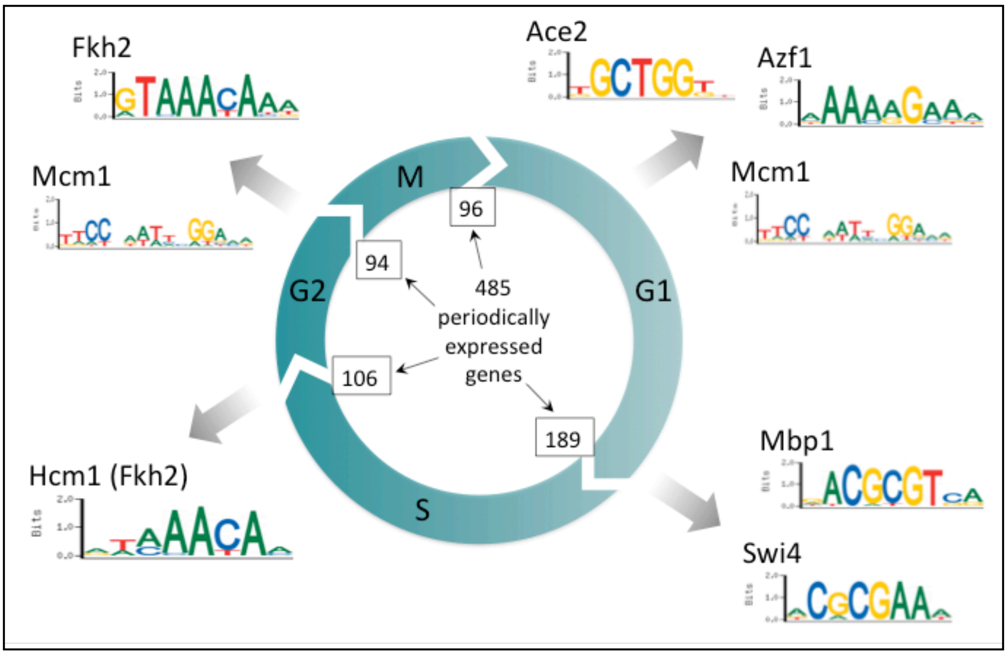
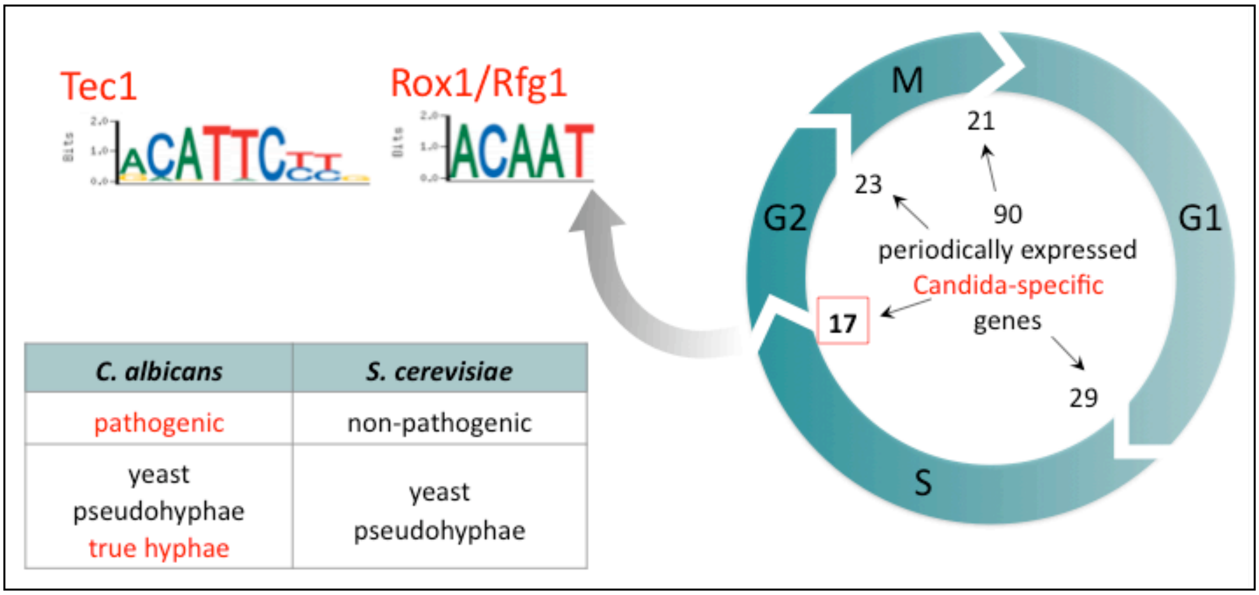


