Improved Gene Ontology Annotation for Biofilm Formation, Filamentous Growth, and Phenotypic Switching in *Candida albicans*

Diane O. Inglis, Marek S. Skrzypek, Martha B. Arnaud, Jonathan Binkley, Prachi Shah, Farrell Wymore, Gavin Sherlock

Department of Genetics, Stanford University Medical School, Stanford, California, USA

The opportunistic fungal pathogen *Candida albicans* is a significant medical threat, especially for immunocompromised patients. Experimental research has focused on specific areas of *C. albicans* biology, with the goal of understanding the multiple factors that contribute to its pathogenic potential. Some of these factors include cell adhesion, invasive or filamentous growth, and the formation of drug-resistant biofilms. The Gene Ontology (GO) (www.geneontology.org) is a standardized vocabulary that the *Candida* Genome Database (CGD) (www.candidagenome.org) and other groups use to describe the functions of gene products. To improve the breadth and accuracy of pathogenicity-related gene product descriptions and to facilitate the description of as yet uncharacterized but potentially pathogenicity-related genes in *Candida* species, CGD undertook a three-part project: first, the addition of terms to the biological process branch of the GO to improve the description of fungus-related processes; second, manual recuration of gene product annotations in CGD to use the improved GO vocabulary; and third, computational ortholog-based transfer of GO annotations from experimentally characterized gene products, using these new terms, to uncharacterized orthologs in other *Candida* species. Through genome annotation and analysis, we identified candidate pathogenicity genes in seven non-*C. albicans* Candida species and in one additional *C. albicans* strain, WO-1. We also defined a set of *C. albicans* genes at the intersection of biofilm formation, filamentous growth, pathogenesis, and phenotypic switching of this opportunistic fungal pathogen, which provides a compelling list of candidates for further experimentation.

The *Candida* Genome Database (CGD) (www.candidagenome.org) is the central repository for the genome sequence and annotation of *Candida albicans* SC5314 (1–3) and is a source of sequence and annotation data for other *Candida* species. At the CGD, Ph.D.-level curators read and analyze the gene-specific literature for *C. albicans*, *C. glabrata*, and *C. parapsilosis* and record detailed information about genes and gene products, including gene names and synonyms, succinct gene descriptions, Gene Ontology (GO) annotations, and mutant phenotypes. This consolidates published gene information into a single, publicly available resource.

The GO is a widely used hierarchical vocabulary for assigning functional information about gene products (4). The GO is comprised of a structured set of terms that describe molecular function (MF), or activity in the cell; biological process (BP), or the larger context in which the gene product acts; and cellular component (CC), which provides the location within (or outside) a cell where a gene product exists. As a standardized vocabulary to describe gene products, the GO was originally developed for the biology of nonpathogenic model organisms. Consequently, pathogenesis-related processes have historically been under-represented in the GO.

Using ortholog relationships in conjunction with experimentally based GO annotation. Manually curated annotations of characterized genes in one species can be used computationally to annotate uncharacterized genes in another species, based on ontology relationships. For example, experimentally determined GO annotations from the well-studied model yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* are computationally transferred to the orthologs in CGD. Such annotations are assigned an evidence code of IEA (inferred by electronic annotation) so they can readily be distinguished from others (3). This approach to genome annotation provides a powerful means by which an organism’s genes that have not been characterized experimentally can be given informative annotations. CGD also uses this process to transfer GO annotations among the *Candida* species. This systematic electronic transfer of GO terms has been invaluable for annotating gene products of *C. glabrata* and *C. parapsilosis*, because compared to *C. albicans*, these species have a much smaller collection of experimental literature available for curation. Thus, the manual curation of the *C. albicans* genes from the literature has a significant impact on the quality of the computational GO terms transferred to these less-well-annotated species.

In this work, we describe the addition of new GO terms to the biological process branch of the ontology and the use of these new terms to improve our descriptions of *Candida* gene products in CGD. Through comprehensive literature re-review and gene annotation, we identified the major genes and pathways that contribute to the pathogenic traits of cell adhesion, biofilm formation, filamentous growth, and phenotypic switching and have experimentally demonstrated roles in pathogenesis of *C. albicans*. Finally, we used these new annotations to make computational, orthology-based predictions of pathogenesis-related functions for uncharacterized genes across multiple *Candida* species, laying the foundation for future work analyzing the contributions to virulence of these orthologs in non-*C. albicans* Candida species.
Ortholog predictions. The ortholog mappings between *Schizosaccharomyces pombe* and *Candida* species (*C. albicans*, *C. glabrata*, and *C. parapsilosis*) were made by pairwise comparisons using InParanoid software (5). The following stringent cutoffs were used: BLOSUM80 and an InParanoid score of 100% (parameters: --Fi "m" S"i" --M BLOSUM80). The data from this comparison are available for download at [http://www.candidagenome.org/download/homology/orthologs/](http://www.candidagenome.org/download/homology/orthologs/). To identify orthologous genes in *C. albicans* SC5314, *C. albicans* WO-1, *C. dubliniensis* CD36, *C. glabrata* CBS138, *C. parapsilosis* CDC317, *C. guilliermondii* ATCC 6260, *C. lusitaniae* ATCC 42720, *C. orthopsilosis* Co 90-125, *C. tropicalis* MYA-3404, and *S. cerevisiae*, we used the orthology predictions made at the *Candida* Gene Order Browser (CGOB) ([http://cgob3.ucd.ie/](http://cgob3.ucd.ie/)) (6), which incorporates genomic positional information and manual evaluation to assign orthologs.

