Journal of
Fungi

MDPI

Tutorial

FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes

Evelina Y. Basenko [1,*,†], Jane A. Pulman [1,†], Achchuthan Shanmugasundram [1,†], Omar S. Harb [2], Kathryn Crouch [3], David Starns [1], Susanne Warrenfeltz [4], Cristina Aurrecoechea [4], Christian J. Stoeckert Jr. [5], Jessica C. Kissinger [4], David S. Roos [2] and Christiane Hertz-Fowler [1,*]

[1] Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK; Jane.Pulman@liverpool.ac.uk (J.A.P.); achchu@liverpool.ac.uk (A.S.); D.E.Starns@liverpool.ac.uk (D.S.)

[2] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA; oharb@sas.upenn.edu (O.S.H.); droos@sas.upenn.edu (D.S.R.)

[3] Wellcome Trust Centre for Molecular Parasitology, Glasgow G12 8TA, UK; kathryn.crouch@glasgow.ac.uk

[4] Center for Tropical and Emerging Global Diseases, Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA; swfeltz@uga.edu (S.W.); aurreco@uga.edu (C.A.); jkissing@uga.edu (J.C.K.)

[5] Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA; stoeckrt@pennmedicine.upenn.edu

* Correspondence: ebasenko@liverpool.ac.uk (E.Y.B.); C.Hertz-Fowler@liverpool.ac.uk (C.H.-F.)

† These authors contributed equally to this work.

Received: 31 January 2018; Accepted: 15 March 2018; Published: 20 March 2018

check for updates

Abstract: FungiDB (fungidb.org) is a free online resource for data mining and functional genomics analysis for fungal and oomycete species. FungiDB is part of the Eukaryotic Pathogen Genomics Database Resource (EuPathDB, eupathdb.org) platform that integrates genomic, transcriptomic, proteomic, and phenotypic datasets, and other types of data for pathogenic and nonpathogenic, free-living and parasitic organisms. FungiDB is one of the largest EuPathDB databases containing nearly 100 genomes obtained from GenBank, *Aspergillus* Genome Database (AspGD), The Broad Institute, Joint Genome Institute (JGI), Ensembl, and other sources. FungiDB offers a user-friendly web interface with embedded bioinformatics tools that support custom in silico experiments that leverage FungiDB-integrated data. In addition, a Galaxy-based workspace enables users to generate custom pipelines for large-scale data analysis (e.g., RNA-Seq, variant calling, etc.). This review provides an introduction to the FungiDB resources and focuses on available features, tools, and queries and how they can be used to mine data across a diverse range of integrated FungiDB datasets and records.

Keywords: fungi; pathogen; bioinformatics; omics; genomics; transcriptomics; proteomics; sequence analysis; RNA-Seq; Galaxy

1. Introduction

The Eukaryotic Pathogen Bioinformatics Resource Center (eupathdb.org) comprises a family of bioinformatics resources that include an integrated functional genomics database for fungi and oomycetes—FungiDB [1,2]. FungiDB (fungidb.org) is a free online resource for data mining and functional genomics analysis. It provides an easy-to-use, interactive interface to explore genomes, annotation, functional data (e.g., transcriptomics, proteomics, phenomic data for selected species), metabolic pathways, and results from numerous genome-wide analyses (i.e., InterPro scan, signal peptide and transmembrane domain predictions, orthology, etc.). FungiDB contains an expanding number of genomes from various taxa including, but not limited to, plant, animal, and human pathogens. FungiDB contains nearly 100 genomes obtained from GenBank [3], Ensembl [4], Joint Genome Institute (JGI) [5], *Aspergillus* Genome Database (AspGD) [6], *Candida* Genome Database (CGD) [7], *Saccharomyces* Genome Database (SGD) [8], PomBase [9], and The Broad Institute [10]. FungiDB continues to acquire new genomes and datasets from data repositories, other fungal databases, and individual providers. The loaded genomes and related datasets can be investigated via a number of search strategies and FungiDB website-embedded tools. Presented here is an overview of FungiDB resources and several tutorials on how to mine FungiDB bioinformatic resources by creating in silico experiments.

2. Overview of the FungiDB Home Page and its Features

The home page of FungiDB can be divided into four sections: the header (Figure 1a), the grey menu bar (Figure 1b), the side bar on the left (Figure 1c), and the main central section with three large panels: *Search for Genes* (left), *Search for Other Data Types* (middle), and *Tools* (right) (Figure 1d).

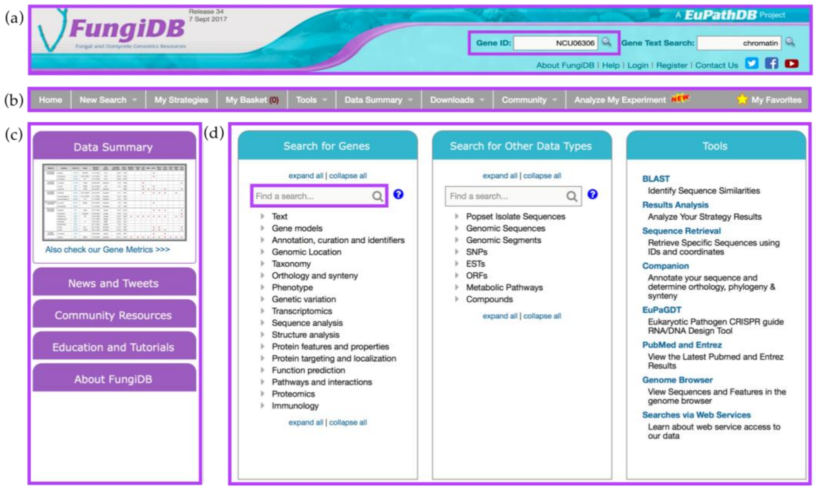

(a)
FungiDB
Release 34
7 Sept 2017
Fungal and Oomycete Genomics Resources
A EuPathDB Project
Gene ID: NCU06306
Gene Text Search: chromatin
About FungiDB | Help | Login | Register | Contact Us
(b)
Home
New Search
My Strategies
My Basket (0)
Tools
Data Summary
Downloads
Community
Analyze My Experiment NEW
My Favorites
(c)
Data Summary
Also check our Gene Metrics >>>
News and Tweets
Community Resources
Education and Tutorials
About FungiDB
(d)
Search for Genes
expand all | collapse all
Find a search...
Text
Gene models
Annotation, curation and identifiers
Genomic Location
Taxonomy
Orthology and synteny
Phenotype
Genetic variation
Transcriptomics
Sequence analysis
Structure analysis
Protein features and properties
Protein targeting and localization
Function prediction
Pathways and interactions
Proteomics
Immunology
expand all | collapse all
Search for Other Data Types
expand all | collapse all
Find a search...
Popset Isolate Sequences
Genomic Sequences
Genomic Segments
SNPs
ESTs
ORFs
Metabolic Pathways
Compounds
expand all | collapse all
Tools
BLAST
Identify Sequence Similarities
Results Analysis
Analyze Your Strategy Results
Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates
Companion
Annotate your sequence and determine orthology, phylogeny & synteny
EuPaGDT
Eukaryotic Pathogen CRISPR guide RNA/DNA Design Tool
PubMed and Entrez
View the Latest Pubmed and Entrez Results
Genome Browser
View Sequences and Features in the genome browser
Searches via Web Services
Learn about web service access to our data

Figure 1. FungiDB Home page and its features. The home page comprises (**a**) the header section with *Gene ID* search highlighted in purple; (**b**) the main menu in grey; (**c**) the side bar with links and various information sections; and (**d**) the main component offering *Search for Genes*, *Search for Other Data Types*, and *Tools* sections. *Find a search* box from *Search for Genes* component is highlighted in purple.

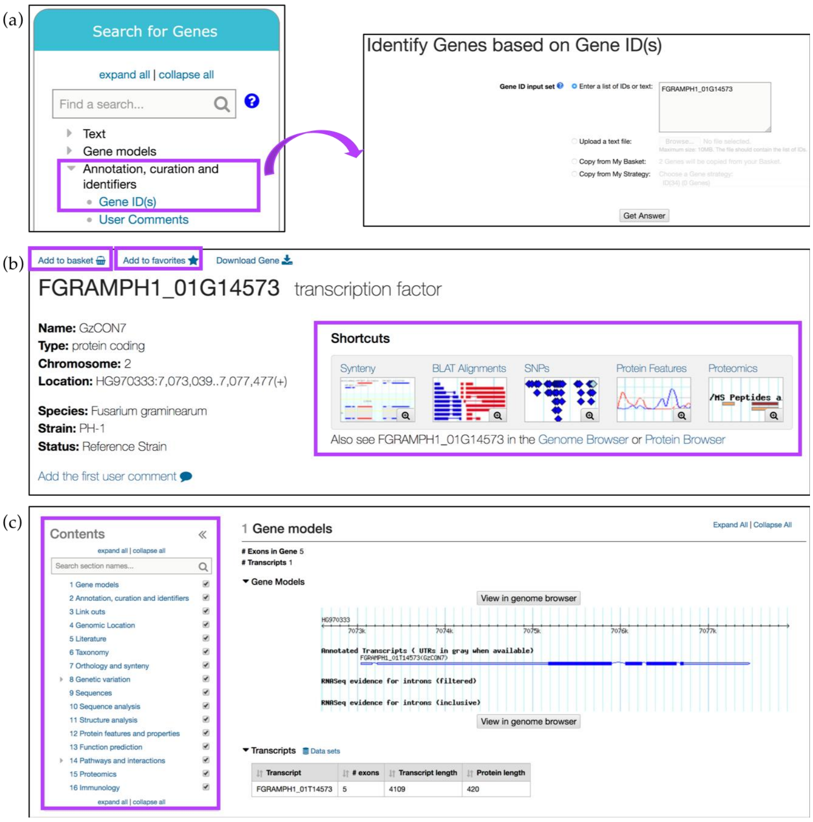
The header section offers quick access to the *Gene ID* and *Gene Text search* and a list of useful links (Figure 1a). To quickly navigate to a specific gene record page, a user can enter a single Gene ID (e.g., NCU06306) into the *Gene ID* search box at the top of the main page (Figure 1a). Alternatively, searching for *chromatin* in the *Gene Text Search* box adjacent to the *Gene ID* search box will return a list of genes from fungal and oomycete organisms that have matches to the searched term in any of the following gene fields: enzyme commission (EC) description, Gene ID, Gene notes, gene product, gene name, gene ontology (GO) terms and definitions, metabolic pathway names and descriptions, phenotype, protein domain names and descriptions, PubMed, and user comments. The links located directly under the search boxes provide quick access to information about the FungiDB project, login and registration, social media links, YouTube tutorial channel, and the *Contact Us* email form. Support inquiries can also be sent to *help@fungidb.org* (Figure 1a).

The grey menu bar is preserved across almost all FungiDB pages and provides easy navigation to all database sections when a user navigates away from the main page (Figure 1b). The side bar section (Figure 1c) consists of the *Data Summary*, *News and Tweets*, *Community Resources*, *Education and Tutorials*, and *About FungiDB* sections.

The central section is the largest section of the home page and consists of three panels: *Search for Genes*, *Search for Other Data Types*, and *Tools* panels (Figure 1d). The *Search for Genes* panel contains links to searches organized into seventeen intuitive categories allowing the user to quickly navigate to a desired search. Each search initiated from this panel will return lists of genes as search results. For example, a search for genes based on a specific organism is found under the *Taxonomy* category. Secondly, the *Search for Other Data Types* panel offers searches that return non-gene entities (e.g., *Single Nucleotide Polymorphisms (SNPs)*, *Metabolic Pathways*, *Compounds*, *Genomic Sequences based on genomic location*). Non-gene entities that have genomic locations can be co-located with each other or with genes (see below). Both the *Search for Genes* and *Search for Other Data Types* panels include a find option that allows users to quickly find a specific search by typing a key word in the *Find a search* box (highlighted in purple). Lastly, the third panel on the right provides access to useful bioinformatics tools that can be used independently of the searches (e.g., BLAST, Sequence Retrieval Tool, or Genome Browser).

3. Exploring FungiDB Records: Gene ID Searches and Gene Record Pages

A *Gene ID* search can also be invoked from the *Search for Genes* panel via the *Annotation, curation, and identifiers* category (Figure 2a). For example, searching for FGRAMPH1_01G14573, a *Fusarium graminearum* (*F. graminearum*) transcription factor important for virulence [11], returns the gene record page for this gene. Gene record pages contain aggregate information about a gene and its function that is generated from integrated datasets and automated pipelines.



Identify Genes based on Gene ID(s)

Gene ID input set ● Enter a list of IDs or text: FGRAMPH1_01G14573
○ Upload a text file: Browse... No file selected. Maximum size: 10MB. The file should contain the list of IDs.
○ Copy from My Basket: 2 Genes will be copied from your Basket.
○ Copy from My Strategy: Choose a Gene strategy: ID(34) (0 Genes)

Get Answer

FGRAMPH1_01G14573 transcription factor

Name: GzCON7
Type: protein coding
Chromosome: 2
Location: HG970333:7,073,039..7,077,477(+)

Species: Fusarium graminearum
Strain: PH-1
Status: Reference Strain

Also see FGRAMPH1_01G14573 in the Genome Browser or Protein Browser

Add the first user comment

1 Gene models

# Exons in Gene 5
# Transcripts 1

Gene Models

Transcripts

Transcript	# exons	Transcript length	Protein length
FGRAMPH1_01T14573	5	4109	420

Figure 2. Simple searches and gene record pages. (**a**) A Gene ID search can be initiated from the main component section, *Search for Genes* option, *Annotation, curation and identifiers* submenu (highlighted in purple); (**b**) Gene record page can be saved or bookmarked via *Add to basket* or *Add to favorites* links, respectively (highlighted in purple). Thumbnail *Shortcuts* (highlighted) provides easy navigation to key datasets; (**c**) Additional gene record sections can be quickly identified via the searchable *Contents* menu on the left (highlighted in purple).

Gene record pages can be saved (via the *Add to basket* link) or bookmarked (by clicking on the *Add to favorites* link) (Figure 2b). The *Add to basket* function saves the gene record to a basket associated with a user's account. Basket items can be found in the *My Strategies* section when a user is logged in. The basket serves as a shopping cart where genes in the basket can be downloaded or transferred to a search strategy. Adding a gene to favorites creates a bookmark to that gene in the *My Favorites* section available within the grey menu bar (Figure 1b). In the *My Favorites* section, users can also add private notes and project descriptions about saved items.

FungiDB gene record pages can be navigated via the thumbnail *Shortcuts* menu at the top (Figure 2b) and *Contents* menu on the left (Figure 2c). For example, when transcriptomics data for a gene of interest is integrated into FungiDB, a corresponding thumbnail appears on the gene page. Each thumbnail shows the number of datasets available within a given category. Clicking on the magnifying glass symbol within the thumbnail will open a preview screen of the evidence, while clicking anywhere within the thumbnail will navigate the user to the corresponding section on the gene page. Sections can be quickly identified by using the *Search section names* search box at the top of the *Contents* menu and clicking on the desired section. In addition, sections in a gene page can be hidden by unchecking the box to the right of any section name in the *Contents* menu.

The *Gene models* section is the first section of the gene record page and it contains information about the structure of the gene such as exon count, transcript number and a visual GBrowse representation of gene location, annotated UTRs and introns, and RNA-Seq evidence (when available). This section can be viewed in more detail, including the supporting data, by clicking on the *View in genome browser* button.

The *Annotation, curation, and identifiers* section offers alternate product descriptions, previous identifiers and aliases, notes from annotator, and user comments associated with a gene record. These sections are populated using data from either other fungal specialized resources, internal curation, or user-submitted data (user comments, see below).

The *Link outs* section offers redirection to other resources including EnsemblFungi [4], Entrez Gene [12], FGSC [13], UniProt [14], JGI [5], PHI-Base [15], AspGD [6], PhylomeDB [16], PomBase [9], CGD [7], SGD [8], and others.

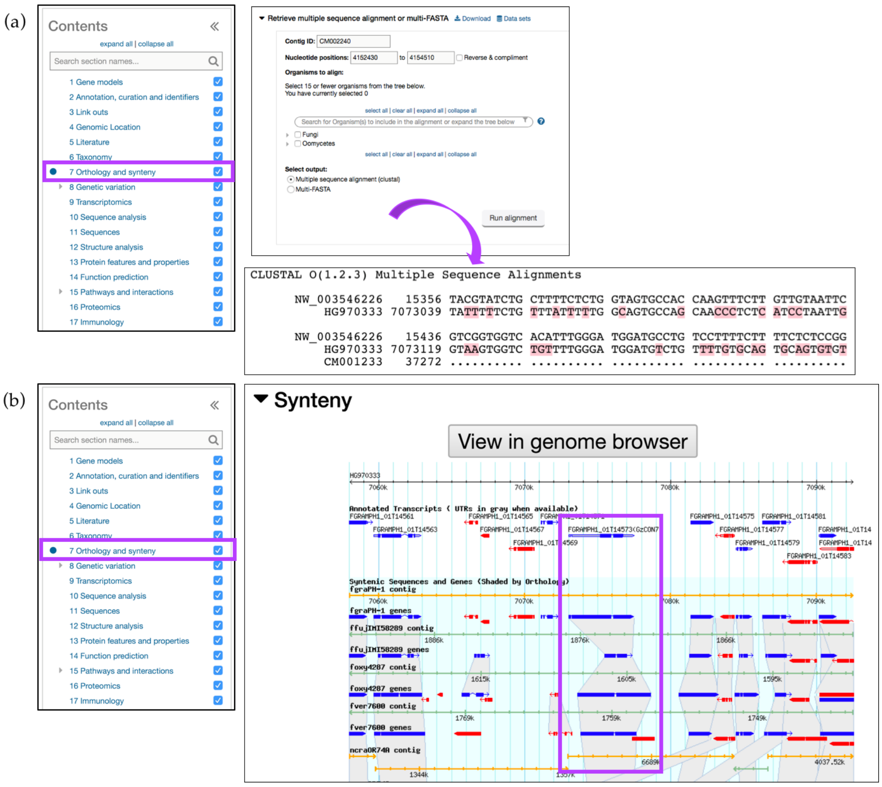
The *Genomic Location* section contains gene or sequence coordinates within a chromosome or contig/scaffold and a link out to GBrowse.

