D428–D432 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 6 October 2009*
doi:10.1093/nar/gkp836

New tools at the *Candida* Genome Database: biochemical pathways and full-text literature search

Marek S. Skrzypek, Martha B. Arnaud*, Maria C. Costanzo, Diane O. Inglis, Prachi Shah, Gail Binkley, Stuart R. Miyasato and Gavin Sherlock

Department of Genetics, Stanford University Medical School, Stanford, CA 94305-5120, USA

Received August 31, 2009; Accepted September 19, 2009

ABSTRACT

The *Candida* Genome Database (CGD, http://www.candidagenome.org/) provides online access to genomic sequence data and manually curated functional information about genes and proteins of the human pathogen *Candida albicans*. Herein, we describe two recently added features, *Candida* Biochemical Pathways and the Textpresso full-text literature search tool. The Biochemical Pathways tool provides visualization of metabolic pathways and analysis tools that facilitate interpretation of experimental data, including results of large-scale experiments, in the context of *Candida* metabolism. Textpresso for *Candida* allows searching through the full-text of *Candida*-specific literature, including clinical and epidemiological studies.

INTRODUCTION

The *Candida* Genome Database (CGD, http://www.candidagenome.org/) is an online resource of gene and protein information for the opportunistic fungal pathogen *Candida albicans*. CGD was launched in 2004 as a resource for the *Candida* research community, to maintain the reference version of the *C. albicans* genome sequence and annotation, and to curate and provide access to gene-specific information for *C. albicans* protein- and RNA-coding genes in the nuclear and mitochondrial genomes, as well as information about other chromosomal features such as repeat sequences and transposons, centromeres and pseudogenes. The CGD curators manually curate published experimental data to collect gene names and aliases, write gene descriptions, make Gene Ontology assignments, collect mutant phenotypes and assemble lists of references specific to each gene. The information content of CGD as of August 2009 is summarized in Table 1.

The CGD website also provides tools to facilitate data access and analysis. Search tools allow easy access to the gene information on the CGD Locus Summary pages via gene name searches or gene property based queries, and the retrieval of DNA and protein sequence for any gene, set of genes or chromosomal region. CGD also provides a genome browser, sequence comparison by BLAST, DNA- or protein-sequence-based pattern matching, a primer design tool and a restriction mapping tool, as well as tools for retrieval of bulk data in batch and data files for download. An at-a-glance, daily updated overview of the status of the *C. albicans* genome is shown by the Genome Snapshot tool, which summarizes the fraction of characterized open reading frames (ORFs), the changes that have been made to the reference sequence and annotation and a high-level overview of Gene Ontology annotation in CGD. CGD is also a community resource for gene naming, conference and job announcements and for sharing colleague and lab information. As of August 2009 more than 800 researchers had registered as CGD Colleagues, and usage of CGD's website had reached almost 3 million hits.

One of CGD's important community functions is to maintain the most up-to-date version of the *C. albicans* sequence and annotation, Assembly 21. In 2008, in collaboration with researchers at the Broad Institute, CGD curators undertook a systematic targeted re-evaluation of problematic regions in the sequence assembly, and incorporated into Assembly 21 a set of sequence and annotation refinements based on comparative genomic analysis of eight closely related *Candida* species and newly generated *C. albicans* sequence data (1). Based on sequence conservation, 73 new ORFs were added to the gene catalog and, due to the lack of conservation, 181 previously annotated ORFs were re-classified as 'Dubious', indicating that they are unlikely to encode a *bona fide* gene product. Curators individually examined all of the sequence traces covering regions where gaps or insertions had been introduced by annotators to compensate for likely sequencing errors that interrupted ORFs. In total, 697 sequence corrections were made, and the coordinates of 63 ORFs were updated.

Two major additions that significantly expand CGD's utility to researchers are discussed in detail below. The *Candida* Biochemical Pathways tool organizes both

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

Table 1. Summary of CGD information content

Feature type	Number of features	Manual GO annotations (number of references used)	Computational GO annotations	Phenotype annotations (number of references used)
ORF, verified	1045	4835 (1163)	3790	5450 (814)
ORF, uncharacterized	4971	979 (104)	13 751	2772 (71)
RNA-coding	166	149 (8)	264	3 (1)

Manual Gene Ontology (GO) annotations are derived from published experiments that focus on specific genes or from published experiments that use high-throughput methods. Computational GO annotations are predictions based on computational analyses and are not individually reviewed.

characterized and putative gene products within the metabolic framework of the cell, summarizes the current state of knowledge about *Candida* metabolism, and, by identifying gaps in our knowledge, suggests directions for future research. Textpresso is a text-mining tool that makes it possible to run user-designed queries through the entire body of *Candida*-relevant literature, including publications of clinical and epidemiological data that are outside the scope of gene-specific literature curation in CGD.