**Orthology-based prediction of GO annotations.** GO annotations were predicted based on orthology when a given gene had a characterized ortholog in CGD, the *Saccharomyces* Genome Database (SGD), or PomBase. Candidate annotations for transfer were selected from those based on experimental evidence, that is, those with the evidence code IDA (inferred from direct assay), IPI (inferred from physical interaction), IGI (inferred from genetic interaction), or IMP (inferred from mutant phenotype). Annotations with the qualifier “NOT” were not transferred. All transferred annotations were given the evidence code IEA (inferred by electronic annotation) to identify the annotation as being computationally derived. The GO annotation files containing the complete list of orthologs are available from the CGD ([http://www.candidagenome.org/download/homology/orthologs/](http://www.candidagenome.org/download/homology/orthologs/)).

**Pathogenicity gene prediction and annotation.** The highlighting of the manually annotated pathogenicity genes of *C. albicans* (see Table S3 in the supplemental material) was based on the strain background(s) described in the publication in which the pathogenicity of a given gene was characterized. Strains CAF2-1, CAI-4, CAI-12, RM100, RM1000, BWP17, SN152, and SN95 were all derived from the SC5314 strain background (see [http://www.candidagenome.org/Strains.shtml](http://www.candidagenome.org/Strains.shtml) for *Candida* strain lineage information).

**RESULTS.** *C. albicans* is the best studied, and thus the most comprehensively annotated, of the *Candida* species, having the largest corpus of literature. Much of this literature has focused on biological processes thought to be related to pathogenic traits, such as adherence to host cells and filamentous growth, and more recently on mating, phenotypic switching, and biofilm formation, with a special emphasis on regulators of these processes. In the process of capturing new functions for *C. albicans* proteins, new GO terms had to be developed in order to make the most accurate and useful GO annotations possible. To improve the quality and specificity of GO annotations for *C. albicans* proteins involved in these processes, we identified *C. albicans* genes for which no GO annotations existed or for which annotations needed to be updated, requesting new GO terms from the GO Consortium (GOC) where needed. We reexamined approximately 450 papers and reviewed the annotations of over 850 *C. albicans* genes (Table 1).

**Gene Ontology biological process vocabulary in areas related to Candida pathogenesis.** Host-pathogen interaction assays with *C. albicans* typically involve the assessment of virulence in the host, filamentous growth within a host or on a host cell surface, or the ability to adhere to or invade host tissues. Genes that have been shown to play a role in virulence have previously been annotated with the term “pathogenesis” (GO:0009405), and no further modifications to these annotations were deemed necessary. This category makes up one of the largest classes of manually annotated genes of *C. albicans*, with 214 genes annotated with the “pathogenesis” term (Fig. 1).

For the genes that function in hyphal morphogenesis *in vivo*, we reviewed and recurred GO terms for all of the genes that had filamentous growth-defective phenotypes. Genes that contribute to the ability of *C. albicans* to adhere to or invade host tissues are now annotated “adhesion to host” (GO:0044406) (Table 2). The term “symbiosis” is used in the GO to cover the entire spectrum of interactions between organisms, ranging from a mutually beneficial relationship to a mutually benign commensal relationship, with parasitism, a relationship in which one organism harms the other, at the other extreme. Consequently, genes involved in growth within the host or filamentous growth within a host or host tissues are now annotated with the terms “growth of symbiont in host” (GO:0044117) and “development of symbiont in host” (GO:0044114) (Table 2; see Table S1 in the supplemental material), which are the GO terms used to describe both pathogenic and commensal relationships between species.

**Biofilm formation.** Biofilms are surface-associated microbial communities that provide a protective environment for growth. Candidiasis in patients is often associated with biofilms on indwelling medical devices (e.g., dental implants, catheters, and other surfaces) that act as substrates for growth (7, 8). Biofilms formed by wild-type *C. albicans* strains are highly resistant to the antifungal agents fluconazole, nystatin, amphotericin B, and chlorhexidine (9–11), making the understanding of biofilm formation in *C. albicans* and other biofilm-producing *Candida* species critical for advances to improve clinical patient outcomes.