The *Literature* section provides access to associated publications gathered by FungiDB via user comments or direct uploads from other resources (The Broad Institute, AspGD, fgmutantDB [17], PHI-Base, The *Neurospora crassa* e-compendium [18]) and FungiDB in-house curation efforts.

The *Taxonomy* section provides detailed information about organism classification within the Eukaryote kingdom following the National Center for Biotechnology Information (NCBI) convention.

The *Orthology and synteny* section provides a table of Orthologs and Paralogs within FungiDB (Figure 3a). This table is searchable and each ortholog in the table is hyperlinked to its gene page in FungiDB. An additional feature to interrogate sequence similarities between species of the same or different genera is offered via the *Retrieve multiple sequence alignment or multi-FASTA* tool. A user can select up to 15 organisms to perform multiple sequence alignment for a particular gene. This query uses pairwise Mercator [19] analysis combined with a Clustal Omega output format. A sequence alignment of FGRAMPH1_01G14573 with *Magnaporthe* and *Sordaria* species, selected from the *Organisms to align* parameter option, generated an output of genomic sequence IDs as identifiers. NW_003546226 and HG970333 identifiers belong to *F. graminearum* PH-1 and *Sordaria macrospora* genomes, respectively (Figure 3a). The gene record page for FGRAMPH1_01G14573 also contains a GBrowse display of synteny between related organisms with syntenic orthologs shaded in grey. An interactive display is also accessible via the *View in genome browser* button and highlighted in Figure 3b.

The *Genetic Variation* section summarizes integrated FungiDB SNP data for a given region and classifies SNPs into groups based on the resulting effect on gene function: noncoding, nonsynonymous or synonymous, and nonsense nucleotide changes.



Contig ID: CM002240
Nucleotide positions: 4152430 to 4154510 ☐ Reverse & compliment
Organisms to align:
Select 15 or fewer organisms from the tree below.
You have currently selected 0
☐ Fungi
☐ Oomycetes
Select output:
● Multiple sequence alignment (clustal)
○ Multi-FASTA

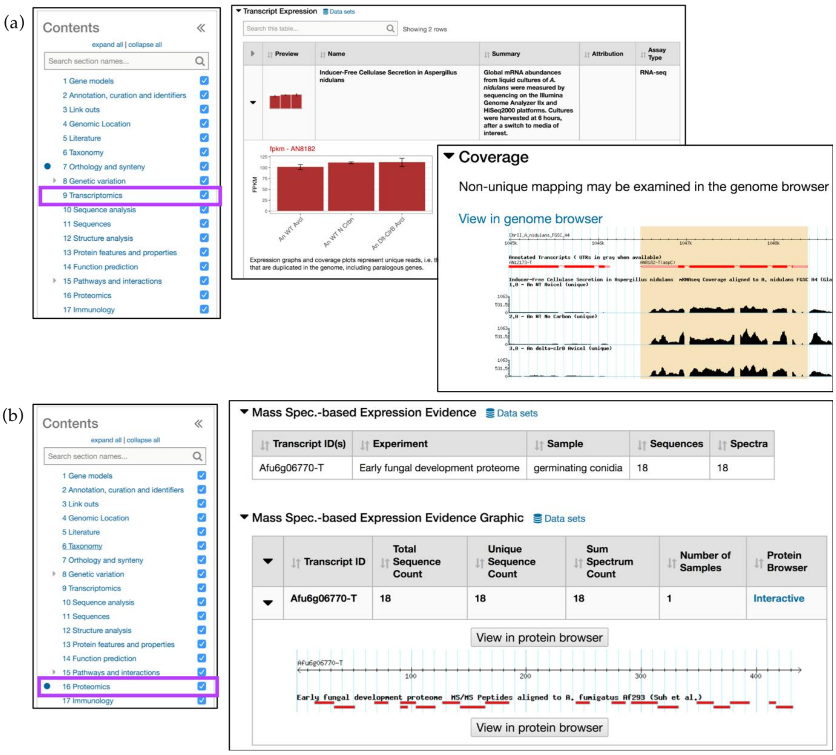
```
CLUSTAL O(1.2.3) Multiple Sequence Alignments

      NW_003546226    15356 TACGTATCTG CTTTTCTCTG GTAGTGCCAC CAAGTTTCTT GTTGTAATTC
          HG970333  7073039 TATTTTTCTG TTTATTTTTG GCAGTGCCAG CAACCCTCTC ATCCTAATTG

      NW_003546226    15436 GTCGGTGGTC ACATTTGGGA TGGATGCCTG TCCTTTTCTT TTCTCTCCGG
          HG970333  7073119 GTAAGTGGTC TGTTTTGGGA TGGATGTCTG TTTTGTGCAG TGCAGTGTGT
          CM001233    37272 .......... .......... .......... .......... ..........
```

Figure 3. Simple searches and gene record pages, continued. *Orthology and Synteny* section (highlighted in purple) for the *F. graminearum* gene FGRAMPH1_01G14573. (http://fungidb.org/gene/FGRAMPH1_01G14573) (**a**) A Multiple Sequence Alignment (MSA) run within the gene record page against *Magnaporthe* and *Sordaria* sequences. Shown as CLUSTAL Omega output; (**b**) Syntenic orthologs can be previewed from the GBrowse window within the gene page. A separate session in GBrowse page can be deployed by clicking on the *View in genome browser* button. Syntenic regions for the gene are highlighted in purple.

Genetic variation tracks can be explored in GBrowse and SNPs visualized by clicking on the *View in genome browser* button and activating appropriate tracks from the *Select Tracks* tab in GBrowse.

The *Transcriptomics* section (RNA-Seq and microarray data) provides expandable rows with tabular data, summaries, coverage, expression graphs, and more. For example, the *Aspergillus nidulans* AN8182 gene record page demonstrates expression profiles for AN8182 where mRNA samples were collected from wild-type and *clr-2*/*clrB* mutants after a culture shift from sucrose/glucose to Avicel (crystalline cellulose) or no carbon media [20] (Figure 4a). The *Expression graph* (fpkm—AN8182) provides an overview of peaks detected in three treatments and *Coverage Section* shows uniquely mapped reads. When reads map to several genome locations and therefore could have been derived from multiple transcripts they are labeled as nonunique. Both unique and nonunique reads are now displayed in fpkm graphs and are available in GBrowse.

Next, the *Sequences*, *Sequence analysis*, *Structure*, and *Protein features and properties* menus offer DNA, RNA, and protein sequence information followed by records about known protein domains (InterPro), signal and Transmembrane predictions, BLASTP hits, BLAT hits against the nonredundant protein sequence database (GenBank), and other information.



	Preview	Name	Summary	Attribution	Assay Type
▼		Inducer-Free Cellulase Secretion in Aspergillus nidulans	Global mRNA abundances from liquid cultures of *A. nidulans* were measured by sequencing on the Illumina Genome Analyzer IIx and HiSeq2000 platforms. Cultures were harvested at 6 hours, after a switch to media of interest.		RNA-seq

Expression graphs and coverage plots represent unique reads, i.e. t
that are duplicated in the genome, including paralogous genes.

Non-unique mapping may be examined in the genome browser

View in genome browser
ChrII_A_nidulans_FGSC_A4
1045k
1046k
1047k
1048k
Annotated Transcripts (UTRs in gray when available)
AN12173-T
AN8182-T(aspC)
Inducer-Free Cellulase Secretion in Aspergillus nidulans mRNAseq Coverage aligned to A. nidulans FGSC A4 (Gla
1.0 - An WT Avicel (unique)
2.0 - An WT No Carbon (unique)
3.0 - An delta-clrB Avicel (unique)
1063
531.5
0
(b)
Contents
expand all | collapse all
Search section names...
1 Gene models ☑
2 Annotation, curation and identifiers ☑
3 Link outs ☑
4 Genomic Location ☑
5 Literature ☑
6 Taxonomy ☑
7 Orthology and synteny ☑
8 Genetic variation ☑
9 Transcriptomics ☑
10 Sequence analysis ☑
11 Sequences ☑
12 Structure analysis ☑
13 Protein features and properties ☑
14 Function prediction ☑
15 Pathways and interactions ☑
● 16 Proteomics ☑
17 Immunology ☑
Mass Spec.-based Expression Evidence
Data sets

Transcript ID(s)	Experiment	Sample	Sequences	Spectra
Afu6g06770-T	Early fungal development proteome	germinating conidia	18	18

▼	Transcript ID	Total Sequence Count	Unique Sequence Count	Sum Spectrum Count	Number of Samples	Protein Browser
▼	Afu6g06770-T	18	18	18	1	Interactive

View in protein browser
Afu6g06770-T
100
200
300
400
Early fungal development proteome MS/MS Peptides aligned to A. fumigatus Af293 (Suh et al.)
View in protein browser

Figure 4. Simple searches and gene record pages, continued. (**a**) *Transcriptomics* menu (highlighted in purple) contains *Transcript Expression* records that show fpkm graphs for *A. nidulans* gene AN8182 (gene record page: http://fungidb.org/fungidb/app/record/gene/AN8182). Coverage plots (an inset to the *Transcript Expression* section) displays unique read alignments and a *View in genome browser* link out; (**b**) *Proteomics* menu (highlighted in purple) displays Mass Spec. Evidence for *A. fumigatus* gene Afu6g06770 (gene record page: http://fungidb.org/fungidb/app/record/gene/Afu6g06770). Proteomics tracks can be also visualized by clicking on the *View in protein browser* button.

The *Function Prediction* section houses Gene Ontology assignments (downloaded from Gene Ontology databases and manually curated by FungiDB for a selected group of organisms). FungiDB gene pages also display GO Slim terms from the generic subset for each GO term associated with that gene. The GO assignments are followed by other functional data and datasets, including the *Pathways and Interactions* (*Metabolic Pathways* and *Metabolic Pathway Reactions* sections), *Proteomics evidence* (Mass Spectrometry-based expression and post-translational modification data in table and graph formats).

Pathways and interactions provide descriptions of metabolic pathways that are loaded into FungiDB from the KEGG [21] and MetaCyc [22] repositories. Genes are linked to individual pathways by EC numbers and clicking on any of the metabolic pathway links will navigate to an interactive metabolic pathway viewer where the user can explore individual reactions or export all known data for a given pathway.

The *Proteomics* section contains Mass Spec. evidence data and phosphoproteomics datasets. In Figure 4b, the gene records page for the *Aspergillus fumigatus (A. fumigatus)* gene Afu6g06770 demonstrates that the protein is found in germinating conidia. A summary table with sequence counts and a link out to peptide alignments against the reference genome can be viewed in GBrowse.

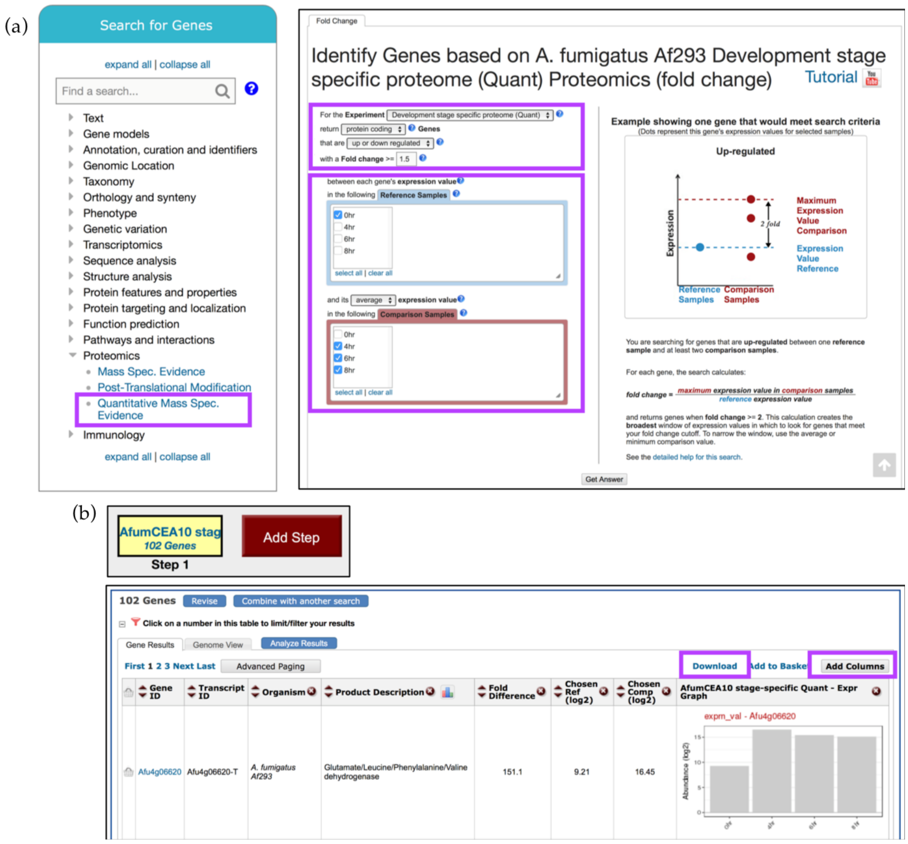
4. Creating in Silico Experiments via the FungiDB Query Interface

4.1. Mining Proteomic Datasets

FungiDB integrates proteomics data that are presented in several ways: peptides are mapped to genes, gene expression levels are loaded based on quantitative Mass Spec. data, and post-translationally modified amino acids are visualized as GBrowse tracks.

Genes may be identified based on proteomics evidence by selecting one of the searches under the *Proteomics* category in the *Search for Genes* section on the home page. In the example described here, quantitative proteomics data from Suh *et al.* [23] are interrogated to identify *A. fumigatus* genes up-regulated in the growing conidia and then cross referenced with *Gene Models* data for *Exon Count* to identify genes with no more than 5 exons.

The *Quantitative Mass Spec. Evidence* search can be accessed from the *Proteomics* section under the *Search for Genes* category (Figure 5a). Then, in the proteomics dataset selection window, click on the *Fold Change* (FC) button next to the Suh *et al.* study. To identify genes that are up-regulated by at least twofold in the 4–8 h time frame compared to the 0 h time point, set the regulation direction to *up-regulated*, modify the Fold Change parameter to 2, and select 0 h from the *Reference Samples* and 4, 6, and 8 h from the *Comparison Samples*, and click *Get Answer* (Figure 5a).



Identify Genes based on A. fumigatus Af293 Development stage specific proteome (Quant) Proteomics (fold change)

For the Experiment: Development stage specific proteome (Quant)
return: protein coding Genes
that are: up or down regulated
with a Fold change >= 1.5

between each gene's expression value in the following Reference Samples
☑ 0hr
☐ 4hr
☐ 6hr
☐ 8hr

and its average expression value in the following Comparison Samples
☐ 0hr
☑ 4hr
☑ 6hr
☑ 8hr

select all | clear all
Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)
Up-regulated
Expression
2 fold
Maximum Expression Value Comparison
Expression Value Reference
Reference Samples
Comparison Samples

You are searching for genes that are **up-regulated** between one **reference sample** and at least two **comparison samples**.

For each gene, the search calculates:

fold change = *maximum expression value in comparison samples* / *reference expression value*

and returns genes when **fold change** >= **2**. This calculation creates the **broadest** window of expression values in which to look for genes that meet your fold change cutoff. To narrow the window, use the average or minimum comparison value.

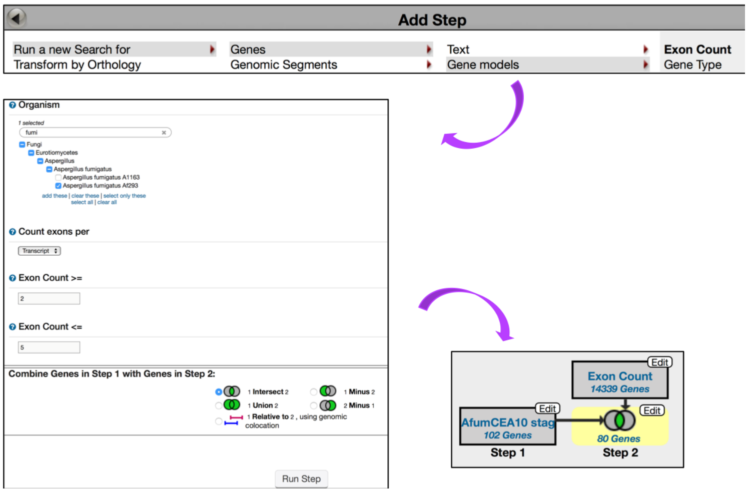
See the detailed help for this search.

