D358–D363 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki003

The *Candida* Genome Database (CGD), a community resource for *Candida albicans* gene and protein information

Martha B. Arnaud*, Maria C. Costanzo, Marek S. Skrzypek, Gail Binkley, Christopher Lane, Stuart R. Miyasato and Gavin Sherlock

Department of Genetics, Stanford University School of Medicine, CCSR 2255, 269 Campus Drive, Stanford, CA 94305-5120, USA

Received August 13, 2004; Revised August 30, 2004; Accepted September 8, 2004

ABSTRACT

The *Candida* Genome Database (CGD) is a new database that contains genomic information about the opportunistic fungal pathogen *Candida albicans*. CGD is a public resource for the research community that is interested in the molecular biology of this fungus. CGD curators are in the process of combing the scientific literature to collect all *C.albicans* gene names and aliases; to assign gene ontology terms that describe the molecular function, biological process, and subcellular localization of each gene product; to annotate mutant phenotypes; and to summarize the function and biological context of each gene product in free-text description lines. CGD also provides community resources, including a reservation system for gene names and a colleague registry through which *Candida* researchers can share contact information and research interests. CGD is publicly funded (by NIH grant R01 DE15873-01 from the NIDCR) and is freely available at http://www.candidagenome.org/.

INTRODUCTION

Candida albicans is the best studied of the human fungal pathogens, and it serves as a model organism for the study of other pathogenic fungi. In recent years, the frequency of fungal infections has steadily grown and although these infections are generally less frequent than bacterial infections, at least two aspects make them increasingly important. First, opportunistic infections in immunocompromised patients represent an increasingly common cause of mortality and morbidity (1,2). Second, many of the currently used antifungal compounds (3,4) are often of limited use because of their toxicity and side effects (5). Furthermore, within the last decade there has been an emergence of anti-fungal drug resistance, which was a rarity in the past (6–10). By serving as a resource for scientists who study fungal biology and pathogenesis, the *Candida* Genome Database (CGD) aims to facilitate progress toward more complete understanding of and effective treatment for fungal diseases.

Before CGD was created, three web sites contained information about the *C.albicans* genome sequence and about *C.albicans* gene products. The Stanford Genome Technology Center undertook the process of sequencing and the difficult challenge of assembling the sequence of this diploid organism (11), and their web site provides options for searching and downloading the genome sequence. CandidaDB, at the Pasteur Institute, was the first freely available *C.albicans* database; it contains sequence-based annotation for assemblies 6 and 19 of the genome sequence (http://genolist.pasteur.fr/CandidaDB/). The third resource was developed by the *Candida* Annotation Working Group, colleagues who came together on a volunteer basis, to analyze the *C.albicans* sequence produced by the Stanford Genome Technology Center. The results of the Annotation Working Group's efforts include a high quality set of gene annotations and gene ontology (GO) terms assigned by sequence-based prediction. The Annotation Working Group's annotation and sequence analysis tools are accessible on a web site hosted at the Biotechnology Research Institute of the National Research Council in Canada (http://candida.bri.nrc.ca/candida/index.cfm).

The *Candida* research community expressed a need for a database with additional features: comprehensive literature curation, to complement the high quality sequence-based annotation already available; a more extensive set of sequence retrieval and analysis tools, similar to those provided at the *Saccharomyces* Genome Database (SGD) (12); and centralized community information, such as a colleague directory

*To whom correspondence should be addressed. Tel: +1 650 736 0075; Fax: +1 650 724 3701; Email: arnaudm@genome.stanford.edu

and a gene name registry. CGD was proposed to meet these needs. CGD is based on the framework of SGD, using the same software, user interfaces, and underlying schema. The format and tools will therefore be familiar to CGD users who are already users of SGD. CGD started with the *Candida* Annotation Working Group's informative data set, and the CGD curators are now adding published material from the literature.

LOCUS PAGE

Similar to SGD, CGD contains gene information organized around locus pages. An example locus page is shown in Figure 1.