Figure 2. Enriched DNA motifs. The DNA binding site motifs of master regulators of the cell cycle are significantly enriched upstream of genes periodically expressed during the *C. albicans* cell cycle.



C. albicans	*S. cerevisiae*
pathogenic	non-pathogenic
yeast pseudohyphae true hyphae	yeast pseudohyphae

Figure 3. Motifs specifically enriched upstream of *Candida*-specific genes. We find that the DNA binding site motifs of *S. cerevisiae* TFs Tec1 and Rox1 (orthologs of *C. albicans* TFs Tec1 and Rfg1, respectively) are significantly enriched upstream of *Candida*-specific S/G2 genes.

3.2. *Tec1 and Rfg1 – regulators of hyphal growth and virulence*

Tec1 is a well-studied TEA/ATTS TF conserved across many fungal species, including *S. cerevisiae* and *C. albicans*. In *S. cerevisiae*, this factor is known to be required for haploid invasive and diploid pseudohyphal growth [17]. In *C. albicans*, Tec1 is involved in regulation of hypha-specific genes and it is required for wild-type biofilm formation [13]. *C. albicans* Tec1 acts downstream of Efg1, a protein that plays a critical role in hyphal morphogenesis [18]. Efg1 is required for expression of all hyphal-specific genes, and Tec1 overexpression has been shown to restore filamentous growth in an *efg1*/*efg1* mutant [12]. Furthermore, Tec1 is also involved in pathogenesis, as *TEC1* homozygous null mutants show decreased or absent hyphal growth and virulence [13].

The *C. albicans* TF Rfg1 is the ortholog of *S. cerevisiae* Rox1. However, unlike Rox1 in *S. cerevisiae*, *C. albicans* Rfg1 is not responsible for hypoxic repression [19]. Instead, it regulates filamentous growth and hyphal genes, acting in both Tup1p-dependent and -independent pathways [13]. Similarly to Tec1, Rfg1 controls both filamentous growth and virulence [19], and *RFG1* homozygous null mutants show decreased filamentous growth and are not virulent [13].

3.3. *Prior literature support for the hypothesis that the 17 S/G2 Candida-specific genes are putative hyphal genes*

Unlike *S. cerevisiae*, *C. albicans* is truly dimorphic: it has the ability to undergo morphological changes between the yeast, pseudohyphal, and the true hyphal forms. The fact that the 17 S/G2 *Candida*-specific genes do not have homologs in *S. cerevisiae* and seem to be regulated by Tec1 and Rfg1, two known regulators of hyphal genes, suggests that these 17 genes are also involved in hyphal growth and they may provide a connection between cell cycle and polymorphism in *C. albicans*. These genes may be involved in a mechanism that acts late during the *C. albicans* cell cycle and provides a last chance for the cell to commit to hyphal growth before entering a new cycle.