Gene ID	Transcript ID	Organism	Product Description	Fold Difference	Chosen Ref (log2)	Chosen Comp (log2)	AfumCEA10 stage-specific Quant - Expr Graph
Afu4g06620	Afu4g06620-T	*A. fumigatus Af293*	Glutamate/Leucine/Phenylalanine/Valine dehydrogenase	151.1	9.21	16.45	

Figure 5. Quantitative Mass Spec. search examining up-regulated genes in *A. fumigatus* Af293 during conidial growth. (**a**) Select the *Quantitative Mass Spec. Evidence* menu (highlighted in purple) to access Suh *et al.* dataset and select *Reference* and *Comparison Samples* as shown to identify genes that are up-regulated throughout conidial development. The selection parameter areas are highlighted in purple; (**b**) Overview of the search strategy. Explore results, *Add Columns*, or *Download* results (highlighted in purple). Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=96851d3ef002899a.

The search strategy returned 102 genes that are up-regulated during conidial growth (Figure 5b). Examine or revise results, explore additional data by adding more columns via the *Add Columns* button, or export records via the *Download* link (Figure 5b and highlighted).

To further expand this search strategy and determine how many of these genes have multiple exons, click on the *Add Step* button and choose to *Run a New Search for, Genes, Gene models, Exon Count* (min >= 2, max <= 5) (Figure 6). Intersecting the second search with the first one returned 80 genes in *A. fumigatus*.



Organism
1 selected
fumi
Fungi
Eurotiomycetes
Aspergillus
Aspergillus fumigatus
☐ Aspergillus fumigatus A1163
☑ Aspergillus fumigatus Af293
add these | clear these | select only these
select all | clear all

Count exons per
Transcript

Exon Count >=
2

Exon Count <=
5

Combine Genes in Step 1 with Genes in Step 2:
● 1 Intersect 2 ○ 1 Minus 2
○ 1 Union 2 ○ 2 Minus 1
○ 1 Relative to 2 , using genomic colocation

Figure 6. Quantitative Mass Spec. query examining up-regulated genes in *A. fumigatus* Af293 during conidial growth with at least 2 and no more than 5 exons. *Exon Count* search is deployed from the *Search for Genes* category, *Gene models* menu, *Exon Count* subcategory. Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=96851d3ef002899a.

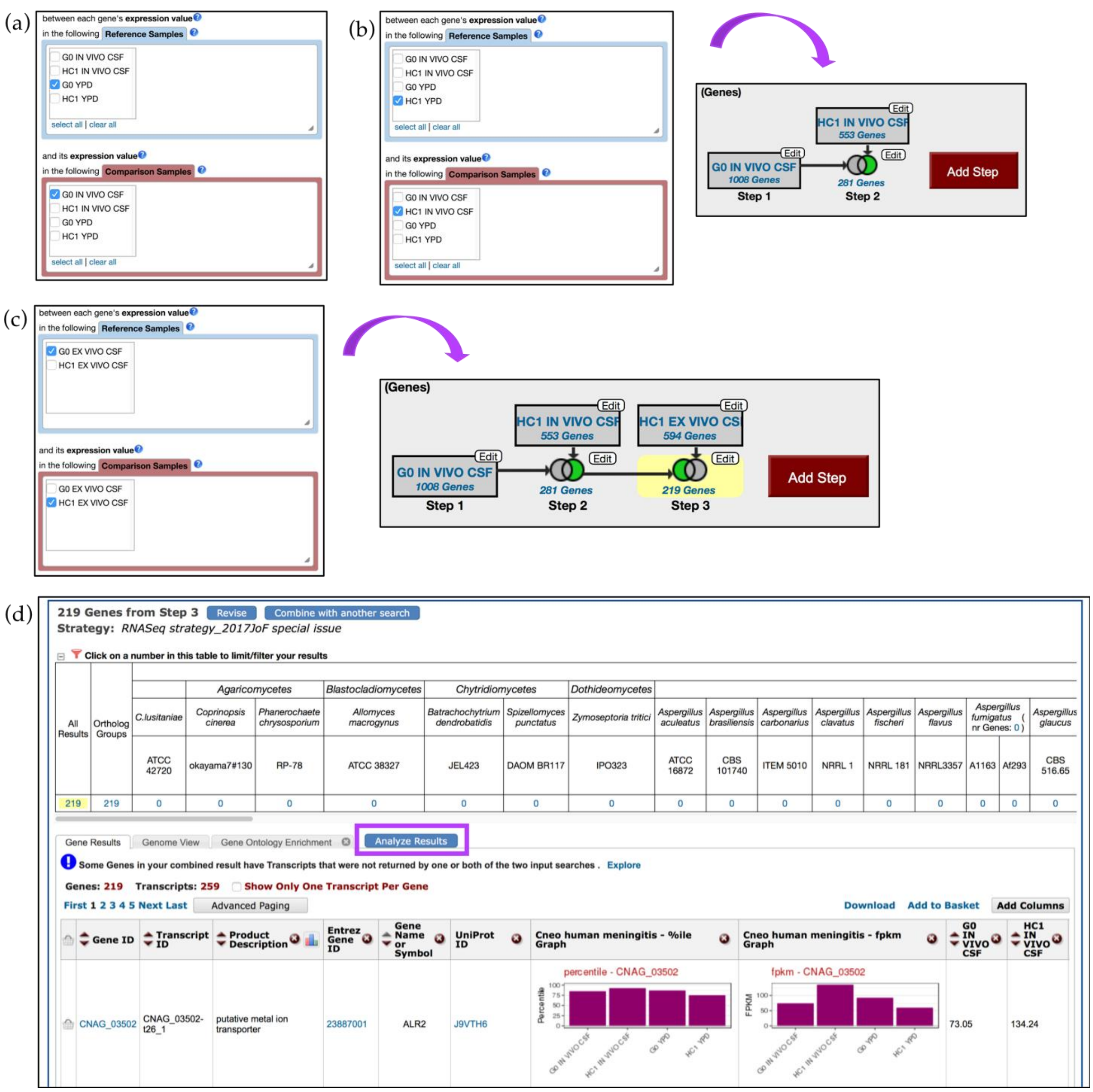
4.2. Creating In Silico RNA-Seq Experiments

There are several ways to identify genes based on RNA-Seq Evidence. Depending on the type of integrated dataset, users can choose to analyze a dataset using Differential Expression (DE), Fold Change (FC), or Percentile (P). In this example, we will query a dataset published by Chen *et al.* [24] and compare transcriptomes of two strains of *Cryptococcus neoformans*, a fungal organism causing fungal meningitis and high mortality rates in immunocompromised patients [25,26]. The HC1 and G0 strains were isolated from the cerebrospinal fluid (*in vivo* CSF) of patients in the U.S. and Uganda, respectively, and then also incubated in pooled human CSF (*ex vivo* CSF) and in yeast extract–peptone–dextrose (YPD).

The strategy described below compares transcriptomes of two strains and aims to identify genes that are up-regulated in the HC1 strain compared with in the G0 strain (in vivo CSF) (Step 1–2) and then identifies those genes that are specific for in vivo expression in the HC1 strain (Step 3) (Figure 7).

To initiate an *in silico* experiment, click on the *Transcriptomics* category on the home page or in the drop-down menu from the grey main menu and then navigate to the *RNA Seq Evidence* page. From the list of available studies, click on the fold change (FC) button in blue for the *Cryptococcus neoformans transcriptome at the site of human meningitis (single-read data)* dataset. Next, select to identify genes that are up-regulated by twofold between each gene's expression value in the *G0 YPD* sample (*Reference Sample*) and its expression in the *G0* in vivo *CSF* sample (*Comparison Sample*) (Figure 7a).

Next, click on the *Add Step* button, *Run a new Search for, Genes*, *Transcriptomics*, *RNA Seq Evidence* and repeat previous steps but this time identify genes that are twofold up-regulated in *HC1* in vivo *CSF* (*Comparison sample*) versus in *HC1 YPD* (*Reference sample*). To set conditions for the next search, select the *2 minus 1* intersect option from the four available set operations (Figure 7b). As a result, Step 2 returned 281 genes that are up-regulated in *HC1* in vivo *CSF* but not in *G0* in vivo *CSF*.



(a)

between each gene's expression value
in the following Reference Samples
☐ G0 IN VIVO CSF
☐ HC1 IN VIVO CSF
☑ G0 YPD
☐ HC1 YPD
select all | clear all

and its expression value
in the following Comparison Samples
☑ G0 IN VIVO CSF
☐ HC1 IN VIVO CSF
☐ G0 YPD
☐ HC1 YPD
select all | clear all

(b)

between each gene's expression value
in the following Reference Samples
☐ G0 IN VIVO CSF
☐ HC1 IN VIVO CSF
☐ G0 YPD
☑ HC1 YPD
select all | clear all

and its expression value
in the following Comparison Samples
☐ G0 IN VIVO CSF
☑ HC1 IN VIVO CSF
☐ G0 YPD
☐ HC1 YPD
select all | clear all

(c)

between each gene's expression value
in the following Reference Samples
☑ G0 EX VIVO CSF
☐ HC1 EX VIVO CSF

and its expression value
in the following Comparison Samples
☐ G0 EX VIVO CSF
☑ HC1 EX VIVO CSF

(d)

219 Genes from Step 3 Revise Combine with another search
Strategy: *RNASeq strategy_2017JoF special issue*

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups		*Agaricomycetes*		*Blastocladiomycetes*	*Chytridiomycetes*		*Dothideomycetes*										
		C.lusitaniae	*Coprinopsis cinerea*	*Phanerochaete chrysosporium*	*Allomyces macrogynus*	*Batrachochytrium dendrobatidis*	*Spizellomyces punctatus*	*Zymoseptoria tritici*	*Aspergillus aculeatus*	*Aspergillus brasiliensis*	*Aspergillus carbonarius*	*Aspergillus clavatus*	*Aspergillus fischeri*	*Aspergillus flavus*	*Aspergillus fumigatus* (nr Genes: 0)		*Aspergillus glaucus*
		ATCC 42720	okayama7#130	RP-78	ATCC 38327	JEL423	DAOM BR117	IPO323	ATCC 16872	CBS 101740	ITEM 5010	NRRL 1	NRRL 181	NRRL3357	A1163	Af293	CBS 516.65
219	219	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene Results | Genome View | Gene Ontology Enrichment | Analyze Results

Some Genes in your combined result have Transcripts that were not returned by one or both of the two input searches . Explore

Genes: 219 **Transcripts: 259** ☐ **Show Only One Transcript Per Gene**

First 1 2 3 4 5 Next Last Advanced Paging **Download** **Add to Basket** **Add Columns**

Gene ID	Transcript ID	Product Description	Entrez Gene ID	Gene Name or Symbol	UniProt ID	Cneo human meningitis - %ile Graph	Cneo human meningitis - fpkm Graph	G0 IN VIVO CSF	HC1 IN VIVO CSF
CNAG_03502	CNAG_03502-t26_1	putative metal ion transporter	23887001	ALR2	J9VTH6			73.05	134.24

Figure 7. RNA-Seq query using *C. neoformans* single-read and paired-end datasets by Chen *et al.* [24] (**a**) RNA-Seq query on single-read data is initiated from the *Search for Genes* panel, *Transcriptomics* menu. Step 1 selects for two-fold up-regulated genes in *G0* in vivo *CSF* compared with in the *G0 YPD* reference sample; (**b**) Step 2 selects for *HC1* genes expressed in vivo *CSF* (using HC1 *YPD* sample as reference). Choose the *2 minus 1* intersect parameter to set up conditions for the final search; (**c**) Add Step and repeat steps as before but this time select the paired-end dataset. Step 3 identifies HC1 genes up-regulated in vivo *CSF* compared to ex vivo *CSF*. Choose the *1 minus 2* intersect parameter for the third step; (**d**) Results table for Step 3. Clicking on the *Analyze Results* tab (highlighted in purple) will redirect users to enrichment analysis using *Gene Ontology*, *Metabolic Pathway*, and *Word Enrichment* tools. Strategy in FungiDB http://fungidb.org/fungidb/im.do?s=892406814924ecdd.

To further delineate in-vivo-specific genes in HC1 strain, one can create an additional step to eliminate genes that were expressed in ex vivo conditions in the HC1 strain. Click on the *Add Step* button, navigate to the RNA-Seq datasets selection window and this time choose the paired-end dataset from Chen *et al.* instead of the single-read data. Configure the search to return genes that are twofold up-regulated in the *HC1* ex vivo *CSF* (*Comparison sample*) versus in the *G0* ex vivo *CSF* sample (*Reference sample*). Choose the *2 minus 3* option and click on the *Run Step* button (Figure 7c).

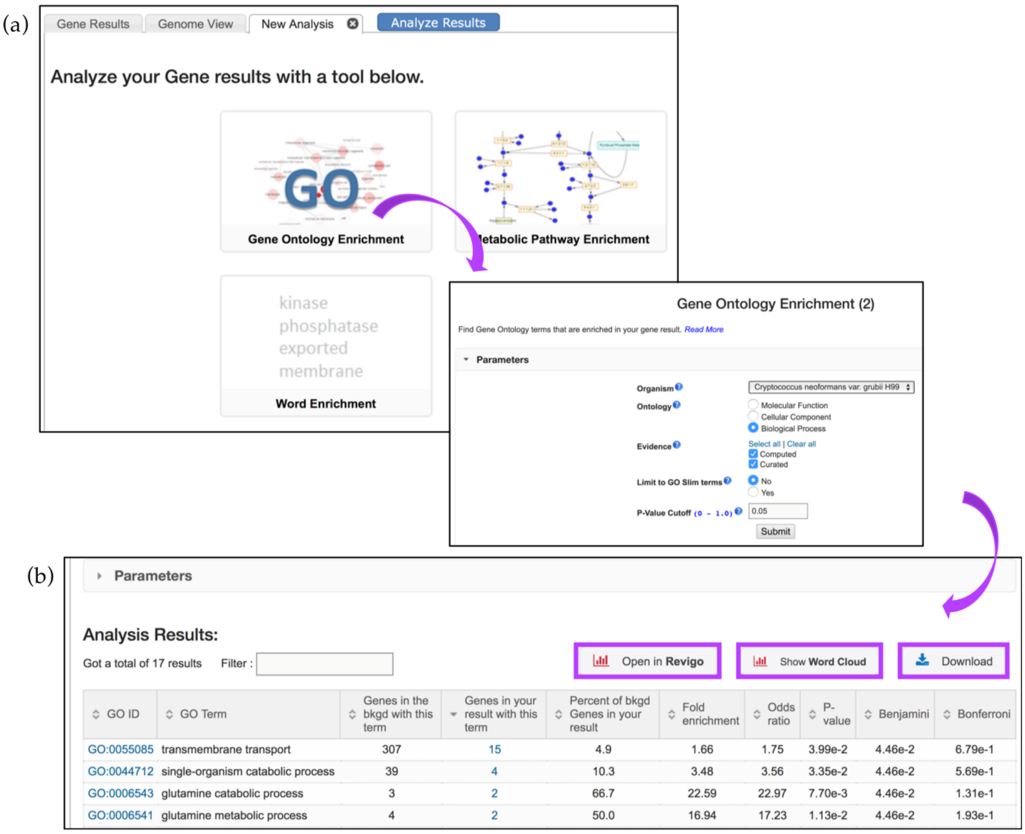
The strategy described above used two different types of data (single-read and paired-end RNA-Seq data) to examine differential gene expression. To further investigate results of gene-specific queries, users can take advantage of a number of enrichment tools located under the *Analyze Results* tab in blue (Figure 7d), which is reviewed further in the text.

5. Additional Tools for Enrichment and Data Analysis

When working with a list of genes such as RNA-Seq results (as described in Figure 6) or user-uploaded gene lists, one can perform several enrichment analyses to further characterize results into functional categories. Enrichment analysis can be accessed via the blue *Analyze Results* tab and it includes *Gene Ontology* [27,28], *Metabolic Pathway* [21,22], and *Word Enrichment* tools. The three types of analysis apply Fisher's Exact test to evaluate ontology terms, over-represented pathways, and product description terms. Enrichment is carried out using a Fisher's Exact test with the background defined as all genes from the organism being queried. The *p*-values corrected for multiple testing are provided using both the Benjamini–Hochberg false discovery rate method [29] and the Bonferroni method [30].

5.1. GO Term Enrichment and Visualization

To determine if a gene list contains enriched functions, a GO enrichment analysis may be performed. Continuing with the RNA-Seq example described in Figure 7, click on the *Analyze Results* tab, and then select the *Gene Ontology Enrichment* button (Figure 8a).



Analyze your Gene results with a tool below.

Gene Ontology Enrichment (2)

Find Gene Ontology terms that are enriched in your gene result. *Read More*

Parameters

Organism: Cryptococcus neoformans var. grubii H99

Ontology: ○ Molecular Function ○ Cellular Component ● Biological Process

Evidence: Select all | Clear all ☑ Computed ☑ Curated

Limit to GO Slim terms: ● No ○ Yes

P-Value Cutoff (0 - 1.0): 0.05

Submit

Analysis Results:

Got a total of 17 results Filter :

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd Genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
GO:0055085	transmembrane transport	307	15	4.9	1.66	1.75	3.99e-2	4.46e-2	6.79e-1
GO:0044712	single-organism catabolic process	39	4	10.3	3.48	3.56	3.35e-2	4.46e-2	5.69e-1
GO:0006543	glutamine catabolic process	3	2	66.7	22.59	22.97	7.70e-3	4.46e-2	1.31e-1
GO:0006541	glutamine metabolic process	4	2	50.0	16.94	17.23	1.13e-2	4.46e-2	1.93e-1

Figure 8. Gene enrichment analysis tools. (**a**) Clicking on the *GO—Gene Ontology Enrichment* button deploys GO enrichment that can be limited based on the following parameters: ontology aspects, GO evidence (*Computed* vs. *Curated*), and *p*-value (0.05 default), or GO Slim terms; (**b**) The enrichment results can be further visualized via Reduce + Visualize Gene Ontology (REViGO), Word Cloud image, or downloaded (highlighted in purple).

