FungiDB: an integrated functional genomics database for fungi

Jason E. Stajich1,*, Todd Harris2, Brian P. Brunk3,4, John Brestelli4,5, Steve Fischer4,5, Omar S. Harb3,4, Jessica C. Kissinger6,7, Wei Li3,4, Vishal Nayak5,4, Deborah F. Pinney4,5, Chris J. Stoeckert Jr4,5 and David S. Roos3,4

1Department of Plant Pathology & Microbiology, University of California, Riverside, CA 92521, 2Hi-Line Informatics, Bozeman, MT, 3Department of Biology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, 4Penn Center for Bioinformatics, 5Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, 6Center for Tropical & Emerging Global Diseases and 7Department of Genetics and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

Received August 16, 2011; Revised October 6, 2011; Accepted October 8, 2011

ABSTRACT

FungiDB (http://FungiDB.org) is a functional genomic resource for pan-fungal genomes that was developed in partnership with the Eukaryotic Pathogen Bioinformatic resource center (http://EuPathDB.org). FungiDB uses the same infrastructure and user interface as EuPathDB, which allows for sophisticated and integrated searches to be performed using an intuitive graphical system. The current release of FungiDB contains genome sequence and annotation from 18 species spanning several fungal classes, including the Ascomycota classes, Eurotiomycetes, Sordariomycetes, Saccharomycetes and the Basidiomycota orders, Pucciniomycetes and Tremellomycetes, and the basal ‘Zygomycete’ lineage Mucormycotina. Additionally, FungiDB contains cell cycle microarray data, hyphal growth RNA-sequence data and yeast two hybrid interaction data. The underlying genomic sequence and annotation combined with functional data, additional data from the FungiDB standard analysis pipeline and the ability to leverage orthology provides a powerful resource for in silico experimentation.

INTRODUCTION

The recent dramatic increase in the number and scale of genome sequence and functional genomic data (i.e. proteomic, microarray, RNA-sequence, ChIP–ChIP, etc.) has made it increasingly challenging for scientists to navigate through the milieu of data. Importantly, it has become essential to be able to interrogate data sets from multiple genomes in an integrated fashion. To this end, FungiDB (http://fungidb.org) was developed as a resource for genomic and functional genomic data across the fungal kingdom.

FungiDB was developed in partnership with the NIAID-funded Eukaryotic Pathogen Bioinformatic Resource Center (http://eupathdb.org) (1). As such this resource uses the same database structural framework and employs the graphical strategies Web Development Kit (WDK) search interface (2). Current genomes in FungiDB are primarily obtained via the Broad Institute (http://broadinstitute.org) and the dedicated Aspergillus, Candida and Saccharomyces Genome resources (3–5).

FungiDB provides a data-mining interface to the comparative and functional genomic data of multiple species of fungi that differs from the species-focused resources of SGD, CGD and AspGD and provides an integrated query system as part of the WDK and GUS database structure. FungiDB differs from other resources such as Ensembl Fungi (6), the Joint Genome Institute’s Mycocosm (http://jgi.doe.gov/fungi) and IMG tools (7) or Microbes Online (8), which provide complementary data query or visualization tools but do not have the data mining capabilities and broad cross-species comparisons that are possible with the WDK search interface.

Data are currently obtained directly from providers at sequencing centers, GenBank and associated functional data repositories (GEO or SRA), or the key model organism databases SGD, CGD and AspGD. The initial 1.0–β release focused on sets of genomes of a cluster of Aspergillus species, and key species available from the

*To whom correspondence should be addressed. Tel: +1 510 642 8441; Fax: +1 510 642 4995; Email: jason.stajich@ucr.edu

© The Author(s) 2011. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Basidiomycota primarily provided by the Broad Institute or the Joint Genome Institute.

GENOMES AND DATA IN THE CURRENT RELEASE OF FUNGIDB

The current release of FungiDB contains genome sequence and annotation from 17 annotated fungal genomes (Aspergillus clavatus, A. flavus, A. fumigatus, A. nidulans, A. niger, A. terreus, Candida albicans, Coccidioides immitis (strain RS), Cryptococcus neoformans, Fusarium graminearum, F. oxysporum, Fusarium verticillioides (Gibberella moniliformis), Magnaporthe oryzae, Neurospora crassa, Puccinia graminis, Rhizopus oryzae and Saccharomyces cerevisiae) and one unannotated strain (H538.4) of C. immitis, providing 650 Mb of sequence and 207111 genes. Specific details about these genomes are available in Table 1.