Testing putative hyphal genes in the laboratory is not trivial. Genetic studies in *C. albicans* are especially challenging because of its diploid genome, chromosomal instability, and incomplete sexual cycle [14, 20]. Nevertheless, several genes involved in hyphal growth have been identified thus far, so we searched the literature for evidence that supports our hypothesis that the 17 *Candida*-specific genes are part of the hyphal morphogenesis program. Except for orf19.3430 (*BUD21*) and orf19.6877 (*PNG2*), the remaining 15 genes are annotated in the *Candida* Genome Database (CGD) [13] as "Uncharacterized ORFs", and thus there is little information about them in the literature. Still, for two of the 17 genes (orf19.5848 and orf19.4905) Nantel *et al.* [21] have reported a significant increase in gene expression during yeast-to-hyphal transition. For one additional gene (*BDA1* or orf19.376), the homozygous deletion mutant shows substantial morphology defects [20], while the null mutant of orf19.4905 (a putative MFS* transporter) shows abnormal infectivity in a mouse infection model [20]. For two additional genes (orf19.3516 and orf19.3430), invasive growth is decreased in a heterozygous null mutant [5]. Relevant information about all 17 genes is summarized in Table 1. Most of these genes are well conserved across related fungi, but their orthologs are also uncharacterized.

We also note that although for some of these genes we have not found strong evidence of their direct involvement in hyphal growth, there is weak evidence supporting our hypothesis (see Table 1). For example, orf19.5549 encodes a protein that is very rich in Ser residues, a characteristic of cell surface proteins (which play an important role in the pathogenic process). Six of the 17 genes encode proteins that contain predicted transmembrane domains [13] (orf19.5549, orf19.5848, orf19.4905, orf19.3430, orf19.1350, orf19.6877), and 4 were predicted to contain signal peptides (orf19.5549, orf19.5848, orf19.7606, orf19.876), which are important for directing proteins to the cell wall and for secretion into the extracellular matrix. The gene *PNG2* was predicted to contain three PRICHEXTENSN domains characteristic of proline-rich extensins (plant cell wall proteins with functions in cell wall strengthening [23]). Another gene, orf19.876, codes for a putative GPI-anchored protein, and many hypha-specific genes encode GPI-anchored cell surface proteins [12].

4. Discussion

In this work, we identify 17 putative hyphal genes in the pathogenic yeast *C. albicans*, which have the specific characteristic that their expression is cell cycle-regulated and peaks at the S/G2 transition. These 17 genes are *Candida*-specific and seem to be regulated by Tec1 and Rfg1, two master regulators of hyphal growth, and thus we hypothesize that the 17 genes are also involved in hyphal growth. A literature search revealed evidence that supports our hypothesis for 10 of the 17 genes, while the other seven remain to be verified in the laboratory.

* Major facilitator superfamily (MFS) is one of the two major superfamilies of plasma membrane efflux proteins involved in antifungal drug resistance [22].

Table 1. *Candida*-specific genes that peak at the S/G2 transition.

ORF (gene name)	**Evidence strength**	**Gene information, protein information, and/or evidence of role in hyphal/invasive growth**
orf19.5549	weak	Proteins rich in Ser residues
orf19.5848	strong	Late-stage biofilm-induced gene [24] Upregulated during yeast-to-hyphal transition [21]
orf19.6238	weak	CGD molecular function: oxidoreductase activity PSI-BLAST: potential FAD-dependent oxidoreductase, similarity to Ser/Thr-protein kinase Chk2 in *P. pastoris*
orf19.376 *(BDA1)*	strong	Mutant shows substantial morphology defects; [20]
orf19.4905	strong	Mutant shows abnormal infectivity; putative MFS transporter [20] Upregulated during yeast-to-hyphal transition [21] CGD biological process: transmembrane transport
orf19.7606	no evidence	
orf19.836	no evidence	CGD description: Protein likely to be essential for growth, based on an insertional mutagenesis strategy
orf19.3516	strong	Invasive growth is decreased in a heterozygous null mutant [5] CGD molecular function: carbonate dehydratase activity
orf19.389	no evidence	CGD description: Hap43p-induced gene
orf19.3430 *(BUD21)*	strong	Invasive growth is decreased in a heterozygous null mutant [5] CGD description: plasma membrane-associated protein
orf19.1350	weak	Included in the “opaque-induced” transcriptional module [25] CGD molecular function: electron carrier activity; protein disulfide oxidoreductase activity PSI-BLAST: Thioredoxin_like superfamily
orf19.3245	no evidence	
orf19.876 *(PGA33)*	weak	CGD description: putative GPI-anchored protein CGD cellular component: cell surface
orf19.1050	no evidence	
orf19.6579	no evidence	
orf19.6877 *(PNG2)*	weak	CGD: transcription upregulated by treatment with caspofungin, ciclopirox olamine, ketoconazole or hypoxia; gene of core caspofungin response; CGD biological process: protein deglycosylation CGD cellular component: plasma membrane CGD conserved domains: 2 PNGaseA domains and 3 PRICHEXTENSN (proline-rich extensin signature) domains.
orf19.3871 *(DAD3)*	no evidence	Subunit of the Dam1 (DASH) complex, which acts in chromosome segregation by coupling kinetochores to spindle microtubules [13]