GO enrichment parameters allow users to limit their analysis on either *Curated* or *Computed* annotations, or both. Those with a GO evidence code inferred from electronic annotation (IEA) are denoted *Computed*, while all others have some degree of curation. Users can also choose to show results for the following functional aspects of the GO ontology: molecular function, cellular component, and biological processes, as well as set a custom *p*-value cut-off. When the GO slim option is chosen, both the genes of interest and the background are limited to GO terms that are part of the generic GO Slim subset (Figure 8a). Users may download a GO enrichment table (with the Gene IDs for each GO term added) as well as view and download a word cloud produced via the GO Summaries R package [31] (Figure 8b, highlighted in purple).

The Analysis Results table from the GO enrichment focusing on biological function shows enrichment for transmembrane transport genes. The table contains columns with GO IDs and GO terms along with the number of genes in the background and those specific to the RNA-Seq analysis results presented. Several additional statistical measurements are also included and are defined below:

- *Fold enrichment*—The ratio of the proportion of genes in the list of interest with a specific GO term over the proportion of genes in the background with that term;
- *Odds ratio*—Determines if the odds of the GO term appearing in the list of interest are the same as that for the background list;
- *p-value*—Assumptions under a null hypothesis, the probability of getting a result that is equal or greater than what was observed;
- *Benjamini–Hochburg* false discovery rate—A method for controlling false discovery rates for type 1 errors;
- *Bonferroni-adjusted p-values*—A method for correcting significance based on multiple comparisons;

GO enrichment analysis returned genes enriched in transmembrane transport (GO:0055085), single-organism catabolic process (GO:0044712), etc. (Figure 8b). These and other terms can be taken to an external REViGO website [32] by clicking on the *Open in Revigo* button. REViGO visualizes the GO enrichment results while removing redundant GO terms. The resulting similarity-based scatterplots are color-coded and provide interactive graphs and tag clouds as alternative visualizations. More information is available at the Revigo website: http://revigo.irb.hr.

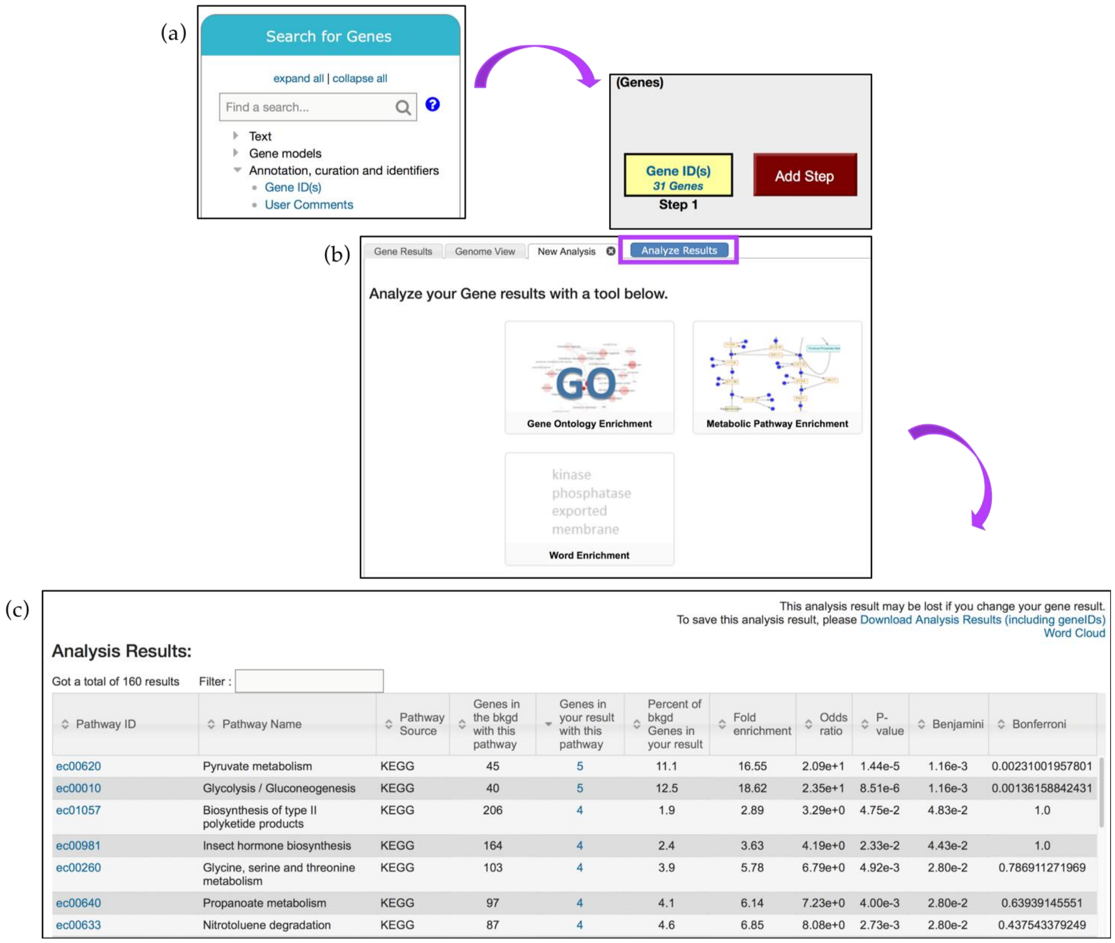
5.2. Metabolic Pathway Enrichment

The Metabolic Pathway Enrichment tool allows the user to determine whether particular metabolic pathways are over-represented in a subset of genes from an organism compared to the background. Described here is a search query using a list of genes from a publication studying pathogen adaptation responses in *Paracoccidioides brasiliensis* (*P. brasiliensis*), one of the most frequent causative agents of systemic mycoses in Latin America [33].

When fungal conidia of *P. brasiliensis* gain access to host tissues and organs they differentiate into a parasitic form that undergoes morphological changes conferring resistance to immune responses of the host. Alternative carbon metabolism is particularly important for the survival and virulence of this fungal pathogen within the host, Lima and colleagues reported that thirty-one genes of *P. brasiliensis* were up-regulated under carbon starvation conditions as determined by both RNA-Seq and proteomics studies [33].

Here, thirty-one genes described above are identified using the *Gene ID(s)* search, deployed via the *Search for Genes* menu and the *Annotation, curation and identifiers* submenu (Figure 9a). Next, deploy the *Metabolic Pathway Enrichment* tool via the *Analyze Results* tab, use *KEGG* pathway enrichment only, and then click *Submit* to start the analysis (Figure 9b). The analysis may take some time to run depending on the size of the gene list. Results will be presented in a new tab (Figure 9c).

KEGG pathway enrichment on the gene list described above highlighted enrichment in carbohydrate, amino acid, and lipid metabolism, corroborating the findings described in the manuscript [33]. These pathways can be explored further by clicking on the link in the *Pathway ID* column. The pyruvate metabolism pathway (ec00620) is one of the pathways enriched amongst *P. brasiliensis* genes that were up-regulated during starvation (Figure 10a).



This analysis result may be lost if you change your gene result.
To save this analysis result, please Download Analysis Results (including geneIDs)
Word Cloud

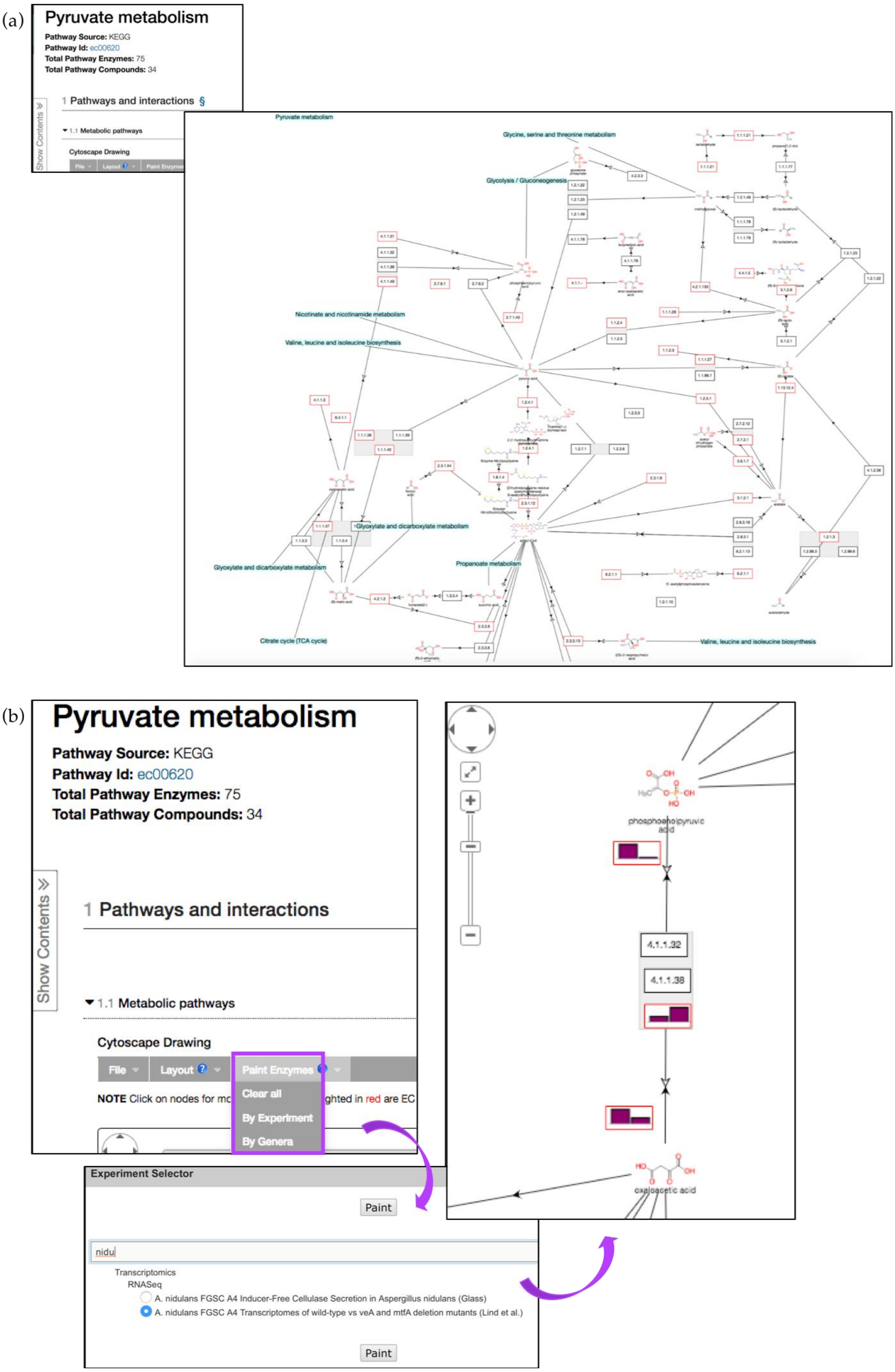
Analysis Results:

Got a total of 160 results Filter :

Pathway ID	Pathway Name	Pathway Source	Genes in the bkgd with this pathway	Genes in your result with this pathway	Percent of bkgd Genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini	Bonferroni
ec00620	Pyruvate metabolism	KEGG	45	5	11.1	16.55	2.09e+1	1.44e-5	1.16e-3	0.00231001957801
ec00010	Glycolysis / Gluconeogenesis	KEGG	40	5	12.5	18.62	2.35e+1	8.51e-6	1.16e-3	0.00136158842431
ec01057	Biosynthesis of type II polyketide products	KEGG	206	4	1.9	2.89	3.29e+0	4.75e-2	4.83e-2	1.0
ec00981	Insect hormone biosynthesis	KEGG	164	4	2.4	3.63	4.19e+0	2.33e-2	4.43e-2	1.0
ec00260	Glycine, serine and threonine metabolism	KEGG	103	4	3.9	5.78	6.79e+0	4.92e-3	2.80e-2	0.786911271969
ec00640	Propanoate metabolism	KEGG	97	4	4.1	6.14	7.23e+0	4.00e-3	2.80e-2	0.63939145551
ec00633	Nitrotoluene degradation	KEGG	87	4	4.6	6.85	8.08e+0	2.73e-3	2.80e-2	0.437543379249

Figure 9. Metabolic pathway enrichment analysis. (**a**) A list of 31 genes is analyzed via the *Gene IDs* tool (from the *Search for Genes* panel) and the (**b**) the *Metabolic Pathway Enrichment* tool from the *Analyze Results* tab (highlighted in purple) (**c**) Metabolic Pathway Enrichment analysis results page. Individual *Pathway ID* (e.g., ec00620) is linked to the interactive CytoScape interface. Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=6825e606be957ead.

Cytoscape interactive map offers zoom-in, zoom-out, and reposition capabilities as well as additional pathway node information. Enzyme nodes can be painted with corresponding available transcriptomic or proteomic data using the *Paint Enzymes* menu. For instance, the pyruvate metabolism pathway is also present in *Aspergillus*. The pathway record pages can be accessed from the gene record page for *A. nidulans* pycA gene AN4462. Gene expression profiles can then be painted onto individual nodes in *A. nidulans* via the *Paint Enzyme* menu option (Figure 10b). Any gene that does not have expression data for the chosen organism/experiment will have the text “None” in the enzyme box.


(a)
1 Pathways and interactions §
1.1 Metabolic pathways
Cytoscape Drawing
Show Contents »
File
Layout
Paint Enzymes
Pyruvate metabolism
Glycine, serine and threonine metabolism
Glycolysis / Gluconeogenesis
Nicotinate and nicotinamide metabolism
Valine, leucine and isoleucine biosynthesis
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Citrate cycle (TCA cycle)
(b)
Clear all
By Experiment
By Genera
phosphoenolpyruvic acid
4.1.1.32
4.1.1.38
oxaloacetic acid
Paint

Pyruvate metabolism

Pathway Source: KEGG
Pathway Id: ec00620
Total Pathway Enzymes: 75
Total Pathway Compounds: 34

NOTE Click on nodes for mo… …ghted in red are EC

Experiment Selector

nidu

Transcriptomics
RNASeq
○ A. nidulans FGSC A4 Inducer-Free Cellulase Secretion in Aspergillus nidulans (Glass)
● A. nidulans FGSC A4 Transcriptomes of wild-type vs veA and mtfA deletion mutants (Lind et al.)

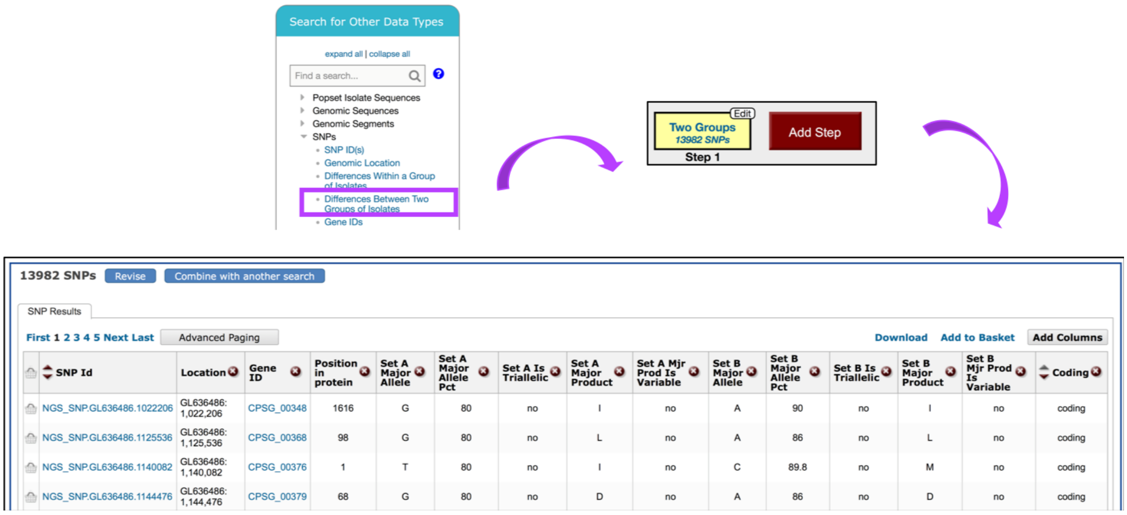
Figure 10. Interactive pathway map for KEGG ec00620 (Pyruvate metabolism). (**a**) Enzyme nodes that are represented by a gene in the database are outlined in red. Clicking on enzyme or compound nodes reveals more information about corresponding entities (**b**) Pyruvate metabolism pathway with RNA-Seq fpkm values painted above individual pathway nodes (*A. nidulans* transcriptomics data) via the *Paint Enzyme* menu (highlighted in purple).