FungiDB also contains several functional genomic data sets: (i) EST data obtained from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) (23) for A. flavus, A. terreus, A. niger and G. moniliformis; (ii) cell cycle microarray data for S. cerevisiae based on different synchronization methods (24, 25); (iii) RNA-sequence data from R. oryzae during hyphal growth (Stajich, J.E., Sain, D., Abramyan, J., unpublished data); and (iv) yeast two hybrid data from S. cerevisiae (5).

Genomes and annotation in FungiDB are processed through the same analysis pipeline, which provides additional data, including InterPro domains (26), gene ontology term association (27), signal peptide predictions (28), transmembrane domain predictions, open reading frame predictions, BLAST against the non-redundant genome database at the National Center for Biotechnology Information (NCBI), orthology prediction based on OrthoMCL (29) and synteny prediction. Pipeline details are available at http://s.fungidb.org/pipeline_methods.

HOW TO USE FUNGIDB

The home page

The FungiDB web interface is designed to provide the user with convenient and straightforward access to the underlying data (Figure 1). The home page is divided into three main sections, the banner (Figure 1A), information and help menus (Figure 1B) and searches and tools (Figure 1C). The banner section appears on all FungiDB pages providing users with quick access to GeneID and text searches, 'contact us' form, registration/login and information from any page on the site. Creating an account and logging in allows search strategies to be saved and shared, gene lists to be saved, and to create gene associated annotation comments that are attributed to the author. The gray tool bar section of the banner consists of a series of mouse-over menus that link to all tools and searches in this resource. The expandable information and help menus on the left-hand side of the home page (Figure 1B) provide access to a data summary table of all data in FungiDB, a news section, community resources (including useful links and upcoming events), web tutorials and additional help and information. The central section contains links to all searches and tools in FungiDB, and clicking on the plus symbols (arrow in Figure 1C) expands the various categories to reveal the underlying searches (Figure 1C). These include searches that return sets of genes (left column of Figure 1C), searches that return other data types such as expressed sequence tags (ESTs), open reading frames (ORFs), genomic segments (these include DNA motifs) and genomic sequences (i.e. scaffolds and chromosomes) (central column of Figure 1C) and tools

Table 1. Species in the current release of FungiDB

<table>
<thead>
<tr>
<th>Phylum</th>
<th>(sub)Class</th>
<th>Species</th>
<th>Strain</th>
<th>Genome size (Mbs)</th>
<th>No. of genes</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascomycota</td>
<td>Eurotiomycetes</td>
<td>Aspergillus clavatus</td>
<td>NRRL1</td>
<td>27.86</td>
<td>9413</td>
<td>(9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aspergillus flavus</td>
<td>NRRL3357</td>
<td>36.79</td>
<td>12730</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aspergillus fumigatus</td>
<td>AF293</td>
<td>29.39</td>
<td>10067</td>
<td>(10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aspergillus nidulans</td>
<td>FGSC A4</td>
<td>30.5</td>
<td>10865</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aspergillus niger</td>
<td>ATCC 1015</td>
<td>37.20</td>
<td>6679</td>
<td>(12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aspergillus terreus</td>
<td>NIH 2624</td>
<td>29.33</td>
<td>10564</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coccidioides immitis</td>
<td>H538.4</td>
<td>27.73</td>
<td>10640</td>
<td>(13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coccidioides immitis</td>
<td>RS</td>
<td>28.95</td>
<td>9878</td>
<td>(14)</td>
</tr>
<tr>
<td>Sordariomycetes</td>
<td>Fusarium graminearum</td>
<td>Fusarium oxysporum</td>
<td>PH-1 (NRRL 31084)</td>
<td>36.45</td>
<td>13605</td>
<td>(15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fusarium verticillioides (Gibberella moniliformis)</td>
<td>7600</td>
<td>41.78</td>
<td>14457</td>
<td>(16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Magnaporthe oryzae</td>
<td>70-15</td>
<td>41.70</td>
<td>11385</td>
<td>(17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neurospora crassa</td>
<td>OR74A</td>
<td>41.04</td>
<td>10154</td>
<td>(18)</td>
</tr>
<tr>
<td>Saccharomyces</td>
<td>Candida albicans</td>
<td>SC5314</td>
<td>14.32</td>
<td>6444</td>
<td></td>
<td>(19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Saccharomyces cerevisiae</td>
<td>S288c</td>
<td>12.16</td>
<td>6926</td>
<td>(20)</td>
</tr>
<tr>
<td>Basidiomycota</td>
<td>Pucciniomycetes</td>
<td>Puccinia graminis f.sp. tritici</td>
<td>CRL 75-36-700-3</td>
<td>88.65</td>
<td>21065</td>
<td>(21)</td>
</tr>
<tr>
<td>Tremellomycetes</td>
<td></td>
<td>Cryptococcus neoformans var. grubii</td>
<td>H99</td>
<td>18.88</td>
<td>7121</td>
<td>NA</td>
</tr>
<tr>
<td>Mucomycotina</td>
<td></td>
<td>Rhizopus oryzae (Rhizopus delemar)</td>
<td>RA 99-880</td>
<td>46.09</td>
<td>17652</td>
<td>(22)</td>
</tr>
</tbody>
</table>

All genomes were downloaded from their sources in January 2011. Additional information is available at http://s.fungidb.org/GenomeDataTypes.
in the right-hand column of Figure 1C, including access to the GMOD genome browser (30), BLAST (31), the sequence retrieval tool, access to web services searches and a list of fungal-related literature (based on PubMed searches).