For all the analyses presented here we used *S. cerevisiae* TF DNA binding site motifs because high-resolution motifs are not yet available for *C. albicans*. However, some *C. albicans* TFs may have slightly different DNA binding specificities as compared to their *S. cerevisiae* orthologs. Furthermore, some TFs do not have orthologs in *S. cerevisiae*. Thus, once high-resolution DNA binding site motifs become available for *C. albicans* TFs, it will be interesting to repeat the analyses presented here to determine whether additional DNA motifs are significantly enriched upstream of *Candida*-specific genes. In addition, future analyses that combine *C. albicans* TF DNA binding data with cell cycle gene expression data and also gene expression data obtained during the yeast-to-hyphal transition could help us to understand how hyphal growth is initiated and maintained during the cell cycle, and what are the roles of these *Candida*-specific genes.

When computing the enrichment of a DNA motif in a set of foreground sequences compared to a set of background sequences, a commonly used approach is to search for good matches to the motif in both the foreground and background sets, and then use Fisher's exact test (*i.e.*, the hypergeometric p-value) to determine whether the motif is overrepresented in the foreground sequences [26, 27]. This approach is sensitive to the size of the two sets and the cutoff used to determine "matches" to the DNA motif, which is why we used an alternative, AUC-based approach (see Section 2.3.) that allowed us to directly compare the enrichment of different DNA motifs in different sets of sequences. Using the AUC-based approach we were able to identify the two motifs (Tec1 and Rox1/Rfg1) that are enriched upstream of *Candida*-specific S/G2 genes and not enriched in general upstream of S/G2 genes. For comparison, we also performed an enrichment analysis using Fisher's exact test (see Supplementary Figure 1, available online at http://thebrain.bwh.harvard.edu/psb2012). The results were inconclusive: we did not find a significant enrichment for either Tec1 or Rox1/Rfg1, and it was unclear whether the motifs were more enriched in the *Candida*-specific promoters or all S/G2 promoters. Furthermore, as expected, the computed p-values for the Tec1 and Rox1/Rfg1 motifs varied widely depending on the motif cutoff. Our AUC-based approach alleviates the need to choose a motif cutoff by taking into account all possible binding sites in a given sequence, weighted according to how well they match the motif of interest. Our approach is not limited to *Candida*-specific genes or to cell cycle data, but rather can be used for any organism to identify regulators of organism-specific genes that exhibit a particular characteristic.

Our finding that the DNA binding site motifs of hyphal TFs Tec1 and Rox1 are enriched upstream of the 17 *Candida*-specific S/G2 genes is intriguing. We had expected to find the DNA binding site motifs of master hyphal growth regulators enriched upstream of genes whose expression peaks early in the cell cycle, most likely in the G1 phase, since it has been proposed that there is a point of phenotypic commitment to hyphal growth in G1 [14, 28]. However, the 17 *Candida*-specific, putative hyphal genes we identified peak at the S/G2 transition, which suggests that they may be involved in the initiation of hyphae formation (or hyphal evagination, given that some genes may encode cell surface proteins) late during the cell cycle. We note that whether hyhae can be induced from all cell cycle stages in *C. albicans* is a matter of some controversy [29]. Still, some studies have shown that *C. albicans* can be induced to start hyphal growth not only in G1, but also in later stages (definitely S and G2, and possible M also) [14], so the 17 *Candida*-specific S/G2 genes could potentially be involved in hyphal initiation.