6. Data Mining via the *Search for Other Data Types* Panel

6.1. Identifying SNPs between Fungal Isolates Collected in Various Geographical Areas

The example described below identifies SNPs in *Coccidioides posadasii (C. posadasii)* str. Silveira isolates collected from patients with Coccidioidomycosis in the U.S. and Latin America [34]. Coccidioidomycosis, also known as Valley fever, is a fungal disease caused by two closely related species—*C. immitis* and *C. posadasii* [34,35]. The disease is associated with high morbidity and mortality rates and affects tens of thousands of people each year. The two fungal species are endemic to several regions in the Western Hemisphere, but recent epidemiological and population studies suggest that the geographic range of these fungal species is becoming wider [36].

To identify SNPs between isolates collected in *Guatemala* and the *U.S.*, navigate to the *Identify SNPs based on Differences Between Two Groups of Isolates* from the *Search for Other Data Types panel* (Figure 11). In the resulting window, scroll through the metadata options on the left and make appropriate *Geographic Location* selections from the *Host* section of *Characteristic* separately for Set A and Set B isolates. Set A isolates should be set to *Guatemala* and Set B to the *United States of America*. All other parameters for both sets should be left as default.



13982 SNPs

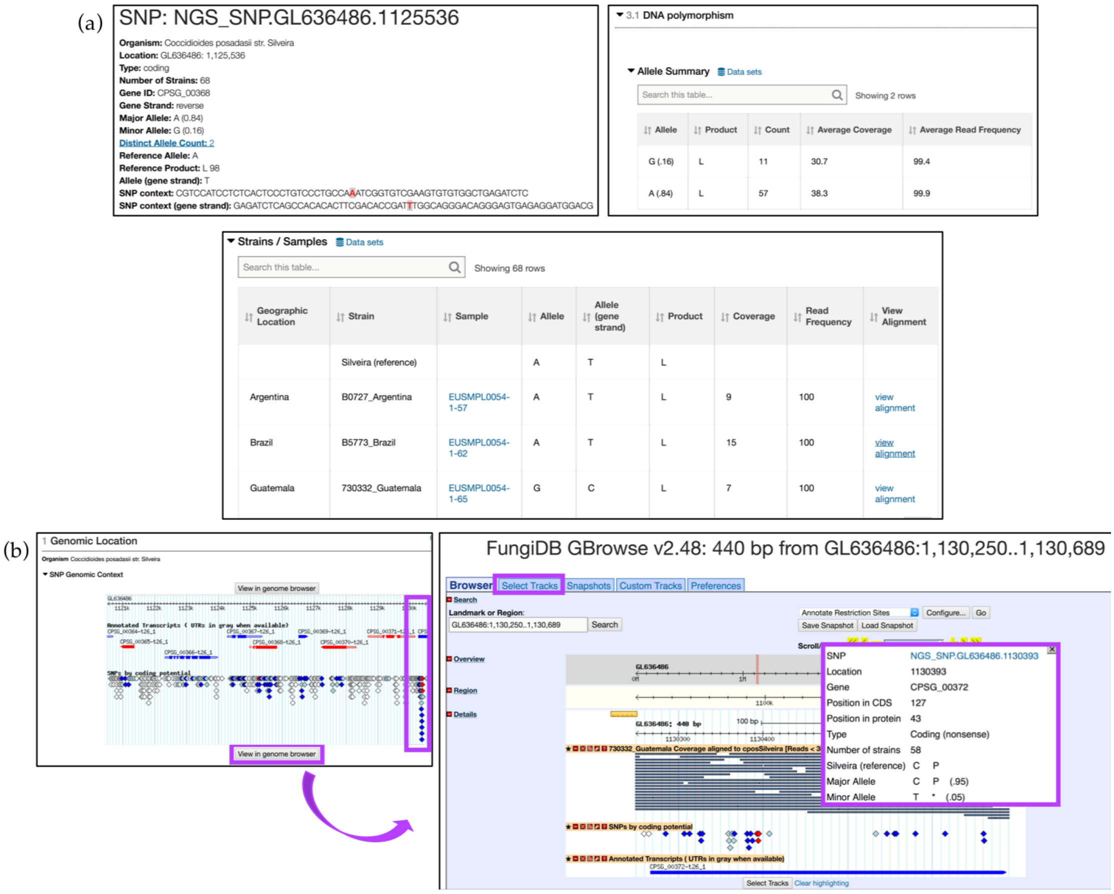
First 1 2 3 4 5 Next Last

SNP Id	Location	Gene ID	Position in protein	Set A Major Allele	Set A Major Allele Pct	Set A Is Triallelic	Set A Major Product	Set A Mjr Prod Is Variable	Set B Major Allele	Set B Major Allele Pct	Set B Is Triallelic	Set B Major Product	Set B Mjr Prod Is Variable	Coding
NGS_SNP.GL636486.1022206	GL636486: 1,022,206	CPSG_00348	1616	G	80	no	I	no	A	90	no	I	no	coding
NGS_SNP.GL636486.1125536	GL636486: 1,125,536	CPSG_00368	98	G	80	no	L	no	A	86	no	L	no	coding
NGS_SNP.GL636486.1140082	GL636486: 1,140,082	CPSG_00376	1	T	80	no	I	no	C	89.8	no	M	no	coding
NGS_SNP.GL636486.1144476	GL636486: 1,144,476	CPSG_00379	68	G	80	no	D	no	A	86	no	D	no	coding

Figure 11. Search for SNPs via the *Differences Between Two Groups of Isolates* option. SNP search is initiated from the *Search for Other Data Types* panel on the home page (highlighted in purple). The SNPs results table shown is sorted by *Coding* column.

The search strategy returns SNPs rather than genes, which are classified by genomic location within the results table. When individual SNPs fall within a gene, the corresponding Gene ID is listed next to the SNP record (Figure 11).

To examine an SNP record page, sort the results by the *Coding* polymorphisms and click on the SNP.GL636486.1125536 SNP in the CPSG_00368 gene, located within the third page of the results. SNP location, allele summary, associated Gene ID, and major and minor allele records can be found at the top of the page, followed by DNA polymorphism summary and an SNP records table that is searchable by isolates (Figure 12a). Genomic location, SNP type, and aligned reads can be displayed in GBrowse by clicking on the *View in genome browser* button (Figure 12b). The SNP track can be activated from the *Select Tracks* tab by selecting *SNPs by coding potential* under *DNA polymorphism* in the *Genetic variation* section. Hover over SNPs labeled with red diamonds (nonsense SNPs) to get more information. Note that individual reads comprising SNP data can be activated from the *Genetic variation* section by selecting either all or specific isolates listed under the *Aligned Genomic Sequence Reads for C. posadasii* str. Silveira.



SNP: NGS_SNP.GL636486.1125536

Organism: Coccidioides posadasii str. Silveira
Location: GL636486: 1,125,536
Type: coding
Number of Strains: 68
Gene ID: CPSG_00368
Gene Strand: reverse
Major Allele: A (0.84)
Minor Allele: G (0.16)
Distinct Allele Count: 2
Reference Allele: A
Reference Product: L 98
Allele (gene strand): T
SNP context: CGTCCATCCTCTCACTCCCTGTCCCTGCCAAATCGGTGTCGAAGTGTGTGGCTGAGATCTC
SNP context (gene strand): GAGATCTCAGCCACACACTTCGACACCGATTTGGCAGGGACAGGGAGTGAGAGGATGGACG

3.1 DNA polymorphism

Allele Summary Data sets

Search this table... Showing 2 rows

Allele	Product	Count	Average Coverage	Average Read Frequency
G (.16)	L	11	30.7	99.4
A (.84)	L	57	38.3	99.9

Strains / Samples Data sets

Search this table... Showing 68 rows

Geographic Location	Strain	Sample	Allele	Allele (gene strand)	Product	Coverage	Read Frequency	View Alignment
	Silveira (reference)		A	T	L			
Argentina	B0727_Argentina	EUSMPL0054-1-57	A	T	L	9	100	view alignment
Brazil	B5773_Brazil	EUSMPL0054-1-62	A	T	L	15	100	view alignment
Guatemala	730332_Guatemala	EUSMPL0054-1-65	G	C	L	7	100	view alignment

FungiDB GBrowse v2.48: 440 bp from GL636486:1,130,250..1,130,689

Landmark or Region: GL636486:1,130,250..1,130,689 Search

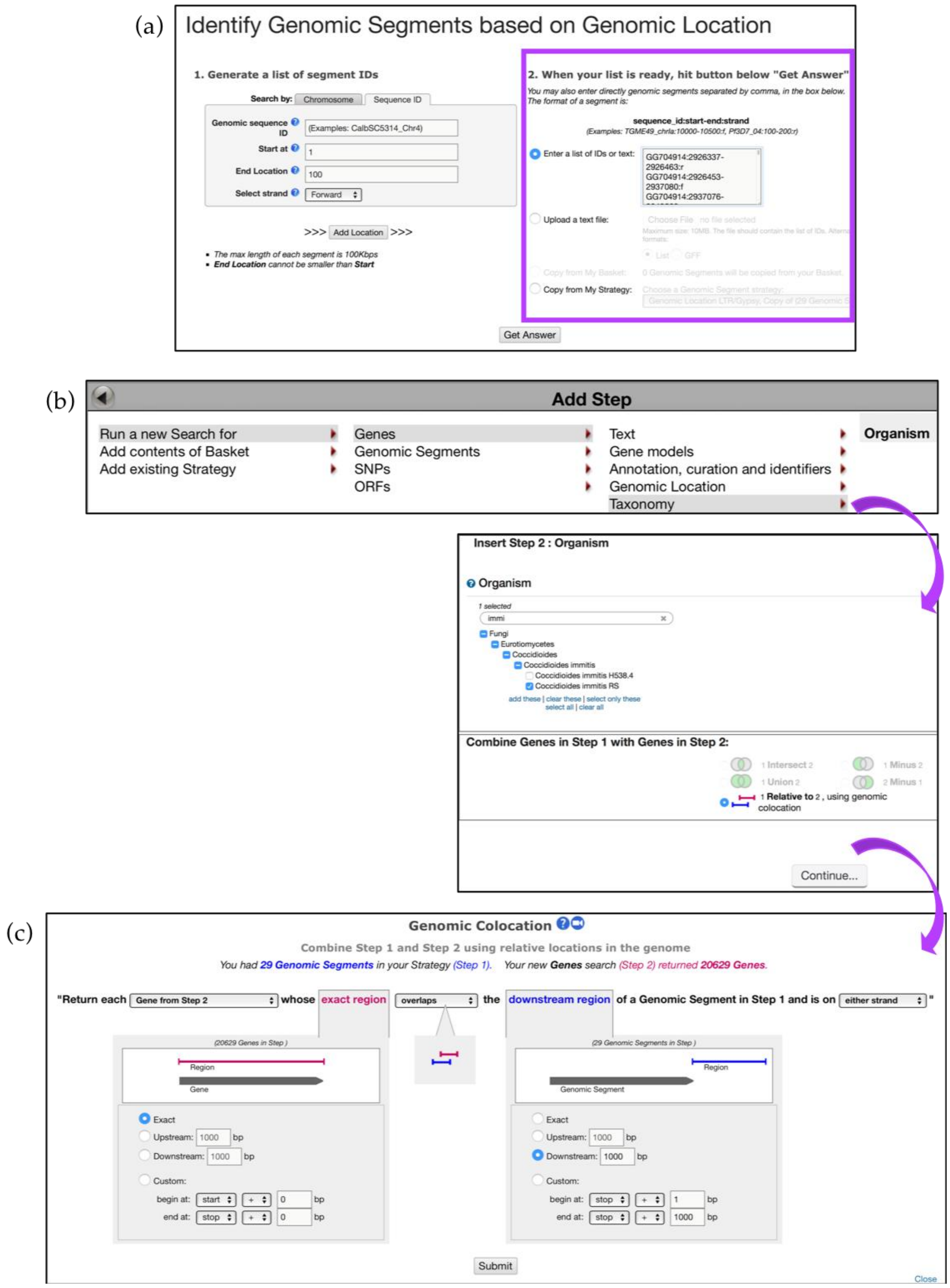
SNP	NGS_SNP.GL636486.1130393
Location	1130393
Gene	CPSG_00372
Position in CDS	127
Position in protein	43
Type	Coding (nonsense)
Number of strains	58
Silveira (reference)	C P
Major Allele	C P (.95)
Minor Allele	T * (.05)

Figure 12. Search for SNPs via the *Differences Between Two Groups of Isolates* option, continued. (**a**) SNP record for SNP.GL636486.1125536 corresponding to the CPSG_00368 gene. *DNA polymorphism* and SNP categorization based on *Strains/Samples* are shown; (**b**) Explore SNPs in FungiDB GBrowse. SNPs are color coded based on their effect on gene function (e.g., nonsense SNPs are shown as red diamonds-highlighted in purple). Click on the *View in genome browser* button to be re-directed to GBrowse. Individual reads can be activated from the *Select Tracks* tab (highlighted in purple). Hover over a SNP to activate a pop-up box with additional information (highlighted in purple). Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=564aa900d161cbbb.

6.2. Utilizing the Genomic Location Search and Genomic Colocation Tool in Custom Queries

The *Genomic Location* search tool empowers researchers to create custom queries for genomic regions of interest (e.g., custom sets of coordinates). The strategy described below investigates the effect of LTR/Gypsy Transposable Elements (TEs) (loci provided by Theo Kirkland; unpublished data) on expression of neighboring genes in *C. immitis* RS. This search strategy takes advantage of two types of queries reviewed above: *Search Other Data Types* and *Search for Genes*.

Twenty-nine LTR/Gypsy genomic coordinates were entered in the search window initiated via the *Search for Other Data Types*, *Genomic Segments*, *Genomic Location* selection menus (Figure 13a). All genomic locations were previously formatted to conform to the following format: *sequence_id:start-end:strand* (e.g., *GG704914:2926337-2926463:r*).



Identify Genomic Segments based on Genomic Location

1. Generate a list of segment IDs

Search by: Chromosome | Sequence ID

Genomic sequence ID: (Examples: CalbSC5314_Chr4)
Start at: 1
End Location: 100
Select strand: Forward

>>> Add Location >>>

- *The max length of each segment is 100Kbps*
- ***End Location** cannot be smaller than **Start***

2. When your list is ready, hit button below "Get Answer"

You may also enter directly genomic segments separated by comma, in the box below. The format of a segment is:

sequence_id:start-end:strand
(Examples: TGME49_chrla:10000-10500:f, Pf3D7_04:100-200:r)

● Enter a list of IDs or text: GG704914:2926337-2926463:r GG704914:2926453-2937080:f GG704914:2937076-

○ Upload a text file: Choose File no file selected
Maximum size: 10MB. The file should contain the list of IDs. Alterna formats:
● List ○ GFF

○ Copy from My Basket: 0 Genomic Segments will be copied from your Basket.