The locus page displays basic information about a gene and its product. The gene name is displayed prominently at the top of the page along with all aliases, including names assigned during sequencing and sequence assembly. Also found near the top of the page is the description, which is a concise statement of the most important information known about the gene and the gene product, especially its function, biological context, and physical characteristics. Each gene product is assigned GO terms (13) that describe its molecular functions, its location within the cell, and the biological processes in which it participates. The GO annotation section of each locus page contains a link to the GO annotation page, which shows all GO terms along with the references that were used to make each assignment and the type of evidence that supports it. An example GO annotation page is shown in Figure 2. Each GO term name, both on the locus page and on the GO annotation page, links to a graphical view that allows users to see parent and child relationships for each term, to navigate within the ontologies, and to view summary information about all of the *Candida* genes assigned to any given GO term.

Initially, CGD GO curation has focused on one or a few references that describe each gene product. With time, CGD will collect GO terms comprehensively, such that the database will list all of the papers that support assignment of each term, rather than listing only a more limited set of representative papers. The rationale for assigning GO terms from each paper is that the number of independent pieces of evidence for assignment to a particular GO term can be a measure of confidence in that assignment.

The locus page also contains a mutant phenotype section. This section lists the type of mutation (e.g. homozygous null, heterozygous null, or overexpression) and any corresponding phenotype. At this time, phenotypes are collected from the literature as free-text descriptions. Each phenotype that is displayed on the locus page is hyperlinked to a list of all *C.albicans* genes that share that mutant phenotype. The locus page also presents a link to a page that lists the references in which specific phenotypes are described. This page also contains phenotype details, including additional information about the specific conditions under which some phenotypes have been observed.

LITERATURE INTERFACE

CGD contains a wealth of information about the *C.albicans* scientific literature. This information is available in several formats within the literature guide. The literature guide, which is accessed from the menu on the right-hand side of each locus page, provides a list of papers that characterize a particular gene. These lists were generated by using an automated search of the PubMed database at NCBI, and have been manually screened to eliminate spurious references. Next to each reference there is a list of all the genes described in the paper, and each gene name is hyperlinked to its corresponding locus page. As each reference is curated, curators note whether the paper pertains to any of a set of 45 'literature topics'. These topics are based on the set that is used by SGD, but have been expanded to include additional topics of special interest to the *Candida* research community. The topics include filamentous growth, phenotypic switching, adherence and biofilms, as well as more generalized topics such as function/process, protein physical properties, protein–protein interactions, protein–nucleic acid interactions, post-translational modifications, transcriptional regulation and translational regulation. The complete set of CGD literature guide topics is listed in Table 1. Within the literature guide interface, the reference list may be sorted according to the topic or by curation status (curated or not yet curated). Alternately, users may choose to focus on individual papers. The curated paper view displays the reference information and the abstract, along with a summary of literature guide topics that are assigned to every gene characterized in the paper.

COMMUNITY RESOURCES

CGD seeks to facilitate an interaction among the members of the *C.albicans* research community. Thus, CGD has implemented a colleague registry, by which researchers may share contact information and find others who share research interests or who are experts in a particular topic.

CGD also serves as the keeper of gene name reservations prior to publication. The community conferred this privilege upon CGD at the ASM meeting on Candida and Candidiasis in March 2004. Having a reservation system for gene names benefits the entire community because it helps to reduce conflicts in gene names and prevents the introduction of confusing synonyms into the literature. CGD does not itself assign gene names, but rather collects and maintains a list of current reservations and attempts to mediate resolution of any disputes that may arise. CGD follows gene name guidelines that are based on those used by the *Saccharomyces cerevisiae* research community. Detailed information about choosing and reserving a gene name is found on the CGD web site under nomenclature guide.

CGD also hosts a web page with *Candida* community news and a list of meetings, courses, and related web sites of interest.

CURRENT PROGRESS AND FUTURE DIRECTIONS

The CGD project began in April 2004, and is progressing rapidly. However, there is still much to be done, and there are plans to add additional information and new features.