The ability of *C. albicans* to switch between yeast and hyphal growth has been linked to its virulence [3], so involvement of the 17 *Candida*-specific genes in hyphal growth suggests that they may also be important for virulence/pathogenesis[†]. We note that the connection between morphological switching and virulence is controversial in the *Candida* literature, as it has recently been demonstrated that the yeast-to-hyphae transition is not necessarily required for infectivity in a mouse model of disseminated candidiasis and infection of the kidney [20]. However, it is still generally accepted that hyphal growth is critical for virulence in many types of *C. albicans* infections [31].

The hypothesis that the 17 *Candida*-specific genes may be important for virulence is supported by the fact that the TFs Tec1 and Rfg1, whose DNA binding site motifs are enriched upstream of the 17 genes, play important roles in both hyphal growth and virulence [13]. Reduced virulence has been observed in both *TEC1* and *RFG1* mutants, and also in *EFG1* mutants [12] (the TF Efg1 acts upstream of Tec1 in several signal transduction pathways that regulate the yeast-to-hyphae transition [18]). Efg1 is known to be important for *C. albicans* pathogenesis in several *in vivo* models of infection [12], and one of pathways known to regulate yeast-to-hyphae transition via Efg1 (a cAMP-dependent protein kinase pathway) plays a major role in both hyphal development and pathogenesis [12]. Furthermore, at least one of the 17 *Candida*-specific genes (orf19.4905) has been shown to play a role in pathogenesis, as a homozygous null mutant showed abnormal infectivity in a mouse model [20]. Thus, a number of the other 16 putative hyphal genes we identified may also be important for *C. albicans* pathogenesis. We note that acquiring new genes is not the only way to develop virulence/pathogenicity. Some protein-coding genes present in both *S. cerevisiae* and *C. albicans* may have acquired coding mutations that allow them to be more virulent. Some genes conserved at the sequence level may have diverged in their gene expression due to changes in non-coding *cis*-regulatory elements. Such genes should also be included in future analyses aimed at deciphering regulatory networks of *C. albicans* virulence.

Identifying genes responsible for hyphal growth and virulence in *C. albicans* is undoubtedly important for understanding how this fungus invades human cells and how it switches from a harmless organism to a pathogen. The work presented here provides a useful resource for future studies of *Candida*-specific hyphal and virulence genes. Such studies may aid in understanding the biology of this important human pathogen.

Acknowledgments

We thank Gerry Fink and James Konopka for helpful discussion. This work was supported by NIH/NHGRI grant # R01 HG003985 (M.L.B.). R.G. was funded in part by an American Heart Association postdoctoral fellowship #10POST3650060.

[†] The terms virulence and pathogenesis are closely related. The former is viewed from the point of view of the fungus, while the latter from its effect on the host [30].