Recent studies from several labs have discovered new genes involved in biofilm formation and adhesion during biofilm formation, including 11 transcriptional regulators of *C. albicans* biofilm formation (12, 13) and 30 regulators of adhesion (8). The identification of these new regulatory factors created the need for new GO terms to annotate these regulators of adhesion and biofilm formation. At the inception of this project, terms for biofilm formation were present in the GO, but no terms for the regulation of biofilm formation or the regulation of adhesion during biofilm formation were available.

Adhesion of yeast cells to a substrate is the first step in biofilm formation. In order to capture the adhesion function of genes during the context of early biofilm development, we replaced annotations with the less specific GO term “cell adhesion” (GO:0007155) with more specific GO terms, such as “cell adhesion involved in single-species biofilm formation” (GO:0043709), and added the regulatory term “positive regulation of cell adhesion involved in single-species biofilm formation” (GO:1900189) (Table 2). Subsequent to the initial adherence stage, mature biofilm formation is driven by selecting the most specific GO term for each gene.

**Functional category** | **No. of genes with GO term assignments** | **No. of papers reexamined**
--- | --- | ---
Cell adhesiona | 42 | 73
Biofilm formation | 122 | 344
Filamentous growth | 521 | 26
Pathogenesis | 214 | 26
Phenotypic switching | 44 | 26

a The number of genes includes child terms for each biological process.
development involves the production of hyphae, the elaboration of an extracellular matrix (14), and the release of new daughter cells (8). To precisely capture the function of genes involved in mature biofilm development, the terms “single-species biofilm formation” (GO:0044010) and “single-species biofilm formation on an inanimate substrate” (GO:0044011) were assigned instead of the less specific parent term, “biofilm formation” (GO:0042710) (Table 2; see Table S1 in the supplemental material).

Whereas intravenous catheter models for biofilm formation within a model host are presumed to represent infection by a single species, oral models of biofilm formation, for example, on acrylic denture material in rodents, involve interactions with other microorganisms of the oral flora (15, 16). New annotations based on such models are made using the term “cell adhesion involved in multispecies biofilm formation” (GO:0043710) (Table 2).

**TABLE 2 Numbers of *C. albicans* manual annotations for specific biological processes in the GO**

| GO ID   | GO term                                         | No. of genes |
|---------|------------------------------------------------|--------------|
| 0044406 | Adhesion to host                                | 41           |
| 0044117 | Growth of symbiont in host                     | 19           |
| 0044118 | Development of symbiont in host cell           | 0            |
| 0044114 | Development of symbiont in host                | 18           |
| 0031589 | Cell-substrate adhesion                         | 8            |
| 0043709 | Cell adhesion involved in single-species biofilm formation | 10 |
| 1900189 | Positive regulation of cell adhesion involved in single-species biofilm formation | 31 |
| 0044011 | Single-species biofilm formation on an inanimate substrate | 107 |
| 1900231 | Regulation of single-species biofilm formation on an inanimate substrate | 1 |
| 1900233 | Positive regulation of single-species biofilm formation on an inanimate substrate | 8 |
| 0043710 | Multispecies biofilm formation on an inanimate substrate | 6 |
| 0044407 | Single-species biofilm formation in or on host organism | 8 |
| 0036166 | Phenotypic switching                            | 31           |
| 1900239 | Regulation of phenotypic switching              | 10           |
| 1900241 | Positive regulation of phenotypic switching     | 8            |
| 1900240 | Negative regulation of phenotypic switching     | 2            |
In total, the terms for “regulation of cell adhesion involved in biofilm formation” (GO:1900187) and “cell adhesion involved in single-species biofilm formation” (GO:0043709) are now used to annotate the 30 adhesion regulators that have been identified to date (8). Biofilm regulators also receive annotations with the basic process term “single-species biofilm formation on an inanimate substrate” (GO:0044011). Basic process terms and regulatory process terms map to parental terms in different branches of the GO hierarchy (Fig. 2). Applying both terms for the regulatory proteins ensures that the genes map appropriately to both “biofilm formation” (GO:0042710) and “regulation of biological process” (GO:0050789), providing more accurate information about the functions of these regulatory proteins than either individual annotation alone.

**Filamentous growth.** Under standard laboratory growth conditions (yeast extract-peptone-dextrose medium [YPD] at 30°C), *C. albicans* grows as a population of unicellular yeast-form cells. In response to specific environmental cues, such as a 37°C temperature, nutrient starvation, neutral pH, or growth in serum, *C. albicans* undergoes a rapid morphological transition, during which growth in the round yeast form is replaced by growth in an elongated, invasive filamentous mode. This transition starts with the formation of a structure referred to as a germ tube and ultimately results in the formation of hyphae and/or pseudohyphae (17, 18). This morphological transition is also important for proper biofilm formation and has extensively been linked to pathogenesis in the experimental literature (19).