○ Copy from My Strategy: Choose a Genomic Segment strategy: Genomic Location LTR/Gypsy, Copy of (29 Genomic S

Get Answer

Insert Step 2 : Organism

Organism

1 selected
immi
- Fungi
  - Eurotiomycetes
    - Coccidioides
      - Coccidioides immitis
        - □ Coccidioides immitis H538.4
        - ☑ Coccidioides immitis RS

add these | clear these | select only these
select all | clear all

Combine Genes in Step 1 with Genes in Step 2:

○ 1 Intersect 2 ○ 1 Minus 2
○ 1 Union 2 ○ 2 Minus 1
● 1 **Relative to** 2 , using genomic colocation

Genomic Colocation

Combine Step 1 and Step 2 using relative locations in the genome

*You had **29 Genomic Segments** in your Strategy (Step 1). Your new **Genes** search (Step 2) returned **20629 Genes**.*

"Return each Gene from Step 2 whose exact region overlaps the downstream region of a Genomic Segment in Step 1 and is on either strand"

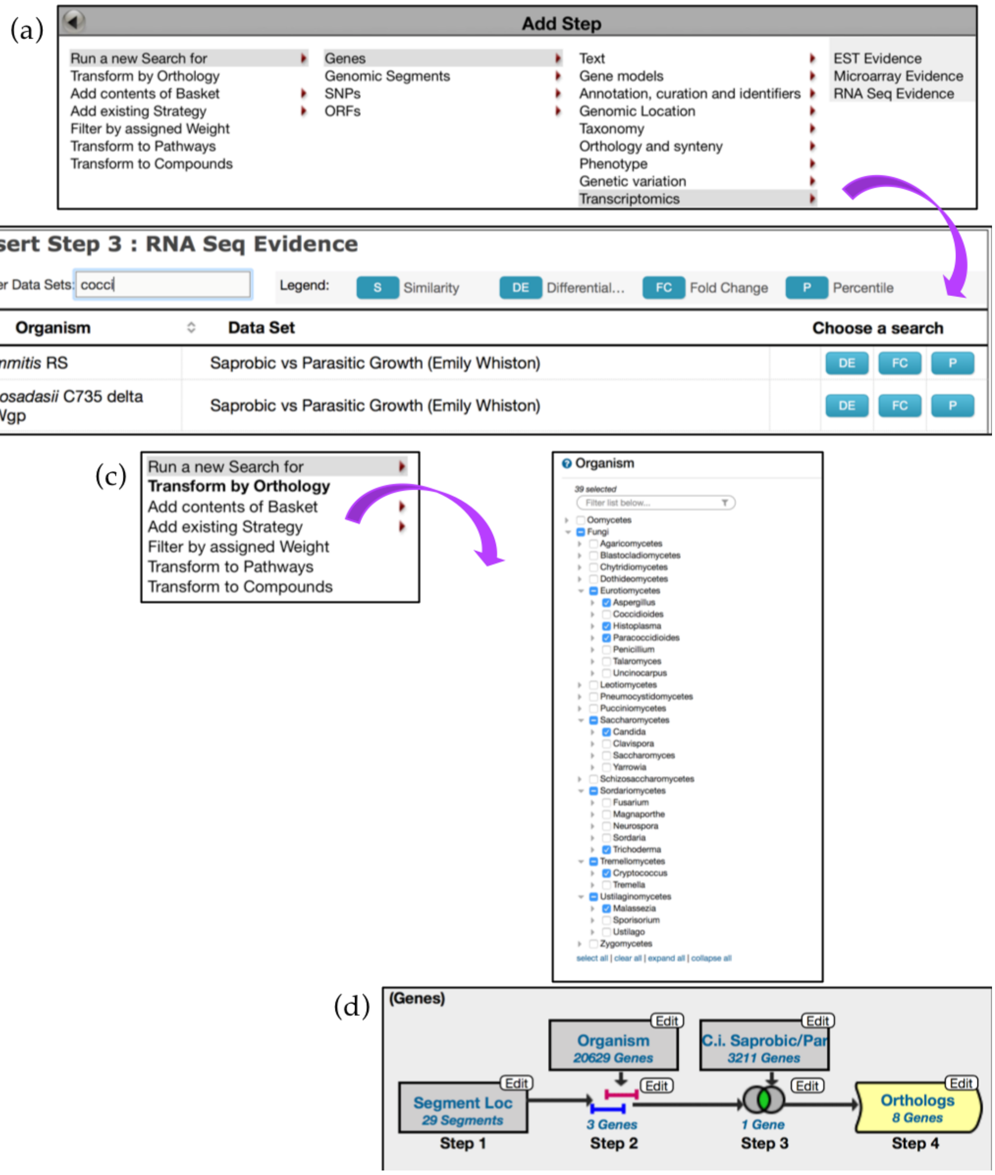
(20629 Genes in Step)
● Exact
○ Upstream: 1000 bp
○ Downstream: 1000 bp
○ Custom:
begin at: start + 0 bp
end at: stop + 0 bp

(29 Genomic Segments in Step)
○ Exact
○ Upstream: 1000 bp
● Downstream: 1000 bp
○ Custom:
begin at: stop + 1 bp
end at: stop + 1000 bp

Figure 13. Using *Genomic Location* search to discover impact of Transposable Elements (TEs) on gene expression. (**a**) Identify TEs based on coordinates formatted to the accepted convention (e.g., *sequence_id:start-end:strand). Paste a list of genomic coordinates in the highlighted section*—Step 1 (**b**) *Add Step* and *Run a new Search* using a *Taxonomy* tool and identify *C. immitis genes that* are (**c**) 1000bp downstream of the previously identified TE elements—Step 2. Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=67dcb3bb26a92111.

Initiate a new search and *Run a new Search for Genes, Taxonomy, Organism*, and choose *C. immitis* and the *1 relative to 2* search option (Figure 13b). The next screen will bring up the *Genomic Colocation* tool that requires selection of several search parameters. Choose to *Return each Gene from Step 2 whose exact region overlaps with downstream region of a Genomic Segments in Step 1 and is on either strand.* Look for genes that are located 1000 bp downstream of the segments identified in Step 1 (Figure 13c).

The second search returned three genes. Next, examine transcriptomic records integrated in FungiDB by adding a dataset from the *RNA Seq Evidence menu* for *C. immitis (Saprobic vs Parasitic Growth (Emily Whiston)) dataset* (Figure 14a,b). Based on the search criteria entered, only one gene was identified—CIMG_13457, a hypothetical protein. One can try to identify orthologs of this gene in other species by adding another step and deploying the *Transform by Orthology* tool for *Eurotiomycetes*, *Saccharomycetes (Candida)*, *Sordariomycetes (Trichoderma)*, *Tremellomycetes (Cryptococcus)*, and *Ustilaginomycetes (Malassezia)* (Figure 14c). The search identified eight orthologs in other pathogens with the majority occurring in *Histoplasma* species and a single ortholog in *P. brasiliensis* Pb03 (Figure 14d).



Filter Data Sets: cocci

Organism	Data Set	Choose a search
C. immitis RS	Saprobic vs Parasitic Growth (Emily Whiston)	DE FC P
C. posadasii C735 delta SOWgp	Saprobic vs Parasitic Growth (Emily Whiston)	DE FC P

39 selected

Filter list below...

- ☐ Oomycetes
- ▣ Fungi
  - ☐ Agaricomycetes
  - ☐ Blastocladiomycetes
  - ☐ Chytridiomycetes
  - ☐ Dothideomycetes
  - ▣ Eurotiomycetes
    - ☑ Aspergillus
    - ☐ Coccidioides
    - ☑ Histoplasma
    - ☑ Paracoccidioides
    - ☐ Penicillium
    - ☐ Talaromyces
    - ☐ Uncinocarpus
  - ☐ Leotiomycetes
  - ☐ Pneumocystidomycetes
  - ☐ Pucciniomycetes
  - ▣ Saccharomycetes
    - ☑ Candida
    - ☐ Clavispora
    - ☐ Saccharomyces
    - ☐ Yarrowia
  - ☐ Schizosaccharomycetes
  - ▣ Sordariomycetes
    - ☐ Fusarium
    - ☐ Magnaporthe
    - ☐ Neurospora
    - ☐ Sordaria
    - ☑ Trichoderma
  - ▣ Tremellomycetes
    - ☑ Cryptococcus
    - ☐ Tremella
  - ▣ Ustilaginomycetes
    - ☑ Malassezia
    - ☐ Sporisorium
    - ☐ Ustilago
  - ☐ Zygomycetes

Segment Loc *29 Segments* Step 1
Organism *20629 Genes*
3 Genes Step 2
C.i. Saprobic/Par *3211 Genes*
1 Gene Step 3
Orthologs *8 Genes* Step 4

Figure 14. Using *Genomic Location* search to discover impact of Transposable Elements (TEs) on gene expression, continued. (**a**,**b**) *Add Step* to mine transcriptomics data for *C. immitis*—Step 3 (**c**) Step 4: *Add step* to deploy the *Transform by Orthology* tool. (**d**) Overview of the search strategy. The shape of the search box highlighted in yellow is different to reflect changes in fungal species/genera. Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=67dcb3bb26a92111.

6.3. Searching for Fungal-Specific Motifs

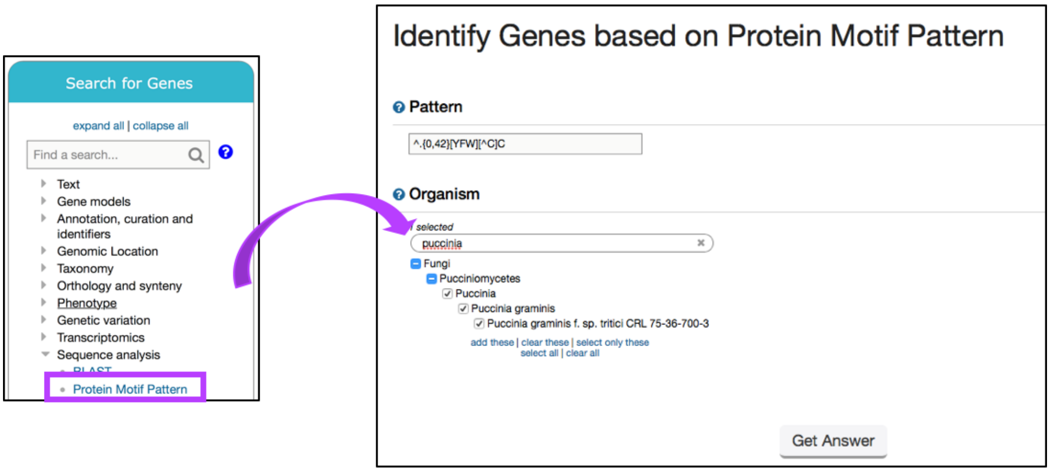
Proteins which perform similar functions often contain conserved sequences such as protein domains or motifs that facilitate important functions such as enzyme binding, post-translational modifications, or cellular trafficking. FungiDB offers tools for finding peptide and DNA motifs in target sequences. The strategy described below takes advantage of the *Protein Motif Pattern* search tool to identify effector proteins in the wheat stem rust fungi *Puccinia gramminis* f. sp. tritici.

Rust fungi are pathogens of crops that depend on developing haustoria in the living plant cells to support nutrient absorption and sporulation. A new class of Y/F/WxC effector proteins was reported in haustoria-producing plant pathogens by Godfrey *et al.* [37]. Before starting a search, the amino acid pattern of interest must be converted into a Perl-style regular expression. Using Perl characters (Table 1), the Y/F/WxC motif is expressed as {0,42}[YFW][ˆC]C, where [YFW] is a 3-amino acid peptide that has an aromatic amino acid, followed by any amino acid except a cysteine, followed by a cysteine, with any amino acid occurring between 0 and 42 times before the 3-amino acid motif.

Table 1. Examples of special characters used in Perl-style regular expressions.

Character	Meaning
A, B, C . . .	Amino acid single letter abbreviation for peptide motifs
[ˆa]	Any character except a
{n,m}	Match the preceding character between n and m times
[ ]	Match any character contained in the bracket

Next, navigate to the *Search for Genes* panel, *Sequence analysis* menu, and then click on the *Protein Motif Pattern* submenu (Figure 15). In the *Identify Genes Based on Protein Motif Pattern* page, enter {0,42}[YFW][ˆC]C in the *Pattern* parameter and select *P. gramminis* from the list of fungal organisms. *Run a new Search for Genes* based on *Protein Targeting and Localization* and identify if any of the genes have a *Predicted Signal Peptide* (Figure 16a). Choosing to intersect your search with previous results will identify several hundred proteins with predicted Y/F/WxC motif and signal peptide domains. One hundred and thirty-two of these genes matched the BLAST analysis reported by Godfrey *et al.* [37] (Figure 16b). Closer examination of additional gene hits would be needed to determine their biological relevance for plant–fungal interactions.



Identify Genes based on Protein Motif Pattern

Pattern

^.{0,42}[YFW][^C]C

Organism

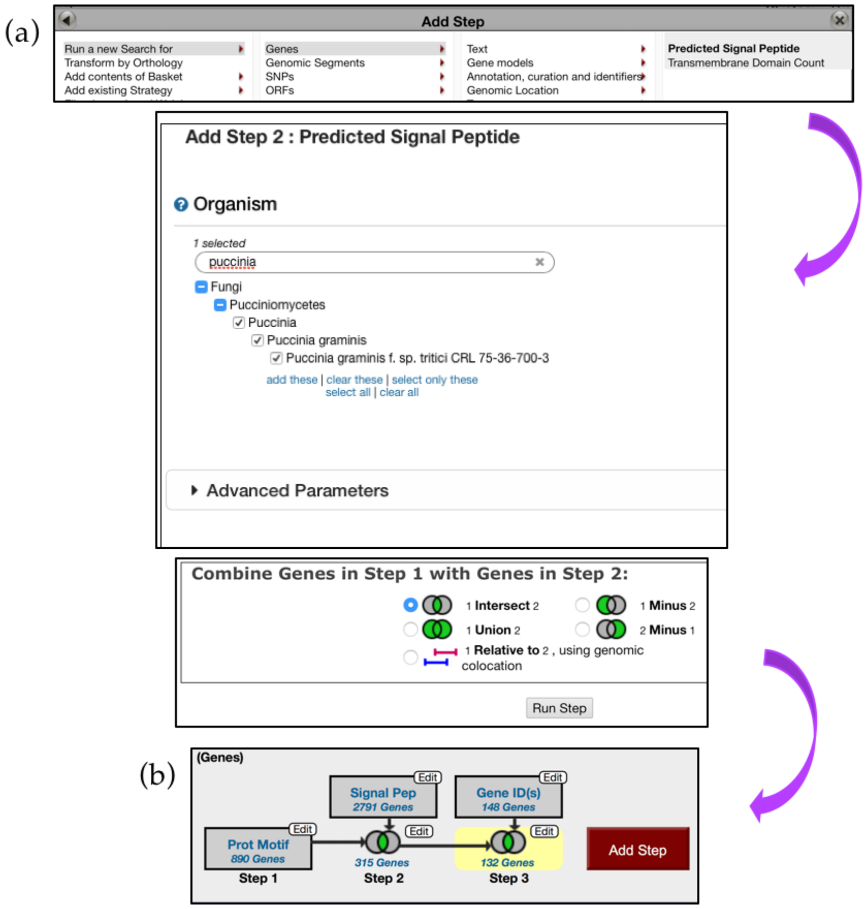
selected

puccinia

Fungi
Pucciniomycetes
☑ Puccinia
☑ Puccinia graminis
☑ Puccinia graminis f. sp. tritici CRL 75-36-700-3

add these | clear these | select only these
select all | clear all

Figure 15. Identifying genes via the *Motif Pattern* tool. Shown is a search for the effector protein motif in *P. gramminis* using a regular expression (Step 1). Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=4e9454cd679286de.



Add Step 2 : Predicted Signal Peptide

Organism

1 selected

puccinia

- Fungi
  - Pucciniomycetes
    - ☑ Puccinia
      - ☑ Puccinia graminis
        - ☑ Puccinia graminis f. sp. tritici CRL 75-36-700-3

Combine Genes in Step 1 with Genes in Step 2:

● 1 Intersect 2 ○ 1 Minus 2
○ 1 Union 2 ○ 2 Minus 1
○ 1 Relative to 2 , using genomic colocation

Figure 16. Identifying genes via the *Motif Pattern* tool, continued. (**a**) *Predicted Signal Peptides* search deployed via the *Add Step* function (Step 2) (**b**) Overview of the search strategy. Search results from Step 2 are intersected with 148 genes reported by Godfrey *et al.* (Step 3) Strategy in FungiDB: http://fungidb.org/fungidb/im.do?s=4e9454cd679286de.

7. EuPathDB/FungiDB Galaxy Workspace

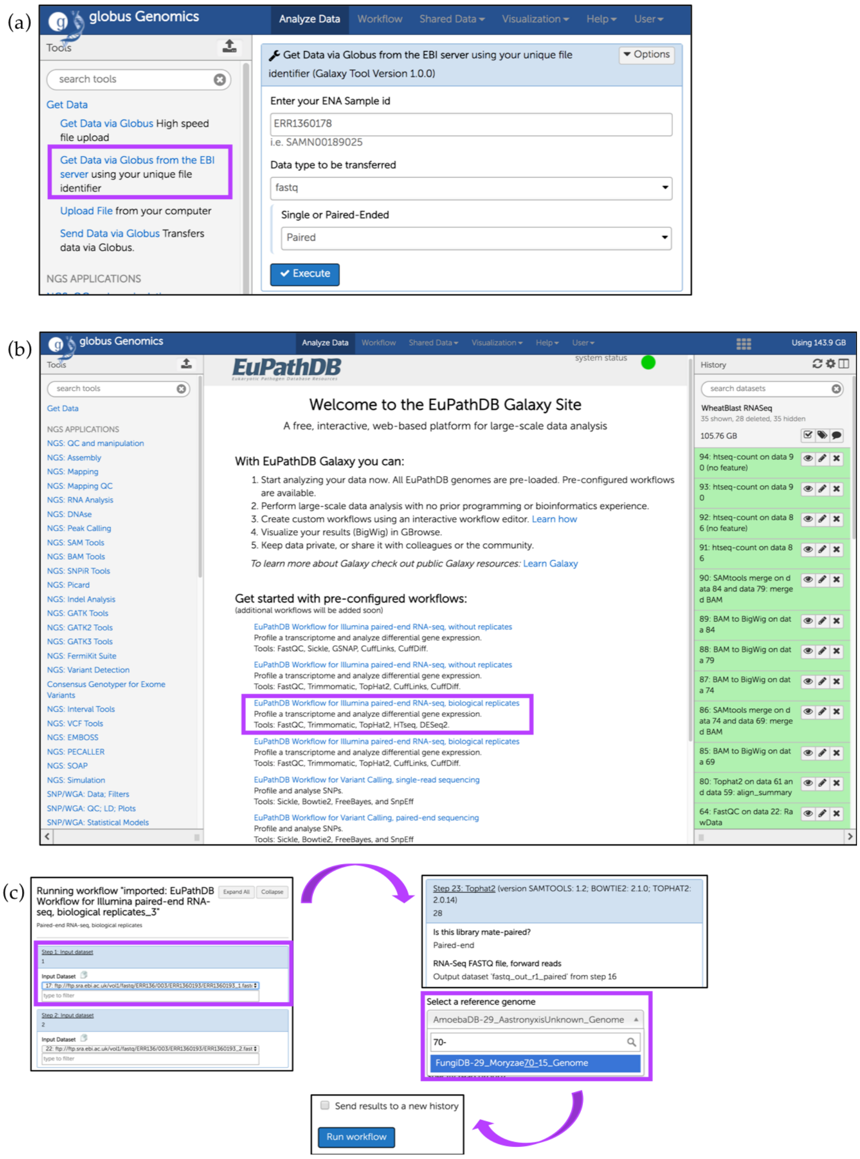
EuPathDB Galaxy-based [38] workspace offers preloaded genomes, private data analysis and display, and the ability to share and export analysis results [2]. Developed in partnership with Globus Genomics [39] and accessed through the *Analyze My Experiment* tab on the home page of FungiDB, the EuPathDB Galaxy instance houses a large variety of bioinformatics tools. Users can create their own private Galaxy workspace by creating a Globus account through FungiDB or any other EuPathDB database. After an account is created, sample workflows can be imported for editing, and new custom workflows can be created which can be stored privately or shared with the wider community or between EuPathDB Galaxy users. Results can be exported or shared with collaborators, but the instance is not meant as a repository for long-term data storage. All genomes available in FungiDB/EuPathDB sites are already preloaded into the EuPathDB Galaxy instance and, alongside these, users can upload their own data for analysis.