BIOCHEMICAL PATHWAYS

The *Candida* Biochemical Pathways in CGD were created using Pathway Tools, a software suite developed by the Bioinformatics Research Group at SRI International (2). Each pathway, reaction, enzyme and compound in CGD has its own web page report. The Pathway page contains a diagram of the pathway, with a user-selectable level of detail: at the most detailed level, the structures of each intermediate and all cofactors are displayed on the pathway diagram, as well as the gene and enzyme names. It also contains a summary, written by CGD curators, that describes what is known in *Candida* species (with a focus on *C. albicans*) about the pathway and the enzymes that participate in it, and contains a list of published references for the pathway information (Figure 1).

The Pathway Tools software suite contains modules for generating organism-specific databases of compounds, enzymes, reactions and metabolic pathways as well as tools for data visualization and analysis. The initial set of predicted biochemical pathways was generated with the PathoLogic module, which used *C. albicans* enzymatic activities identified by Gene Ontology curation in CGD in conjunction with two reference datasets: SRI's reference database of biochemical reaction and pathway information, MetaCyc (3), and the set of pathways curated at the *Saccharomyces* Genome Database (SGD) (4). The software predicted that a pathway existed in *C. albicans* if at least one enzyme from that pathway in MetaCyc or SGD was identifiable among the *C. albicans* gene products. Since not all of the enzymes were found in many of the predicted pathways, another module, the Pathway Hole Filler, was used to identify genes encoding candidate enzymes for the missing reactions (the 'pathway holes'). The Pathway Hole Filler was configured to compare GenBank sequences associated with each of the enzymes known to carry out the reaction in other organisms to the ORF sequences from CGD, and where significant similarity was found, it assigned candidate *C. albicans* genes to these activities, thus predicting which gene might fill the 'pathway hole' (5).

The parameters for the automatic pathway generation were intentionally set at a relatively low stringency so that borderline predictions could be subjected to curatorial review rather than being automatically rejected. Consequently, the initial pathway set also contained a number of spurious and redundant pathways. CGD curators reviewed the pathway list, identified relevant literature for the pathways, removed spurious predictions and collected lists of relevant citations that are displayed on each pathway page in the database. A number of new *Candida* pathways were added, such as those for farnesol, oxylipin, selenocysteine, xylose/xylitol and glucosylceramide metabolism. In an ongoing effort, CGD curators are reviewing each pathway in detail, making updates to the pathway structure or reactions where necessary and linking the CGD Pathway page to the corresponding pathway(s) in SGD. The literature relevant to the pathway in *C. albicans* and other related species is reviewed and summarized on the Pathway page. In many cases, information about a pathway is synthesized from a broad-based survey of the literature that includes characterization performed in *C. albicans* and *Candida*-related species, as indicated in the text of the summary on the Pathway page. In total, 181 pathways were added to CGD from the initial predicted set of 408 pathways, an additional 15 pathways were added *de novo*, and subsequent curation has refined the list to 159 pathways that are currently represented in CGD as of September 2009.

The Biochemical Pathways in CGD can be accessed via the Pathways link under the Search Options section on the CGD home page. This link opens the main Pathway Tools Query Page (http://pathway.candidagenome.org/), which provides tools for searching and browsing pathway data. The Query box allows searching for a pathway, a protein name, a reaction or a compound; reactions and proteins can be searched by name or E.C. number. The Browse Ontology box allows browsing of the pathways, E.C. numbered reactions and compounds in the Pathway Tools, and the hierarchical structures in which they are organized. For example, the pathways are classified into categories including Biosynthesis, Degradation/Utilization/Assimilation and Generation of Precursor Metabolites and energy, with each of these classes being broken down further into more specific subclasses. The query page also provides an option to choose from an