“Filamentous growth” is defined in the GO as “the process in which a multicellular organism, a unicellular organism or a group of unicellular organisms grow in a threadlike, filamentous shape.” The definition of “hyphal growth” is “growth of fungi as threadlike, tubular structures that may contain multiple nuclei and may or may not be divided internally by septa, or cross-walls.” CGD now uses new specific filamentous growth process terms to annotate genes involved in the actual process of making filaments (i.e., for mutants that do not form filaments) and reserves the GO term “hyphal growth,” a child term of “filamentous growth” (see Table S2 in the supplemental material), for genes that affect hyphal morphology (i.e., for mutants that display twisted, bent, wide, or excessively branched filaments).

Another annotation issue that we addressed is that some genes are observed to influence filamentous growth under some conditions but not others. Thus, new, more specific filamentous growth terms were needed for accurate annotation of this type of genes. The new filamentous growth annotations are now broken down by categories of stimuli so that the GO annotations now reflect the precise type of process described, as defined by the type of experimental assay, which presumably represents particular classes or distinct aspects of *Candida* growth in vivo (Table 3; see Fig. S1 and Table S1 in the supplemental material). For example, a gene that is involved in the response to only one type of stimulus, such as serum, is now annotated “filamentous growth of a population of unicellular organisms in response to biotic stimulus” (GO:0036180) and not in response to other types of stimuli. Additionally, since genes that are involved in responses to serum, starvation, increased temperature, neutral pH, and chemical stimuli are considered to function in response to these stimuli, filamentous growth-related genes are also annotated by cellular process: “cellular response to biotic substance” (GO:0071216), “cellular response to starvation” (GO:0009267), “cellular response to heat” (GO:0034605), “cellular response to neutral pH” (GO:0036244), and “cellular response to chemical stimulus” (GO:0070887), re-
In addition, the GO terms “cellular response to N-acetyl-D-glucosamine” (GO:0097316) and “cellular response to farnesol” (GO:0097308) are used, as appropriate, to describe genes that participate in response to these chemical stimuli.

To describe the filamentous growth of Candida growing either on a solid surface as colonies or in liquid culture, terms such as “filamentous growth of a population of unicellular organisms in response to heat” (GO:0036168) and “filamentous growth of a population of unicellular organisms in response to neutral pH” (GO:0036178) are used, as appropriate for the biological conditions examined. These new specific terms are child terms of “filamentous growth of a population of unicellular organisms” (GO:0034605), which is defined as “unipolar budding to a threadlike, filamentous shape.” Growth mode assays that examine the yeast-to-filament transition at the cellular level are typically performed in liquid medium, and the morphological transition is monitored on a population of individual cells. To specifically address filamentous growth functions that have been observed at the level of the single cell, we created the new GO terms “cell growth mode switching, budding to filamentous” (GO:0036187) and “cell growth mode switching, filamentous to budding” (GO:0097321) to annotate genes involved in the filament-to-yeast transition.

**Phenotypic switching.** Phenotypic switching was originally observed in *C. albicans* clinical isolates that underwent dramatic and reversible changes in colony morphology (20). One particular type of phenotypic switch, the W-O switch, was discovered, where “white”-form cells (considered the default growth state) convert to “opaque”-form cells that are characterized by changes in cellular morphology (20, 21), gene expression (22), environmental responses (23), host-immune cell interactions (24, 25), and the ability to mate (26).

Mutations in 31 *C. albicans* genes have been described in the literature as resulting in defects in W-O switching, and 14 of these genes have characterized roles in pathogenesis (see Table S2 in the supplemental material). Since no GO terms previously existed to describe the process of phenotypic switching, we added necessary terms to the GO vocabulary and then annotated these genes with new terms (Table 2; see Table S2 and Fig. S1). Research on the W-O switch has predominantly involved elucidation of the regulatory circuits that control the switch. The WOR1 gene was discovered to be a “master regulator” of W-O switching; however, a role for WOR1 in pathogenesis has not been reported. Other genes involved in regulating the W-O switch include WOR2, CZF1, and MYO5, among others (27, 28). Thus, both positive and negative phenotypic switching regulatory terms were needed to describe the regulators of phenotypic switching. All genes related to phenotypic switching can be found at CGD by searching with the GO term “phenotypic switching” (GO:0036166).