CGD literature curation is now in progress. As of August 2004, CGD contained more than 900 gene product descriptions, and ~1500 mutant phenotype descriptions and 1500 GO term assignments. The initial release of CGD contained locus

pages for genes that have been characterized in the literature. The *C.albicans* genome contains ~6400 homologous pairs of genes (11); the majority of these genes have not yet been characterized. CGD will contain the entire SC5314 gene complement, with locus pages and sequences for all the genes that were identified in the genome-sequencing project, although not all of this information was included in CGD upon the initial database release. CGD will contain the reference

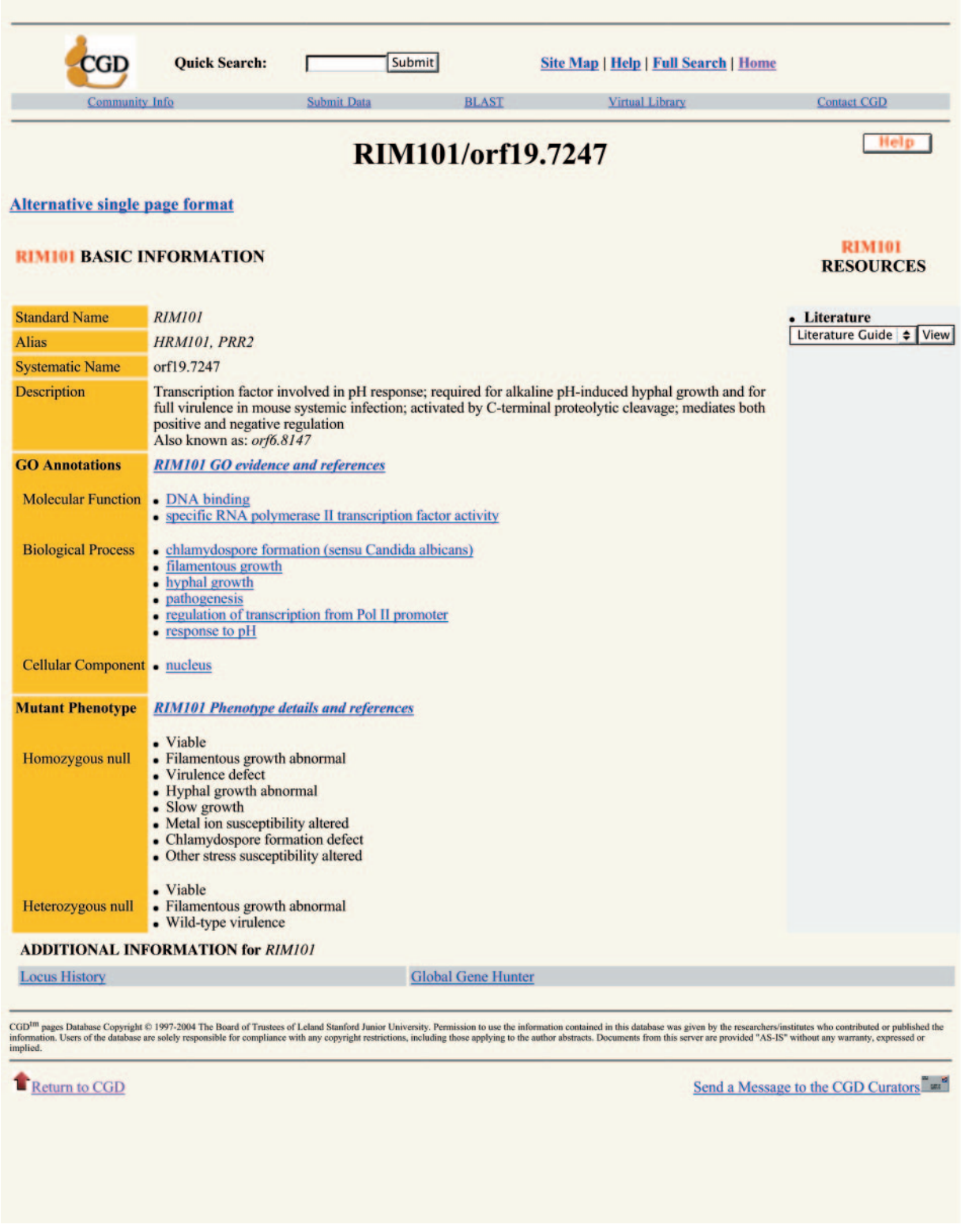

Figure 1. CGD locus page. The locus page presents the basic information about a gene and its product, including names and aliases, a concise description, GO term assignments and mutant phenotypes. The locus page also provides links to additional resources.