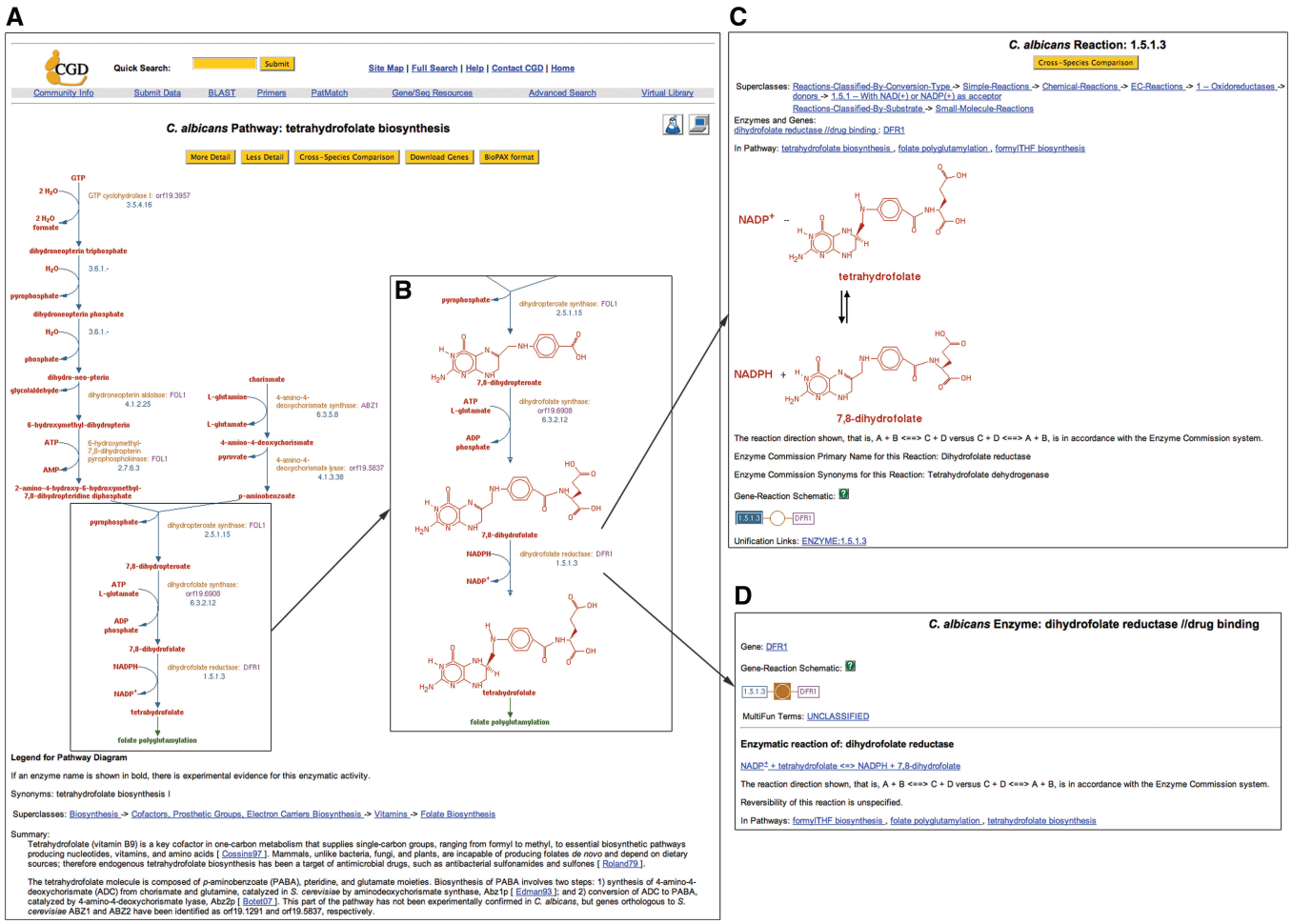

Figure 1. *C. albicans* biochemical pathways. Pathways may be viewed at five different levels of detail, and the names of each compound, enzyme and protein are hyperlinked to their respective report pages. (**A**) Part of the web page report for the tetrahydrofolate biosynthesis pathway, displayed at an intermediate detail level. (**B**) A fragment of the pathway shown at high detail. (**C**) The report page for the reaction catalyzed by Dfr1p. (**D**) The report page for the enzymatic activity of Dfr1p.

alphabetical list of all the pathways, proteins or compounds present in the database.