**Leveraging manual GO annotations to generate orthology-based GO annotations.** The use of GO to describe gene products provides a powerful handle for analysis of gene functions and comparison across the fungal clade and throughout the animal kingdom. To augment the manual curation, we leveraged orthology relationships to infer GO annotations for genes with an experimentally characterized ortholog in CGD, SGD, or PomBase. Predictions for *C. albicans* were made based on *S. cerevisiae*, *S. pombe*, *C. glabrata*, and *C. parapsilosis* orthologs, whereas predictions for *C. glabrata* were based on orthologs from *S. cerevisiae*, *S. pombe*, *C. albicans*, and *C. parapsilosis* and predictions for *C. parapsilosis* were based on orthologs from *S. cerevisiae, S. pombe, C. albicans*, and *C. glabrata*. These new annotations were computationally predicted and assigned based on manual annotations. The total numbers of computational GO annotations that were predicted

---

**TABLE 3 New GO terms that specify the filamentous growth-inducing stimulus assayed**

| GO ID    | GO term                                                                 | Filamentous growth-inducing media                                                                 | No. of *C. albicans* genes |
|----------|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|---------------------------|
| 0036168  | Filamentous growth of a population of unicellular organisms in response to heat | 37°C YPD, other solid and liquid media at 37°C                                                     | 20                        |
| 0036165  | Invasive growth in response to heat                                      | Solid and liquid media at 37°C                                                                    | 1                         |
| 0034605  | Cellular response to heat                                                | Solid and liquid media at 37°C                                                                    | 38                        |
| 0036170  | Filamentous growth of a population of unicellular organisms in response to starvation | Spider medium, SLAD                                                                               | 225                       |
| 0009267  | Cellular response to starvation                                          | Spider medium, SLAD                                                                               | 215                       |
| 0036171  | Filamentous growth of a population of unicellular organisms in response to chemical stimulus | GlcNAc, proline, farnesol                                                                          | 41                        |
| 0070887  | Cellular response to chemical stimulus                                    | Other chemical stimuli                                                                            | 27                        |
| 0097308  | Cellular response to farnesol                                            | Farnesol                                                                                            | 5                          |
| 0097316  | Cellular response to N-acetyl-D-glucosamine                               | GlcNAc                                                                                             | 7                          |
| 0036177  | Filamentous growth of a population of unicellular organisms in response to pH | M199, pH 8.0; other alkaline media                                                                | 21                        |
| 0071467  | Cellular response to pH                                                  | M199, pH 8.0; other alkaline media                                                                | 21                        |
| 0071469  | Cellular response to alkalinity                                          | M199, pH 8.0; other alkaline media                                                                | 16                        |
| 0036178  | Filamentous growth of a population of unicellular organisms in response to neutral pH | Lee’s medium, pH 6.8; M199, pH 6.8–7.0                                                            | 63                        |
| 0036244  | Cellular response to neutral pH                                          | Lee’s medium, pH 6.8; M199, pH 6.8–7.0                                                            | 52                        |
| 0036180  | Filamentous growth of a population of unicellular organisms in response to biotic stimulus | Solid and liquid media with serum                                                                  | 258                       |
| 0071216  | Cellular response to biotic stimulus                                     | Solid and liquid media with serum                                                                  | 251                       |
| 0044182  | Filamentous growth of a population of unicellular organisms              | Embedded conditions, unspecified or ambiguous stimulus                                              | 151                       |
and transferred from the characterized species to the noncharacterized species are summarized in Table 4.

**Prediction of non-*C. albicans* Candida pathogenicity genes by orthology.** Nosocomial infections and drug resistance due to non-*C. albicans* Candida species are on the rise (29). In addition to *C. albicans* SC5314, CGD currently curates the literature and manually annotates the genes of *C. glabrata* CBS138 and *C. parapsilosis* CDC327. To identify new candidate genes that may play a role in pathogenesis, we determined the experimentally characterized set of 224 *C. albicans*, *C. glabrata*, and *C. parapsilosis* genes that were manually annotated as having a role in pathogenesis with phenotype-based evidence (IMP) (highlighted in Table S3 in the supplemental material). We then used the orthology relationships for *Candida* species at CGOB (http://cgob3.ucd.ie/) (6) to identify genes orthologous to the pathogenicity genes in seven non-*C. albicans* Candida species and two *C. albicans* strains (see Table S3). In total, 213 ortholog groups including 1,516 predicted pathogenesis genes were identified (see Table S3). Comparative studies of pathogenicity gene orthologs may reveal important differences in gene function between species and strains, while the prediction of putative non-*C. albicans* Candida pathogenicity genes presents researchers with opportunities for targeted research on candidate virulence genes in these emerging pathogens.