Performing a search in FungiDB

A search in FungiDB is initiated by selecting a search type, which opens a new page with additional parameters that can be set. For example, selecting a 'genes by species' search opens a new page where a user can select which organisms to include in the search (Figure 1C insert, green box). Once a search is executed, results are displayed as the first step in a search strategy (see building a search strategy below). In the example shown in Figure 2, a search for all genes in all species in FungiDB returns 207,111 genes (note the number of genes is displayed in the first step of the strategy, yellow box Figure 2A). The distribution of the gene results from step 1 among the various species in FungiDB is displayed in a filter table below the strategy (Figure 2B). A user can toggle between the various species-specific results by selecting the number in the desired cell. The actual results list of what is selected (yellow highlighted step and cell) is displayed in a dynamic table below the filter table (Figure 2C). Users may navigate through the results by selecting specific page numbers, clicking on next/previous (or first/last), and choosing the number of results to display/page by selecting 'advanced paging'. In addition, columns may be added to the results table by clicking on 'select columns' and choosing which columns to display (green insert, Figure 2C). The columns in the results table may be removed (click on the 'x' to the right of a column title), moved (drag and drop) or sorted (clicking on the up/down arrows to the left of a column title). Individual items in the results table may be added to the basket (see below) by clicking on the basket icon (arrow, Figure 2C). Also, individual records (for example gene pages) may be accessed by clicking on the item ID in the first column of each result list (Figure 2D).

BUILDING A SEARCH STRATEGY IN FUNGIDB

The search strategy system implemented in FungiDB (1) is designed to encourage users to run in silico experiments. After running the initial search, the search strategy may be expanded by clicking on add step (red button, Figure 2A). Once the add step button is engaged, a popup (Figure 2A, green box) of all searches in FungiDB is revealed and the results of any chosen search may be combined with the results of the previous search using a set operation (intersect, union or minus). Steps in a strategy may be viewed, revised, renamed and developed further by nesting or deleting. Furthermore, entire search strategies may be renamed, copied, saved and shared with a unique strategy URL or deleted.
The multi-step search strategy in Figure 3 defines potential drug targets in *C. neoformans* by identifying enzymes that may be associated with growth using cross-species comparisons. Step 1 of the strategy identifies all genes with an enzyme commission (EC) number (available data in FungiDB is from *S. cerevisiae*) to start with a set of known enzymes. Step 2 combines the results of Step 1 with all genes in FungiDB that have an associated metabolic process gene ontology (GO) term (automatically assigned to all genomes in FungiDB as part of the standard analysis pipeline) using a union operation to identify additional potential enzymes with this GO annotation. Step 3 transforms the results from Step 2 into their orthologs in all fungal organisms in FungiDB into *C. neoformans* genes. In Step 4, the results are intersected with genes that are upregulated during hyphal growth in *R. oryzae* [note that this step is actually a nested strategy (Figure 3B) where the *R. oryzae* results are transformed into their FungiDB orthologs] to find genes that are likely involved during active growth of the fungus and its cell wall. Since an actively growing RNA-Seq data set was not available for *C. neoformans*, one can use the *R. oryzae* set to filter the genes. Step 5 is a result of intersecting Step 4 with all genes in FungiDB that do not have orthologs in bacteria, archaean and non-fungal eukaryotes (based on OrthoMCL results) in order to find genes which are unique to Fungi. The final list of results is then filtered on *C. neoformans* using the filter table as described above to reveal 32 genes. While clearly a hypothetical list of potential drug targets, it is intriguing to note that several
known drug targets appear in the list including, the glucan 1,3 β-glucosidase gene (S. cerevisiae FKS1 homolog), which is the target of echinocandins, and an oxysterol-binding protein involved in ergosterol synthesis (target of amphotericin B). This strategy maybe viewed and shared using the following URL: http://fungidb.org/fungidb/im.do?s=c485b8fe6e1769bd.