Running a Preconfigured (Sample) Workflow

The EuPathDB Galaxy workspace has several workflows for RNA-Seq analysis and SNP calling (available from the EuPathDB Galaxy Welcome page). Each pipeline can be initiated directly from the welcome page; however, necessary input files must be uploaded prior to deploying a preconfigured workflow.

The analysis described here uses files obtained from the EBI BioProject PRJEB13569, which is a RNA-Seq dataset deposited by the Wheat Open Blast initiative [40]. This dataset comprises RNA-Seq samples collected from wheat leaf blades harvested from wheat-blast-infected fields in Bangladesh [41]. Wheat blast is a fungal disease caused by *Magnaporthe oryzae* (*M. oryzae*), a plant pathogen threatening crops globally.

To load the following files into the Galaxy instance, click on the *Get Data* link in the *Tools* section, select the *Get Data via Globus from the EBI server* option (Figure 17a), and enter Sample IDs. As an example, we will use two Asymptomatic samples F12 (Run accession numbers: ERR1360178, ERR1360179) and Symptomatic samples 12 (Run accession numbers: ERR1360192, ERR1360193). Specify *paired-end* sequencing for each accession number and click *Execute*.


(a)
globus Genomics
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
search tools
Get Data
Get Data via Globus High speed file upload
Get Data via Globus from the EBI server using your unique file identifier
Upload File from your computer
Send Data via Globus Transfers data via Globus.
NGS APPLICATIONS
Options
Execute

Get Data via Globus from the EBI server using your unique file identifier (Galaxy Tool Version 1.0.0)

Enter your ENA Sample id: ERR1360178
i.e. SAMN00189025

Data type to be transferred: fastq

Single or Paired-Ended: Paired

Welcome to the EuPathDB Galaxy Site

A free, interactive, web-based platform for large-scale data analysis

With EuPathDB Galaxy you can:

1. Start analyzing your data now. All EuPathDB genomes are pre-loaded. Pre-configured workflows are available.
2. Perform large-scale data analysis with no prior programming or bioinformatics experience.
3. Create custom workflows using an interactive workflow editor. Learn how
4. Visualize your results (BigWig) in GBrowse.
5. Keep data private, or share it with colleagues or the community.

To learn more about Galaxy check out public Galaxy resources: Learn Galaxy

Get started with pre-configured workflows:
(additional workflows will be added soon)

EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Sickle, GSNAP, CuffLinks, CuffDiff.

EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Trimmomatic, TopHat2, CuffLinks, CuffDiff.

EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Trimmomatic, TopHat2, HTseq, DESeq2.

EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Trimmomatic, TopHat2, CuffLinks, CuffDiff.

EuPathDB Workflow for Variant Calling, single-read sequencing
Profile and analyse SNPs.
Tools: Sickle, Bowtie2, FreeBayes, and SnpEff

EuPathDB Workflow for Variant Calling, paired-end sequencing
Profile and analyse SNPs.
Tools: Sickle, Bowtie2, FreeBayes, and SnpEff

WheatBlast RNASeq
35 shown, 28 deleted, 35 hidden
105.76 GB

94: htseq-count on data 90 (no feature)
93: htseq-count on data 90
92: htseq-count on data 86 (no feature)
91: htseq-count on data 86
90: SAMtools merge on data 84 and data 79: merged BAM
89: BAM to BigWig on data 84
88: BAM to BigWig on data 79
87: BAM to BigWig on data 74
86: SAMtools merge on data 74 and data 69: merged BAM
85: BAM to BigWig on data 69
80: Tophat2 on data 61 and data 59: align_summary
64: FastQC on data 22: RawData

Running workflow "imported: EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates_3"
Paired-end RNA-seq, biological replicates

Step 1: Input dataset
1
Input Dataset: 17: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR136/003/ERR1360193/ERR1360193_1.fastq
type to filter

Step 2: Input dataset
2
Input Dataset: 22: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR136/003/ERR1360193/ERR1360193_2.fast
type to filter

Step 23: Tophat2 (version SAMTOOLS: 1.2; BOWTIE2: 2.1.0; TOPHAT2: 2.0.14)
28

Is this library mate-paired?
Paired-end

RNA-Seq FASTQ file, forward reads
Output dataset 'fastq_out_r1_paired' from step 16

Select a reference genome
AmoebaDB-29_AastronyxisUnknown_Genome
70-
FungiDB-29_Moryzae70-15_Genome

□ Send results to a new history

Figure 17. Running a preconfigured workflow for paired-end RNA-Seq data with biological replicates. (**a**) Upload files from EBI server (highlighted in purple); (**b**) Initiate a sample workflow by clicking on the corresponding workflow link (highlighted); (**c**) Choose preloaded paired datasets from the *Input dataset* drop down menu (highlighted). Some tools (e.g., Tophat2) require specification of the reference genome (highlighted) which will be used to map reads. Use search box to find and select the reference genome. Click on *Run workflow* to initiate the analysis.

After all samples have been loaded successfully (highlighted in green), click on *Analyze Data* at the top of the EuPathDB Galaxy screen to return to the main page and click on the third workflow within the sample workflows list (Figure 17b); choose input datasets so that each pair forms a paired-end sample. This workflow uses FastQC [42], Trimmomatic [43], TopHat2 [44], HTseq-count [45], and DESeq2 [46]. Choose *M. oryzae* 70–15 genome for all workflow steps when prompted and then click on the *Run workflow* button (Figure 17c). To send this run into a new history, check the box *Send results to a new history* (Figure 17c). After the workflow has finished running, users can visualize the expression graphs (as BigWig files) in FungiDB GBrowse (Figure 18). To import BigWig files into FungiDB GBrowse, click on the *Display in FungiDB GBrowse* link, which will become visible once a user clicks to expand the *Bam to BigWig* step in the workflow EuPathDb Galaxy history section on the right (Figure 17b).

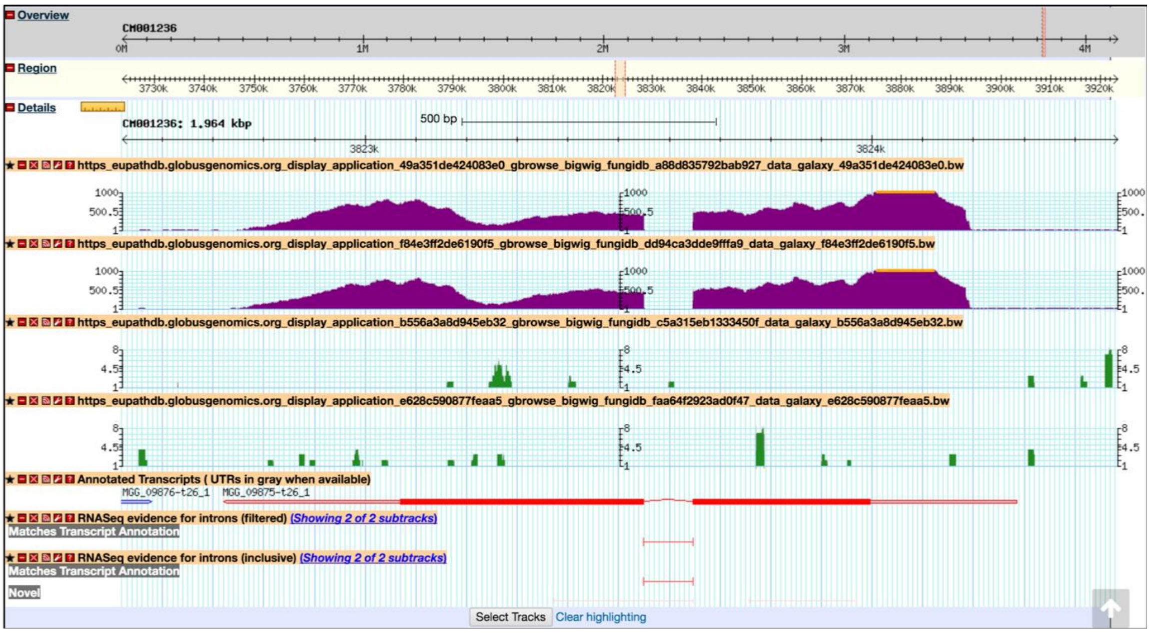

Overview
CM001236
0M
1M
2M
3M
4M
Region
3730k 3740k 3750k 3760k 3770k 3780k 3790k 3800k 3810k 3820k 3830k 3840k 3850k 3860k 3870k 3880k 3890k 3900k 3910k 3920k
Details
CM001236: 1.964 kbp
500 bp
3823k
3824k
https_eupathdb.globusgenomics.org_display_application_49a351de424083e0_gbrowse_bigwig_fungidb_a88d835792bab927_data_galaxy_49a351de424083e0.bw
1000
500.5
1
https_eupathdb.globusgenomics.org_display_application_f84e3ff2de6190f5_gbrowse_bigwig_fungidb_dd94ca3dde9fffa9_data_galaxy_f84e3ff2de6190f5.bw
1000
500.5
1
https_eupathdb.globusgenomics.org_display_application_b556a3a8d945eb32_gbrowse_bigwig_fungidb_c5a315eb1333450f_data_galaxy_b556a3a8d945eb32.bw
8
4.5
1
https_eupathdb.globusgenomics.org_display_application_e628c590877feaa5_gbrowse_bigwig_fungidb_faa64f2923ad0f47_data_galaxy_e628c590877feaa5.bw
8
4.5
1
Annotated Transcripts (UTRs in gray when available)
MGG_09876-t26_1
MGG_09875-t26_1
RNASeq evidence for introns (filtered) (Showing 2 of 2 subtracks)
Matches Transcript Annotation
RNASeq evidence for introns (inclusive) (Showing 2 of 2 subtracks)
Matches Transcript Annotation
Novel
Select Tracks
Clear highlighting

Figure 18. Importing BigWig files into the FungiDB GBrowse from the EuPathDB Galaxy instance. The data is scaled to local min/max. Shown are tracks from symptomatic and asymptomatic samples (both replicates), followed by Annotated Transcripts and FungiDB pre-configured RNASeq evidence tracks.

Previously, Mathioni *et al.* [47] reported that *MAS3*, a *CAS1*-like domain-containing protein, was implicated in virulence in the blast fungus *M. oryzae* infecting barley. To determine if this gene is also up-regulated during infection in wheat in the RNA-Seq samples analyzed here, navigate to *MAS3* gene (MGG_09875) by entering the following coordinates in the Landmark or region search window in GBrowse: CM001236:3,822,721.3,824,284. As a result, the GBrowse window along with all corresponding BigWig tracks will be rescaled to the region specified. Next, zoom out to 50 kbp and deploy the synteny option for Sordariomycetes from the *Select Tracks* tab. As demonstrated in Figure 19, MGG_09875 is syntenic with an *M. oryzae* BR32 gene as one would expect; however, it is lacking syntenic orthologs in other species. To access nonsyntenic orthologs of *MAS3* in other species, navigate to the *Orthology and Synteny* section within the gene record page.

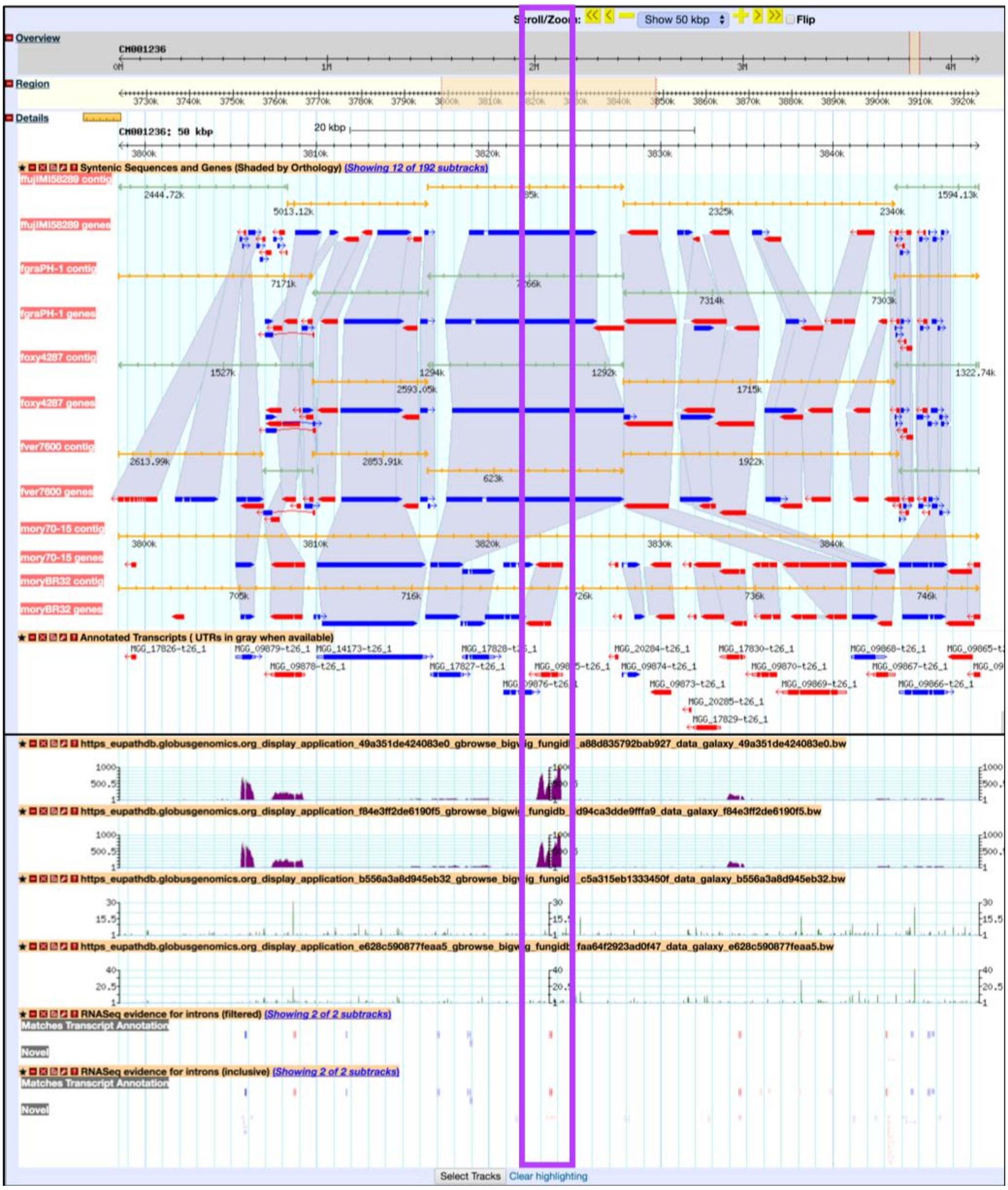
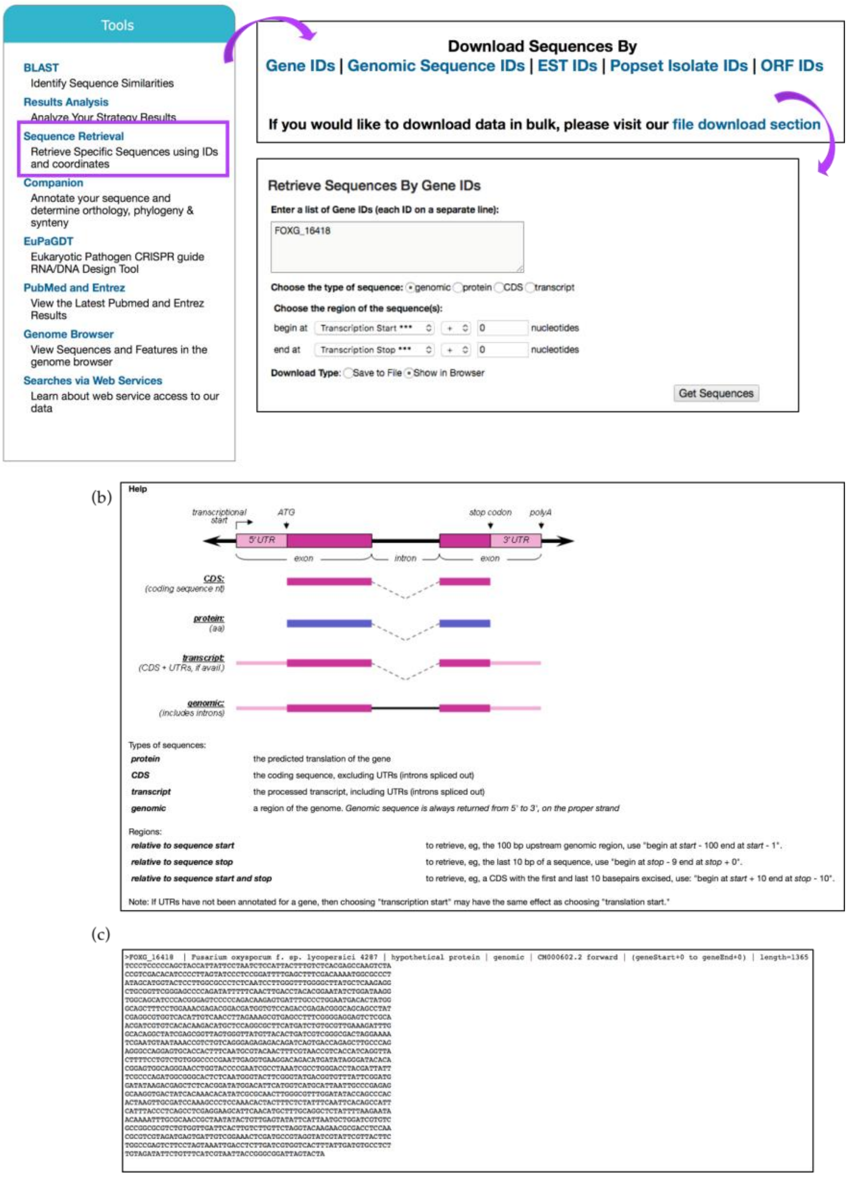