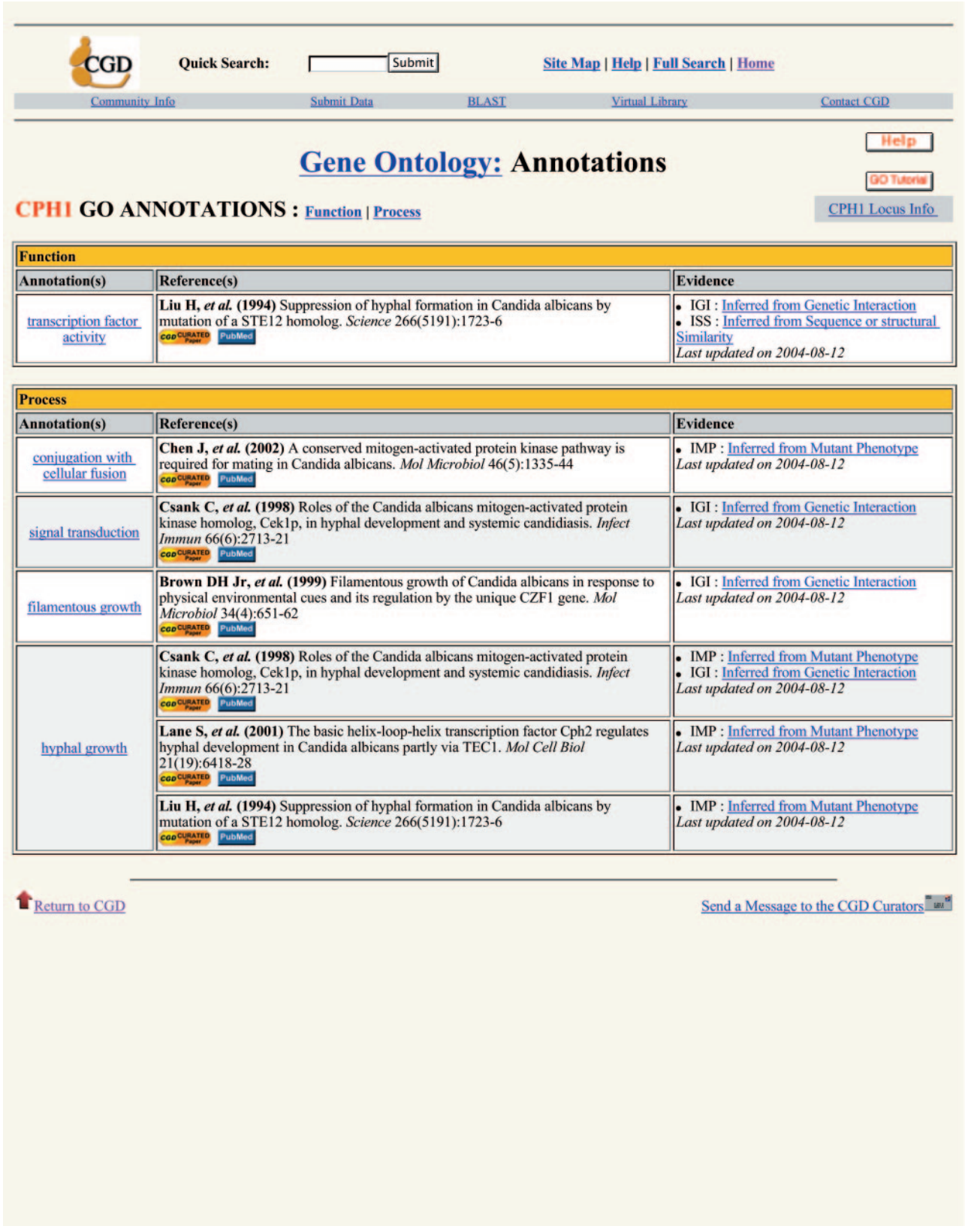

Figure 2. CGD gene ontology (GO) annotation page. The GO annotation page displays each of the GO term assignments along with the references from which these assignments were made, and the types of evidence that support assignment of each GO term.

Table 1. CGD literature topic curation

Genetics/Cell Biology	
Cellular Location	180
Function/ Process	530
Genetic interactions	126
Mutants/ Phenotypes	515
Regulatory Role	69
Signal Transduction	49

Nucleic Acid Information	
DNA/RNA Sequence Features	151
Variant Alleles	26
Mapping	62

Regulation	
Transcriptional Regulation	517
Translational Regulation	2
Regulation of RNA Degradation	4
Regulation of Protein Degradation	1
Regulation of Activity	23
Regulation (Other)	9
RNA Levels and Processing	145
Regulation (Unspecified)	18

Research Aids	
Alias	44
Other Features	0
Strains/ Constructs	649
Techniques and Reagents	86
Genome-Wide Analysis	224

Gene Product Information	
Protein Domains/ Motifs	285
Protein Physical Properties	206
Protein-Protein Interactions	37
Protein-Nucleic Acid Interactions	14
Nucleic acid-Nucleic acid interactions	0
Protein/Nucleic Acid Structure	5
Post-Translational Modifications	77
Substrates/ Ligands/ Cofactors	94
Protein Processing	83

Life cycle	
Filamentous Growth	155
Phenotypic Switching	13

Virulence-related Information	
Virulence	79
Adherence	20
Biofilms	2
Drug Resistance/ Susceptibility	62
Drug Effects	15
Animal Model	92
Clinical Data	39

Related Genes/Proteins	
Fungal Related Genes/ Proteins	428
Non-Fungal Related Genes/ Proteins	126
Disease Gene Related	6
Cross-Species Expression	182