**Candida genes at the intersection of biofilm formation, filamentous growth, and pathogenesis.** From the manual annotation of the experimental literature, we examined the *C. albicans* genes annotated with the terms “biofilm formation,” “filamentous growth,” and “pathogenesis.” Among the 32 gene products that shared annotations for all three of these processes, we identified 10 transcription factors (Ace2, Ada2, Ahrl, Brgl1, Cas5, Cph1, Efg1, Ndt80, Soh1, and Tec1), 12 proteins with cell wall functions (Als1, Als2, Als3, Dse1, Ecm33, Hwp2, Mp65, Pga26, Pmt1, Pmt4, Pmt6, and Sun41), 7 signaling pathway components (Chk1, Cylr1, Mds3, Mkcl, Pde2, Ras1, and Tpk2), and 3 additional factors (Kem1, Not4, and Vam3). Selecting for genes that are involved in all three of these processes, and in addition are also annotated with both “cell adhesion” (or cell adhesion-related terms) and “phenotypic switching,” identified Ras1, Crl1 (Cdc35), and Efg1 as critical regulators of an entire group of processes, including cell adhesion, biofilm formation, filamentous growth, pathogenesis, and phenotypic switching (Fig. 3). Additional genes involved in the RAS signaling pathway that regulate all of the pathogenic traits examined here, except for phenotypic switching, which has not been shown to play a direct role in pathogenesis, include the transcription factor Tec1p and a cyclic AMP (cAMP)-dependent protein kinase catalytic subunit, Tpk2. Overall, our new GO annotations have dramatically improved the ability to predict genes involved in numerous biological processes and have enabled the identification of genes at the intersection of all of the major pathogenic traits of *C. albicans*. These genes play a pivotal role in multiple critical processes of *C. albicans* and may play important roles

**TABLE 4 Numbers of annotations that were inferred computationally from manual annotations of other species**

| Organism with manual annotations | C. albicans | C. glabrata | C. parapsilosis |
|----------------------------------|-------------|-------------|-----------------|
| *S. cerevisiae*                  | 16,240      | 19,772      | 16,843          |
| *S. pombe*                       | 5,262       | 5,092       | 5,343           |
| *C. albicans*                    | 3,661       | 4,690       |                 |
| *C. glabrata*                    | 120         | 161         |                 |
| *C. parapsilosis*                | 21          | 18          |                 |

**FIG 3** Functional relatedness of biofilm formation, filamentous growth, and pathogenesis. (A) Venn diagram showing overlap among genes involved in biofilm formation, filamentous growth, and pathogenesis. (B) Genes involved in biofilm formation, filamentous growth, and pathogenesis, as well as the additional processes of cell adhesion (ca) and phenotypic switching (ps).
in pathogenesis and other virulence traits in non-\textit{C. albicans} \textit{Candida} species.

**DISCUSSION**

The direct curatorial annotation of pathogen genomes is critical for researchers to interpret their experimental data. Analysis of large-scale data sets, such as those obtained from RNA-Seq, microarray, proteomic, and other types of high-throughput experiments, often results in lists of genes of interest, but interpreting these lists is a Herculean and error-prone task in the absence of a central repository of high-quality manual GO annotations. CGD serves this function for \textit{C. albicans} and the other \textit{Candida} species. The entire \textit{C. albicans} literature has been curated for gene information, and CGD curators have made all possible phenotype and GO annotations for every gene described in the literature for \textit{C. albicans}. Such a comprehensive set not only represents a wealth of useful knowledge but also identifies gaps in the current set of questions that have been addressed experimentally, and thus suggests areas for future exploration. Ongoing curation of the literature as it is published, as well as reexamination of existing annotations when the GO is updated with new terms and automated transfer of new GO terms from related species, ensures that the CGD GO annotations are as up-to-date and informative as possible.

During the process of curation, we identified areas of the GO where more informative terms were needed to capture the characterized functions of \textit{Candida} genes. We worked with the GO Consortium to develop and contribute new terms to the ontology and then used these terms to annotate or reannotate specific classes of genes in CGD. Due to the concentrated efforts of the Plant-Associated Microbe Gene Ontology (PAMGO) project (30), the GO is quite well developed to cover areas related to host interactions and pathogenesis, though new GO terms were needed to more precisely capture the functions of genes related to filamentous growth and regulation of adhesion and biofilm formation. Certain biological processes, such as phenotypic switching, lacked representation in the GO altogether. Because these new terms describe important aspects of \textit{Candida} biology, their addition to the GO has enhanced the accuracy and representation of biological processes in GO enrichment analyses with tools such as the GO Term Finder.

GO annotations are useful for making predictions of functions for gene products across numerous species. As part of this project, we made orthology-based predictions of GO annotations for orthologs of experimentally characterized pathogenesis genes. The prediction of these candidate pathogenicity genes in the non-\textit{C. albicans} \textit{Candida} species may aid in the characterization of the pathogenesis of non-\textit{C. albicans} \textit{Candida} species.

We also examined the \textit{C. albicans} genes involved in biofilm formation, filamentous growth, and pathogenesis and identified genes that are annotated for all three processes in \textit{C. albicans}. For the gene products annotated with cell adhesion and phenotypic switching terms, we further identified a set of genes that regulate the constellation of pathogenic traits represented by the terms “cell adhesion,” “biofilm formation,” “filamentous growth,” and “pathogenesis,” and also the process of “phenotypic switching.”