Any specific pathway page can easily be found by a name-based query using CGD's Quick Search box, which is present at the top of most pages in CGD. This tool performs keyword searches through major categories of information, including pathway names. Individual pathways are also accessible via hyperlinks from the Locus Summary pages of the participating genes.

The ability to analyze gene functions in the context of biochemical pathways is particularly important in *Candida* research because of the major focus in this field on finding drug targets and investigating mechanisms of drug resistance. The Pathway Tools suite provides a module for such analyses, the Pathway Tools Omics Viewer, which is accessible from the main Pathway Tools Query Page. This tool allows the results of large-scale experiments (e.g. microarray expression, proteomics) to be superimposed on the diagram of biochemical pathways, thus presenting a graphical overview of the response of the entire *Candida* metabolic profile to a particular condition or treatment. In addition, a collection of datasets can be used and the diagram can be animated to show, for instance, changes in gene expression over time.

TEXTPRESSO FOR *CANDIDA*

Textpresso is a text-mining tool developed at Wormbase (6). Adapted at CGD, it allows keyword searches through >16 500 pre-screened, *Candida*-related full-text journal articles from the CGD literature database. The tool conducts the search in the entire text of each of the articles, making use of CGD's collection of full-text literature. Although each result displays only a small snippet of text from the article, it allows users to efficiently identify papers of interest for additional follow-up.

The main literature curation pipeline at CGD focuses on information pertaining to specific genes or proteins, which have Locus Summary pages in the database. However, there is a vast amount of relevant *Candida* literature that does not deal with specific genes or identifiable proteins, for instance, clinical or epidemiological