The literature topics are displayed above, along with the number of times each topic had been assigned, as of August 5, 2004. As each paper is curated, literature topics are assigned to all of the genes described in the paper. Each gene has a set of topic assignments from every curated paper that describes the gene. This information is shown in the literature guide interface, which is accessed from the menu on the right-hand side of each locus page.

sequence of the strain SC5314 (11). The *C.albicans* genomic sequence data is scheduled to be added to CGD in the autumn of 2004. Once this information has been incorporated, CGD will provide access to sequence analysis and visualization tools that are similar to those available at SGD, including tools for viewing multiple versions of sequences that have been updated since the original sequence was published. Each locus page currently provides a hyperlink to the *C.albicans* BLAST tool at the Biotechnology Research Institute of the National Research Council in Canada.

In addition, CGD will also provide links between CGD and SGD locus pages, which will provide instant access to information about the *S.cerevisiae* orthologs of *C.albicans* proteins. All CGD data will available for free download at an ftp site that will be linked from our home page.

The current curation efforts are focused on the body of scientific literature that deals with specific *C.albicans* genes by name. However, an additional set of literature exists that concerns more generalized *C.albicans* biology, e.g. drug sensitivity studies or morphological descriptions that do not examine the role of any specific gene product. CGD plans to include these papers in the database and to make literature guide topic assignments. The current set of literature topics may need to be expanded to capture information from this set of papers more effectively. The CGD group seeks input from the research community as to what types of information would be most useful for CGD to collect from such papers.