GO enrichment analysis provides a powerful means for identifying processes that are statistically overrepresented among a gene set’s annotations. One caveat to consider with evaluation of such analyses is that the population of manual annotations is biased by the prevalence of particular research topics in the published experimental literature. Indeed, this effect is illustrated by the set of annotations described in this work. For example, the genes that are annotated as contributing to biofilm development are largely dominated by regulators of this process, a fact that reflects the strategy of using transcription factor libraries to specifically identify regulators of biofilm formation (8, 13, 31, 32).

The large number of manual annotations of “filamentous growth” also reflects the attention that this research topic has received in the \textit{C. albicans} literature, and this in turn is reflected in the results of GO enrichment analysis of the \textit{C. albicans} pathogenicity genes. Media such as serum, Lee’s medium, and spider medium (abiotic stimulus, neutral pH, and starvation, respectively) are highly represented in the literature, and manual annotations at CGD reflect these experimental data in the GO annotations. Thus, a large number of annotations for “cellular response to abiotic stimulus” (GO:0071214), “cellular response to starvation” (GO:0009267), and “cellular response to neutral pH” (GO:0036244) are retrieved during GO enrichment analysis of the set of pathogenesis genes.

Our reannotation efforts in this project focused specifically on manually assigned GO annotations in the defined BP categories of adhesion, biofilm formation, filamentous growth, pathogenesis, and phenotypic switching. However, all three GO aspects (molecular function, cellular component, and biological process) are extensively curated at CGD, and CGD also includes a comprehensive set of computational annotations derived from InterPro domain matches, which have particularly high utility in representing predictions within the MF branch of the GO. Thus, GO enrichment analysis at CGD provides researchers the ability to evaluate both function and localization terms, as well as annotations within the biological process branch of the hierarchy.

**ACKNOWLEDGMENTS**

We thank Gail Binkley for CGD Oracle database administration, Stuart Miyasato and Matt Simison for CGD database software and hardware maintenance, and Jane Lomax, Rebecca Foulger, and Paola Roncaglia, editors at the GO Consortium, for assistance in creating new GO terms. We give special thanks to Geraldine Butler, Sarah Maguire, and other members of the Butler laboratory for the ortholog groupings they developed and curated for use in the CGOB (http://cgob3.ucd.ie), which were used for orthology-based GO term predictions.

This work was supported by the National Institute of Dental and Craniofacial Research at the National Institutes of Health (grant R01 DE015873 to G.S.).

**REFERENCES**

1. Arnaud MB, Costanzo M, Skrzypek MS, Binkley G, Lane C, Miyasato SR, Sherlock G. 2005. The \textit{Candida} Genome Database (CGD), a community resource for \textit{Candida albicans} gene and protein information. Nucleic Acids Res. 33:D358–D363.

2. Costanzo MC, Arnaud MB, Skrzypek MS, Binkley G, Lane C, Miyasato SR, Sherlock G. 2006. The \textit{Candida} Genome Database: facilitating research on \textit{Candida albicans} molecular biology. FEMS Yeast Res. 6:671–684.

3. Inglis DO, Arnaud MB, Binkley J, Shah P, Skrzypek MS, Wymore F, Binkley G, Miyasato SR, Simison M, Sherlock G. 2012. The \textit{Candida} genome database incorporates multiple \textit{Candida} species: multispecies search and analysis tools with curated gene and protein information for \textit{Candida albicans} and \textit{Candida glabrata}. Nucleic Acids Res. 40:D667–D674.

4. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Feil R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drake B, Eppig JT, Hill DP, Jona L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS,
Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hancock I, Wortman J, Worman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:D258–D261.

5. Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314:1041–1052.

6. Fitzpatrick DA, O’Gaora P, Byrne KP, Butler G. 2010. Analysis of gene evolution and metabolic pathways using the Candida Gene Order browser. BMC Genomics 25:289–291.

7. Chandra J, Mukherjee PK, Ghannoum MA. 2008. In vitro growth and analysis of Candida biofilms. Nat. Protoc. 3:1909–1924.

8. Finkel JS, Xu W, Huang D, Hill EM, Desai JV, Woolford CA, Nett JE, Taff H, Norice CT, Andes DR, Lanni F, Mitchell AP. 2012. Portrait of Candida albicans adherence regulators. PLoS Pathog. 8:e1002525. doi:10.1371/journal.ppat.1002525.

9. Chandra J, Kuhn DM, Mukherjee PK, Hoyer LL, McCormick T, Ghannoum MA. 2001. Biofilm formation by the fungal pathogen Candida albicans: development, architecture, and drug resistance. J. Bacteriol. 183:5385–5394.

10. Chandra J, Mukherjee PK, Leidich SD, Faddoul FF, Hoyer LL, Douglas LJ, Ghannoum MA. 2001. Antifungal resistance of candidal biofilms formed on denture acrylic in vitro. J. Dent. Res. 80:903–908.