### ADDITIONAL FEATURES IN FUNGIDB

#### Favorites

The favorites tool allows users of FungiDB to bookmark their favorite genes for quick future access. Adding or removing a gene to the favorites can be done by clicking on the favorites icon (star) on gene pages (Figure 2D). Accessing the favorites page is achieved via the favorites link in the gray tool bar (Figure 1A). Genes in the favorites page can be assigned to user-defined projects and free text can be added to each gene.

#### Basket

The basket tool allows a user to cherry pick individual results (i.e. genes, ESTs and genomic sequences) and place them in the basket for further analysis. Adding items to the basket can be achieved by clicking on the basket icon in a list of results (Figure 2C) or at the top of a record page (Figure 2D). Once a desired set of items have been added to the basket, a user may add the basket contents to a search strategy and analyze the results by combining with other data in FungiDB.

#### Weighted searches

Weighting allows a user to add arbitrary weights to steps in a strategy. As the strategy grows, the results are sorted by the sum of their weights. For this feature to function properly, steps need to be combined using the union operation. The benefit of using this feature is that items that do not meet all the criteria in a strategy are not lost but rather appear in the final result list, albeit ranked lower down the list. As an example, the strategy described in Figure 3A was weighted (each step was weighted on a scale from 1 to 10): details are available at http://fungidb.org/fungidb/im.do?s=3921e20d384bd503. Although the final result list expands dramatically (as expected since the union operation was used to combine steps), it is returned as a ranked list. Additional high-interest genes that appear as top hits in the weighted search include sulfite reductase.

#### Genomic colocation

This tool enables users to identify genomic features based on their relative location to each other on the genome. For example, a user can identify a DNA motif in their favorite fungal genome and then find genes that have one of these motifs within 1000 bp of their 5′-end. One can then ask questions about this list of motif containing genes such as when they are expressed or what GO terms are associated with them.

#### User comments

Comments by users may be added to record pages (such as gene pages) in FungiDB by clicking on the add comment link (Figure 2D and E). The comment form provides a user with a quick and straightforward mechanism to enhance FungiDB with their information, which may include free text, references (by entering PubMed or digital object IDs), NCBI accession numbers, images and documents, and genomic location coordinates. Moreover, a comment may be linked to multiple records in the database using their IDs. Once a user submits a comment it appears immediately for all users on the
record page and becomes searchable via the text search functionality.

DATA DOWNLOAD AND SEQUENCE RETRIEVAL

Data in FungiDB are conveniently available for bulk download from the ‘Data Files’ section accessible from the ‘Downloads’ menu item in the gray tool bar (Figure 1A). Data files are in folders organized by database release version number and species. Within each species-specific directory, data may be downloaded in FASTA, text or GFF formats.

The sequence retrieval tool, accessible from the tools section (Figure 1C) allows users to specify exact coordinates to be downloaded. Additionally, results of searches may be downloaded in bulk and in a defined manner. For example, one may choose to download 300 nt upstream of the translation start site of a set of genes in a search strategy to allow dumping of promoter sequences for the genes that resulted from a query.

FUTURE DIRECTIONS

FungiDB is expected to dramatically expand over the next few years to include >100 genome sequences and annotation. This will fill more phylogenetic diversity and cover additional plant and animal pathogenic fungi to allow for comparisons of common genomic traits among related and independent origins of pathogenicity. Importantly, the functional data portion is expected to accumulate multiple data sets from high-throughput proteomics (32), transcriptomics and metabolomics (33) studies to enable data mining and querying within a species data that can connect genes to function. The housing of many sets of functional genomics information will allow data to be applied to those organisms with less experimental data through comparative analyses thus assisting informing potential gene function in less tractable study organisms.

The work underway to further develop FungiDB is focused on enabling a ‘franchise model’ of the EuPathDB system where the site runs entirely independently from the present core EuPathDB system. This will greatly simplify the efforts needed to deploy instances of EuPathDB for groups of species as independent instances of the database and website since the tools are general purpose and the WDK search interfaces useful to a variety of groups of organisms. This will provide more generic installation of the software that runs EuPathDB to support clade-specific databases and websites such as FungiDB. Future work will also provide standalone installation documentation to instruct on best practices to deploy, configure and load data into the system.

ACKNOWLEDGEMENTS

The decision to launch this project emerged from a BWF community resource to provide better access to the fungal genome resources available.

FUNDING

The Burroughs Wellcome Fund (BWF) to build a Pan-Fungal genome database and the Alfred P. Sloan Foundation (to J.E.S.) to support creation of a data analysis core (MoBeDAC) for microbial community analyses of the Built Indoor Environment to support the curation and deployment of data in FungiDB for metagenomic applications. Funding for open access charge: Burroughs Wellcome Fund.

Conflict of interest statement. None declared.

REFERENCES