Within the next year, CGD plans to begin curation of metabolic pathway information. CGD will use the Pathway Tools software (14) to make pathway predictions, and will supplement and validate these predictions by curating pathway information from the published literature.

SUMMARY AND AVAILABILITY

In summary, the *Candida* Genomic Database is a resource modeled after the *Saccharomyces* Genome Database. The CGD contains information about *C.albicans* genes and gene products. CGD is freely available on the web at www.candidagenome.org. CGD also facilitates community interaction by providing a colleague registry and a gene name registry. CGD is being actively developed, and the CGD project staff would like to solicit advice from *Candida* researchers about ways in which CGD may best serve the *C.albicans* research community. Users are encouraged to contact CGD at candida-curator@genome.stanford.edu with comments or suggestions.

ACKNOWLEDGEMENTS

CGD thanks the *Candida* Annotation Working Group, especially Andre Nantel, for their generosity in sharing their entire data set and their enthusiasm for letting CGD house this resource; Judith Berman, Burk Braun, Neil Gow, Pete Magee and Aaron Mitchell for their advice and support; and the *Saccharomyces* Genome Database group for their resources and invaluable assistance. CGD is supported by grant R01 DE15873-01 from the National Institute of Dental and Craniofacial Research at the US National Institutes of Health.

REFERENCES

1. Fisher-Hoch,S.P. and Hutwagner,L. (1995) Opportunistic candidiasis: an epidemic of the 1980s. *Clin. Infect. Dis.*, **21**, 897–904.
2. Groll,A.H., De Lucca,A.J. and Walsh,T.J. (1998) Emerging targets for the development of novel antifungal therapeutics. *Trends Microbiol.*, **6**, 117–124.
3. Vanden Bossche,H. (1995) In Lyr,H. (ed.), *Modern Selective Fungicides: Properties, Applications, Mechanism of Actions*. Gustav Fisher Verlag, Jena, pp. 431–484.
4. Bennet,J.E. (1996) In Taylor,P. (ed.), *Goodman and Gilman's The Pharmacological Basis of Therapeutics*. Pergamon, Elmsford, NY, pp. 1165–1181.
5. Georgopapadakou,N.H. and Walsh,T.J. (1996) Antifungal agents: chemotherapeutic targets and immunologic strategies. *Antimicrob. Agents Chemother.*, **40**, 279–291.
6. Smith,D., Boag,F., Midgley,J. and Gazzard,B. (1991) Fluconazole resistant candida in AIDS. *J. Infect.*, **23**, 345–346.
7. Siegman-Igra,Y. and Rabaw,M.Y. (1992) Failure of fluconazole in systemic candidiasis. *Eur. J. Clin. Microbiol. Infect. Dis.*, **11**, 201–202.
8. Johnson,E.M., Warnock,D.W., Luker,J., Porter,S.R. and Scully,C. (1995) Emergence of azole drug resistance in *Candida* species from HIV-infected patients receiving prolonged fluconazole therapy for oral candidosis. *J. Antimicrob. Chemother.*, **35**, 103–114.
9. Denning,D.W. (1995) Can we prevent azole resistance in fungi? *Lancet*, **346**, 454–455.
10. Boschman,C.R., Bodnar,U.R., Tornatore,M.A., Obias,A.A., Noskin,G.A., Englund,K., Postelnick,M.A., Suriano,T. and Peterson,L.R. (1998) Thirteen-year evolution of azole resistance in yeast isolates and prevalence of resistant strains carried by cancer patients at a large medical center. *Antimicrob. Agents Chemother.*, **42**, 734–738.
11. Jones,T., Federspiel,N.A., Chibana,H., Dungan,J., Kalman,S., Magee,B.B., Newport,G., Thorstenson,Y.R., Agabian,N., Magee,P.T. *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl Acad. Sci. USA*, **101**, 7329–7334.
12. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al.* (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32** (Database issue), D311–D314.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology consortium. *Nature Genet.*, **25**, 25–29.
14. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18** (Suppl. 1), S225–232.