11. Kuhn DM, George T, Chandra J, Mukherjee PK, Ghannoum MA. 2002. Antifungal susceptibility of Candida biofilms: unique efficacy of amphotericin B lipid formulations and echinocandins. Antimicrob. Agents Chemother. 46:1773–1780.

12. Nobile CJ, Andes DR, Nett JE, Smith FJ, Yue F, Phan QT, Edwards JE, Filler SG, Mitchell AP. 2006. Critical role of Bcr1-dependent adhesins in C. albicans biofilm formation in vitro and in vivo. PLoS Pathog. 2:e63. doi:10.1371/journal.ppat.0020063.

13. Nobile CJ, Fox EP, Nett JE, Sorrells TR, Mitrovich QM, Hernady AD, Tuch BB, Andes DR, Johnson AD. 2012. A recently evolved transcriptional network controls biofilm development in Candida albicans. Cell 148:126–138.

14. Blankenship JR, Mitchell AP. 2006. How to build a biofilm: a fungal perspective. Curr. Opin. Microbiol. 9:588–594.

15. Nett JE, Marchillo K, Spiegel CA, Andes DR. 2010. Development and validation of an in vivo Candida albicans biofilm denture model. Infect. Immun. 78:3650–3659.

16. Zijinge V, van Leeuwen MB, Degener JE, Abbas F, Thurnheer T, Gmur R, Harmsen HJ. 2010. Oral biofilm architecture on natural teeth. PLoS One 5:e9321. doi:10.1371/journal.pone.0009321.

17. Gow NA. 1997. Germ tube growth of Candida albicans. Curr. Top. Med. Mycol. 8:43–55.

18. Sudbery P, Gow NA, Berman J. 2004. The distinct morphogenic states of Candida albicans. Trends Microbiol. 12:317–324.

19. Brand A. 2012. Hyphal growth in human fungal pathogens and its role in virulence. Int. J. Microbiol. 2012:517529. doi:10.1155/2012/517529.

20. Slutsky B, Buflo J, Soll DR. 1985. High-frequency switching of colony morphology in Candida albicans. Science 230:666–669.

21. Anderson JM, Soll DR. 1987. Unique phenotype of opaque cells in the white-opaque transition of Candida albicans. J. Bacteriol. 169:5579–5588.

22. Lan CY, Newport G, Murillo LA, Jones T, Scherer S, Davis RW, Agabian N. 2002. Metabolic specialization associated with phenotypic switching in Candida albicans. Proc. Natl. Acad. Sci. U. S. A. 99:14907–14912.

23. Kolotila MP, Diamond RD. 1990. Effects of neutrophils and in vitro oxidants on survival and phenotypic switching of Candida albicans WO-1. Infect. Immun. 58:1174–1179.

24. Geiger J, Vessels D, Lockhart SR, Soll DR. 2004. Release of a potent polymorphonuclear leukocyte chemoattractant is regulated by white- opaque switching in Candida albicans. Infect. Immun. 72:667–677.

25. Lohse MB, Johnson AD. 2008. Differential phagocytosis of white versus opaque Candida albicans by Drosophila and mouse phagocytes. PLoS One 3:e1473. doi:10.1371/journal.pone.0001473.

26. Miller MG, Johnson AD. 2002. White-opaque switching in Candida albicans is controlled by mating-type locus homeodomain proteins and allows efficient mating. Cell 110:293–302.

27. Kachurina N, Turcotte B, Whiteway M. 2012. Motor protein Myo5p is required to maintain the regulatory circuit controlling WOR1 expression in Candida albicans. Eukaryot. Cell 11:626–637.

28. Zordan RE, Miller M, Galgoczy DJ, Tuch BB, Johnson AD. 2007. Interlocking transcriptional feedback loops control white-opaque switching in Candida albicans. PLoS Biol. 5:e256. doi:10.1371/journal.phb.0052056.

29. Pfaffer M. 1996. Nosocomial candidiasis: emerging species, reservoirs, and modes of transmission. Clin. Infect. Dis. 22:589–594.

30. Torto-Alabio T, Collmer CW, Gwinn-Giglio M, Anderson JM, Soll DR. 2009. The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. BMC Microbiol. 9:51. doi:10.1186/1471-2180-9-S1-S1.

31. Homann OR, Dea J, Noble SM, Johnson AD. 2010. A phenotypic profile of the Candida albicans regulatory network. PLoS Genet. 6:e1001070. doi:10.1371/journal.pgen.1001070.

32. Richard ML, Nobile CJ, Bruno VM, Mitchell AP. 2005. Candida albicans biofilm-defective mutants. Eukaryot. Cell 4:1493–1502.

33. Day-Richter J, Harris MA, Haendel M, Gene Ontology OBO-Edit Working Group, Lewis S. 2007. OBO-Edit—an ontology editor for biologists. Bioinformatics 23:2198–2200.