References

1. T. E. Zaoutis, J. Argon, J. Chu, J.A. Berlin, T. J. Walsh and C. Feudtner, *Clin. Infect. Dis.* **41**, 1232-9 (2005)
2. P. Sudbery, N. Gow and J. Berman, *Trends Microbiol.* **12**, 317-24 (2004)
3. H. J. Lo, J. R. Kohler, B. DiDomenico, D. Loebenberg, A. Cacciapuoti and G. R. Fink, *Cell* **90**, 939-49 (1997)
4. P. Côte, H. Hogues and M. Whiteway, *Mol. Biol. Cell.* **20**, 3363-73 (2009)
5. J. Oh, E. Fung, U. Schlecht, R. W. Davies, G. Giaever, R. P. St Onge, A. Deutschbauer and C. Nislow, *PLoS Pathog.* **6**, e1001140 (2010)
6. M. F. Berger, A. A. Philippakis, A. M. Qureshi, *et al.*, *Nat. Biotechnol.* **24**, 1429-35 (2006)
7. A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, M. Bonke, K. Palin, *et al.*, *Genome Res*. **20**, 861-73 (2010)
8. S. J. Maerkl and S. R. Quake, *Science* **315**, 233-7 (2007)
9. C. Zhu, K. Byers, R. McCord, Z. Shi, M. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, *et al.*, *Genome Res.* **19**, 556-66 (2009).
10. G. Badis, E.T. Chan, H. van Bakel, L. Peña-Castillo, *et al.*, *Mol. Cell* **32**, 878-87 (2008)
11. R. Gordân, A. J. Hartemink and M. L. Bulyk, *Genome Res.* **19**, 2090-100 (2009).
12. H. Liu, *Int. J. Med. Microbiol.* **292**, 299-311 (2002)
13. M. S. Skrzypek, M. B. Arnaud, M. C. Costanzo, D. O. Inglis, P. Shah, G. Binkley, S. R. Miyasato and G. Sherlock, *Nucleic Acids Res.* **38**, D428-32 (2010)
14. I. Hazan, M. Sepulveda-Becerra and H. Liu, *Mol. Biol. Cell* **13**, 134-45 (2002)
15. G. D. Stormo, *Bioinformatics* **16**, 16-23 (2000)
16. J. A. Granek and N. D. Clarke, *Genome Biol.* **6**, R87 (2005).
17. T. Köhler, S. Wesche, N. Taheri, G. H. Braus, H. U. Mösch, *Eukaryot. Cell* **1**, 673-86 (2002)
18. J. Berman and P. E. Sudbery, *Nat. Rev. Genet.* **3**, 918-30 (2002)
19. D. Kadosh and A. D. Johnson, *Mol. Cell. Biol.* **21**, 2496-505 (2001)
20. S. M. Noble, S. French, L. A. Kohn, V. Chen and A. D. Johnson, *Nat. Genet.* **42**, 590-8 (2009)
21. A. Nantel, D. Dignard, C. Bachewich, D. Harcus, A. Marcil, A.-P. Bouin, C. W. Sensen, H. Hogues, M. van het Hoog, P. Gordon, T. Rigby, *et al.*, *Mol. Biol. Cell* **13**, 3452-65 (2002)
22. R. Cannon, E. Lamping, A. Holmes, K. Niimi, *et al.*, *Clin. Microbiol. Rev.* **22**, 291-321 (2009)
23. V. Stiefel, L. Perez-Grau, F. Albericio, E. Giralt, L. Ruiz-Avila, M. D. Ludevid and P. Puigdomenech, *Plant Mol. Biol.* **11**, 483-93 (1988)
24. J. Bonhomme, M. Chauvel, S. Goyard, P. Roux, T. Rossignol and C. d'Enfert, *Mol. Microbiol.* **80**, 995-1013 (2011)
25. J. Ihmels, S. Bergmann, J. Berman and N. Barkai, *PLoS Genet.* **1**, e39 (2005)
26. S. Tavazoie, J. Hughes, M. J. Campbell, R. Cho and G. Church, *Nat. Genet*. **22**, 281-5 (1999)
27. J. D. Hughes, P. W. Estep, S. Tavazoie and G. M. Church, *J. Mol. Biol.* **296**, 1205-14 (2000)
28. L. H. Mitchell and D. R. Soll, *Exp. Cell Res.* **120**, 167-79 (1979)
29. J. Berman, *Curr. Opin. Microbiol.* **9**, 595-601 (2006)
30. R.A. Weiss, *Trends Microbiol.* **10**, 314-17 (2002)
31. J. Shareck, A. Nantel and P. Belhumeur, *Eukaryot. Cell* **10**, 565–77 (2011)