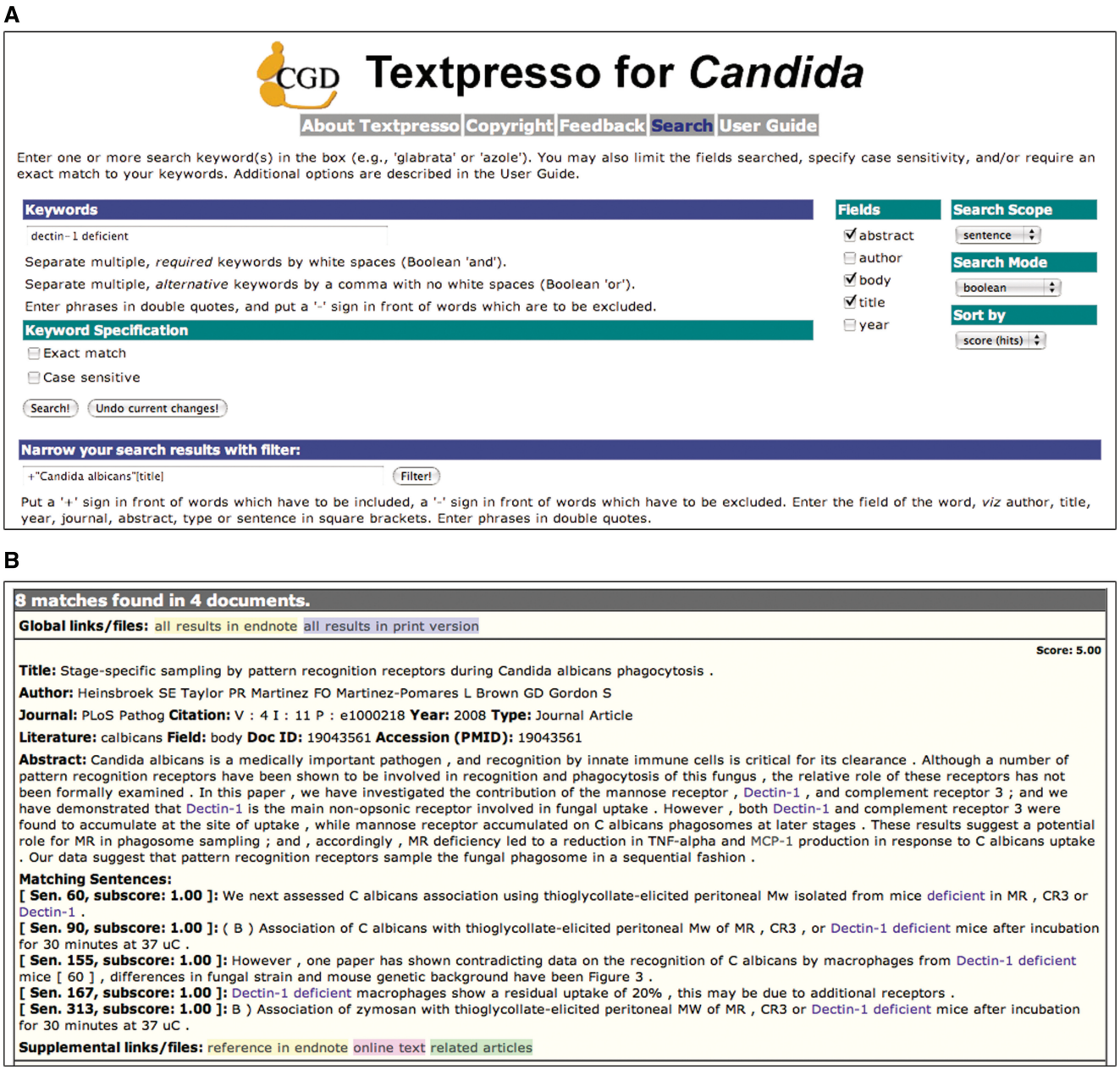

Figure 2. Textpresso for *Candida*. (**A**) Entry of search criteria. An example search for the keywords 'dectin-1' and 'deficient' is shown. The results were also filtered to show only those articles that contain '*C. albicans*' in the title. (**B**) A portion of the output, which is customizable, showing the abstract and matching sentences for an article.

studies and analyses of drug treatments or host responses to infection. The Textpresso for *Candida* tool now makes it possible to access this literature in CGD, enabling equally powerful searches within the text of gene-based papers as well as the more topic-based, gene-independent *Candida* papers that are included in the CGD literature collection.

The Textpresso for *Candida* search engine (http://textpresso.candidagenome.org/cgi-bin/textpresso/search) performs keyword or phrase searches. In constructing the query, a user can input multiple keywords combined either with Boolean "AND", where all the keywords are required, or with Boolean 'OR', where the keywords are treated as alternatives. The query can also specify words to exclude from the search results. The default search mode includes the entire body of text, but the search can be limited to particular fields in the article, such as abstracts or titles. Search results are returned in a form that a user can easily customize to highlight the most desirable pieces of information (Figure 2). For example, the output may show entire sentences containing the search keywords, so that it is possible for the user to quickly evaluate the context in which the keyword was found.

FUTURE DIRECTIONS

Curation of *Candida* Biochemical Pathways is an ongoing project. CGD curators will continue reviewing the existing and newly published literature relevant to *C. albicans* metabolism, incorporating new pathways as they are characterized, and augmenting the data associated with the pathways already entered in CGD. Pathways associated with pathogenicity and drug resistance are of particular interest.

The corpus of literature available to Textpresso for *Candida* will be periodically updated to include newly published articles. In addition, we will introduce biological topic-based curation into our pipeline to augment the accessibility of the literature to CGD users, including papers that do not deal with particular genes, and which therefore fall outside the scope of the current locus-based curation pipeline. We will extend the manual literature curation procedures already in place to construct high-quality bibliographies of articles pertaining to specific topics, for instance, *Candida* response to drugs, clinical studies or host response to *Candida* infection.

CGD strives to facilitate the progress of the *Candida* research community by providing high-quality curation of the *Candida* literature along with tools for accessing and analyzing genomic information. The CGD curators welcome comments or suggestions from the research community at any time, at candida-curator@genome.stanford.edu.

ACKNOWLEDGEMENTS

The authors would like to thank the Pathway Tools support team from the Bioinformatics Research Group at SRI International, the Textpresso developers at WormBase and Mike Cherry and the other members of the SGD group for their help and support.

FUNDING

National Institute of Dental and Craniofacial Research at the US National Institutes of Health [grant no. R01 DE015873]. Funding for open access charge: National Institutes of Health [grant no.RO1 DE015873].

Conflict of interest statement. None declared.

REFERENCES

1. Butler,G., Rasmussen,M.D., Lin,M.F., Santos,M.A., Sakthikumar,S., Munro,C.A., Rheinbay,E., Grabherr,M., Forche,A., Reedy,J.L. *et al*. (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, **459**, 657–662.
2. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**, S225–S232.
3. Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.
4. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al*. (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
5. Green,M.L. and Karp,P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.
6. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.