Scroll/Zoom:
Show 50 kbp
Flip
Overview
CM001236
Region
Details
CM001236: 50 kbp
20 kbp
Syntenic Sequences and Genes (Shaded by Orthology) (Showing 12 of 192 subtracks)
ffujIMI58289 contig
ffujIMI58289 genes
fgraPH-1 contig
fgraPH-1 genes
foxy4287 contig
foxy4287 genes
fver7600 contig
fver7600 genes
mory70-15 contig
mory70-15 genes
moryBR32 contig
moryBR32 genes
Annotated Transcripts (UTRs in gray when available)
https_eupathdb.globusgenomics.org_display_application_49a351de424083e0_gbrowse_bigwig_fungidb_a88d835792bab927_data_galaxy_49a351de424083e0.bw
https_eupathdb.globusgenomics.org_display_application_f84e3ff2de6190f5_gbrowse_bigwig_fungidb_d94ca3dde9fffa9_data_galaxy_f84e3ff2de6190f5.bw
https_eupathdb.globusgenomics.org_display_application_b556a3a8d945eb32_gbrowse_bigwig_fungidb_c5a315eb1333450f_data_galaxy_b556a3a8d945eb32.bw
https_eupathdb.globusgenomics.org_display_application_e628c590877feaa5_gbrowse_bigwig_fungidb_faa64f2923ad0f47_data_galaxy_e628c590877feaa5.bw
RNASeq evidence for introns (filtered) (Showing 2 of 2 subtracks)
Matches Transcript Annotation
Novel
RNASeq evidence for introns (inclusive) (Showing 2 of 2 subtracks)
Matches Transcript Annotation
Novel
Select Tracks
Clear highlighting

Figure 19. Viewing BigWig files in the FungiDB GBrowse. Shown are synteny tracks for selected Sordariomycetes (highlighted in purple). Tracks can be activated from the *Orthology and synteny* section, which is located under the *Select Tracks* tab.

8. Retrieve Sequences via *Sequence Retrieval Tool*

The *Sequence Retrieval Tool* (SRT) retrieves sequences in FASTA format using genome or proteome coordinates (Figure 20). SRT is accessible from the *Tools* component on the home page or from the *Tools* tab within the main grey main menu. This tool works independently from the data mining infrastructure used when creating queries and results can be shown in the browser or downloaded as a text file. The tool can be further customized by changing transcription/translation start and stop site settings (Figure 20a,b). *Fusarium oxysporum* f.sp. lycopersici 4287 is an important tomato pathogen. The FOXG_16418 gene encodes an effector protein that is required for virulence. To retrieve the genomic sequence of this gene, enter the gene ID in the text box within the *Retrieve Sequences By Gene IDs* window and select *genomic* sequence type.


Tools
BLAST
Identify Sequence Similarities
Results Analysis
Analyze Your Strategy Results
Sequence Retrieval
Retrieve Specific Sequences using IDs and coordinates
Companion
Annotate your sequence and determine orthology, phylogeny & synteny
EuPaGDT
Eukaryotic Pathogen CRISPR guide RNA/DNA Design Tool
PubMed and Entrez
View the Latest Pubmed and Entrez Results
Genome Browser
View Sequences and Features in the genome browser
Searches via Web Services
Learn about web service access to our data
Download Sequences By
Gene IDs | Genomic Sequence IDs | EST IDs | Popset Isolate IDs | ORF IDs
Get Sequences
(b)
Help
transcriptional start
ATG
stop codon
polyA
5′UTR
3′UTR
exon
intron
exon
CDS:
(coding sequence nt)
protein:
(aa)
transcript:
(CDS + UTRs, if avail.)
genomic:
(includes introns)
(c)

If you would like to download data in bulk, please visit our file download section

Retrieve Sequences By Gene IDs

Enter a list of Gene IDs (each ID on a separate line):

FOXG_16418

Choose the type of sequence: ● genomic ○ protein ○ CDS ○ transcript

Choose the region of the sequence(s):

begin at Transcription Start *** + 0 nucleotides

end at Transcription Stop *** + 0 nucleotides

Download Type: ○ Save to File ● Show in Browser

Types of sequences:

protein	the predicted translation of the gene
CDS	the coding sequence, excluding UTRs (introns spliced out)
transcript	the processed transcript, including UTRs (introns spliced out)
genomic	a region of the genome. *Genomic sequence is always returned from 5′ to 3′, on the proper strand*

Regions:

relative to sequence start	to retrieve, eg, the 100 bp upstream genomic region, use "begin at *start* - 100 end at *start* - 1".
relative to sequence stop	to retrieve, eg, the last 10 bp of a sequence, use "begin at *stop* - 9 end at *stop* + 0".
relative to sequence start and stop	to retrieve, eg, a CDS with the first and last 10 basepairs excised, use: "begin at *start* + 10 end at *stop* - 10".

Note: If UTRs have not been annotated for a gene, then choosing "transcription start" may have the same effect as choosing "translation start."

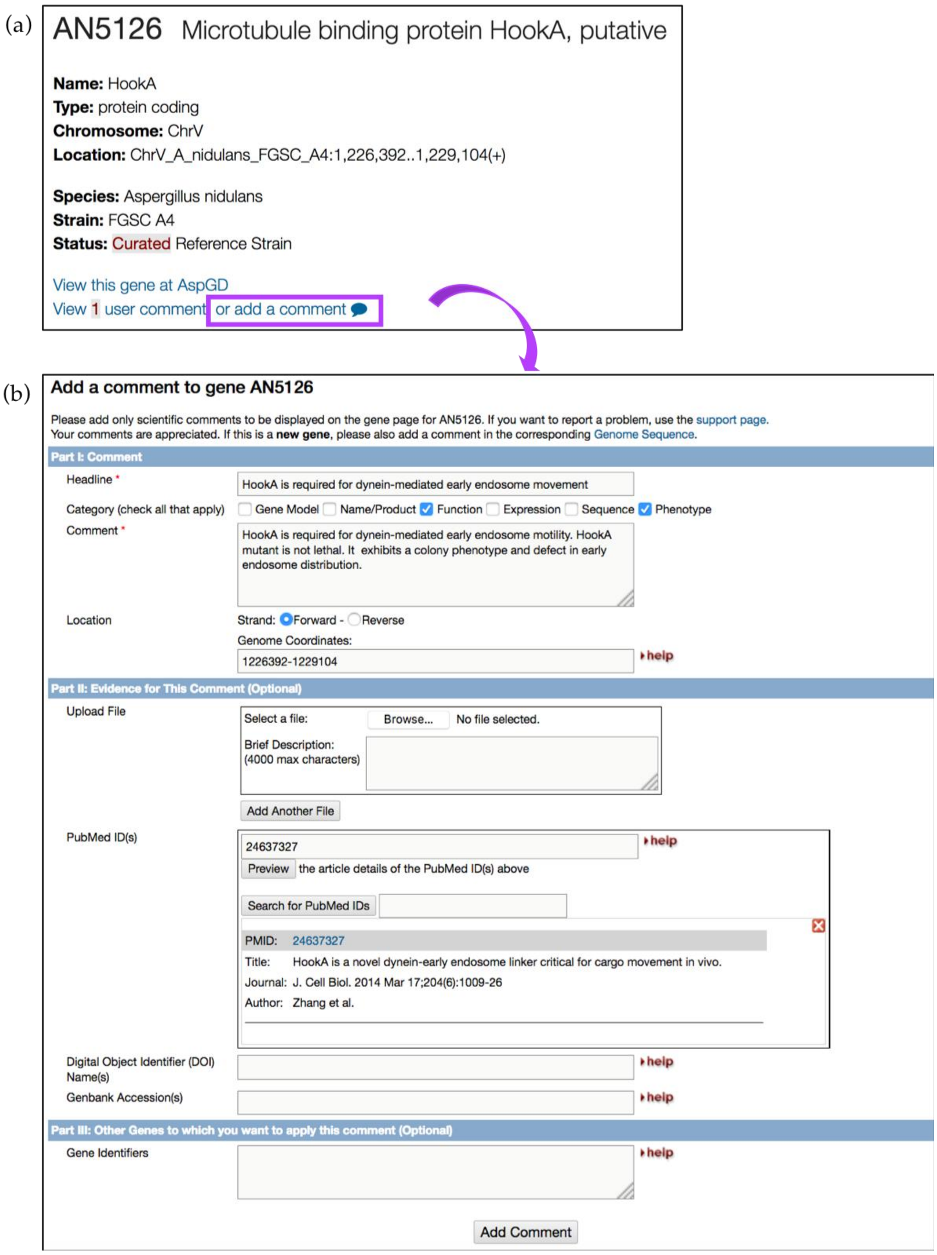
```
>FOXG_16418  | Fusarium oxysporum f. sp. lycopersici 4287 | hypothetical protein | genomic | CM000602.2 forward | (geneStart+0 to geneEnd+0) | length=1365
TCCCTCCCCCAGCTACCATTATTCCTAATCTCCATTACTTTGTCTCACGAGCCAAGTCTA
CCGTCGACACATCCCCTTAGTATCCCTCCGGATTTTGAGCTTTCGACAAAATGGCGCCCT
ATAGCATGGTACTCCTTGGCGCCCTCTCAATCCTTGGGTTTGGGGCTTATGCTCAAGAGG
CTGCGGTTCGGGAGCCCCAGATATTTTTCAACTTGACCTACACGGAATATCTGGATAAGG
TGGCAGCATCCCACGGGAGTCCCCCAGACAAGAGTGATTTGCCCTGGAATGACACTATGG
GCAGCTTTCCTGGAAACGAGACGGACGATGGTGTCCAGACCGAGACGGGCAGCAGCCTAT
CGAGGCGTGGTCACATTGTCAACCTTAGAAAGCGTGAGCCTTTCGGGGAGGAGTCTCGCA
ACGATCGTGTCACACAAGACATGCTCCAGGCGCTTCATGATCTGTGCGTTGAAAGATTTG
GCACAGGCTATCGAGCGGTTAGTGGGTTATGTTACACTGATCGTCGGGCGACTAGGAAAA
TCGAATGTAATAAACCGTCTGTCAGGGAGAGAGACAGATCAGTGACCAGAGCTTGCCCAG
AGGGCCAGGAGTGCACCACTTTCAATGCGTACAACTTTCGTAACCGTCACCATCAGGTTA
CTTTTCCTGTCTGTGGCCCCGAATTGAGGTGAAGGACAGACATGATATAGGGATACACA
CGGAGTGGCAGGGAACCTGGTACCCCGAATCGCCTAAATCGCCTGGGACCTACGATTATT
TCGCCCAGATGGCGGGCACTCTCAATGGGTACTTCGGGTATGACGGTGTTTATTCGGATG
GATATAAGACGAGCTCTCACGGATATGGACATTCATGGTCATGCATTAATTGCCCGAGAG
GCAAGGTGACTATCACAAACACATATCGCGCAACTTGGGCGTTTGGATATACCAGCCCAC
ACTAAGTTGCGATCCAAAGCCCTCCAAACACTACTTTCTCTATTTCAATTCACAGCCATT
CATTTACCCTCAGCCTCGAGGAAGCATTCAACATGCTTTGCAGGCTCTATTTTAAGAATA
ACAAAATTTGCGCAACCGCTAATATACTGTTGAGTATATTCATTAATGCTGGATCGTGTC
GCCGGCGCGTCTGTGGTTGATTCACTTGTCTTGTTCTAGGTACAAGAACGCGACCTCCAA
CGCGTCGTAGATGAGTGATTGTCGGAAACTCGATGCCGTAGGTATCGTATTCGTTACTTC
TGGCCGAGTCTTCCTAGTAAATTGACCTCTTGATCGTGGTCACTTTATTGATGTGCCTCT
TGTAGATATTCTGTTTCATCGTAATTACCGGGCGGATTAGTACTA
```

Figure 20. *Sequence Retrieval Tool*. (**a**) The tool can be initiated from the *Tools* component of the main page (highlighted in purple). Retrieving Sequences by Gene IDs is shown for gene FOXG_16418. Additional retrieval features are available: *Retrieve Sequences By Genomic Sequence IDs*, *Retrieve Sequences By EST IDs*, *Retrieve Sequences By Popset Isolate IDs*, *Retrieve Sequences By Open Reading Frame IDs*; (**b**) *Help* menu at the bottom of the page offers guidance in identification and selection of transcriptional and translation start and stop sites; (**c**) Browser view of the results showing FASTA sequence for the genomic sequence for gene FOXG_16418.

9. Empowering Users to Improve Genome Annotation

FungiDB has nearly 100 genomes and counting. Several *Aspergillus* species (*A. fumigatus* Af293 and *A. nidulans* FGSC A4), *Cryptococcus* (*C. neoformans* JEC21 and H99), and *Neurospora crassa* are under active curation (via Chado relational database [48]). Other genomes, such as *A. niger*, *A. oryzae*, *C. immitis*, *C. posadasii*, *Ustilago maydis*, and *F. graminearum* are either already loaded or being loaded into the curation infrastructure, which is maintained in collaboration with the Wellcome Trust Sanger Institute.

User comments offer the fastest way to improve gene records and to alert the curation team of critical literature sources. We strongly encourage the community to offer their expertise to improve gene records. User comments about new findings or publications, or even negative results, help improve genome annotations and provide evidence for alternative gene models. To add user comments, click on the *Add a comment* link available on all gene record pages (Figure 21a). Figure 21b shows a user comment form where a user can enter descriptive information about gene function, upload a reference and files (e.g., images of protein tracking), and/or other gene identifiers. Several examples of user comments are shown in Table 2. All user comments become immediately visible on the gene pages, searchable via the text search, and can be modified at any time from a user account. To upload user comments in bulk, email *help@fungidb.org*.



AN5126 Microtubule binding protein HookA, putative

Name: HookA
Type: protein coding
Chromosome: ChrV
Location: ChrV_A_nidulans_FGSC_A4:1,226,392..1,229,104(+)

Species: Aspergillus nidulans
Strain: FGSC A4
Status: Curated Reference Strain

Add a comment to gene AN5126

Please add only scientific comments to be displayed on the gene page for AN5126. If you want to report a problem, use the support page.
Your comments are appreciated. If this is a **new gene**, please also add a comment in the corresponding Genome Sequence.

Part I: Comment

Headline *: HookA is required for dynein-mediated early endosome movement

Category (check all that apply): □ Gene Model □ Name/Product ☑ Function □ Expression □ Sequence ☑ Phenotype

Comment *: HookA is required for dynein-mediated early endosome motility. HookA mutant is not lethal. It exhibits a colony phenotype and defect in early endosome distribution.

Location: Strand: ● Forward - ○ Reverse
Genome Coordinates: 1226392-1229104

Part II: Evidence for This Comment (Optional)

Upload File: Select a file: No file selected.
Brief Description: (4000 max characters)

PubMed ID(s): 24637327
the article details of the PubMed ID(s) above

PMID: 24637327
Title: HookA is a novel dynein-early endosome linker critical for cargo movement in vivo.
Journal: J. Cell Biol. 2014 Mar 17;204(6):1009-26
Author: Zhang et al.

Digital Object Identifier (DOI) Name(s):

Genbank Accession(s):

Part III: Other Genes to which you want to apply this comment (Optional)

Gene Identifiers:

Figure 21. Improving gene records via user comments. (**a**) Shown is a gene record page for a putative microtubule-binding protein HookA AN5126 gene in *A. nidulans*. To add a comment, click on the *add a comment* link, which is located at the top of the gene record page (highlighted in purple); (**b**) User comment forms can be supplemented with scientific or experimental information (see comment examples in Table 2); files can be uploaded as well.

Table 2. Examples of user comments.

Comment type	Example Comment
Gene name, gene synonym	Cell division control protein Cdc48 is also known as dsc-6
Functional characterization	This is a transcription factor involved in the regulation of copper import
Subcellular localization	GFP tagging demonstrates that ham-5 co-localize with mak-2. Images attached
Phenotype	Deletion of rim101 has resulted in capsule defects and increased virulence in mice models

10. Conclusions

FungiDB is a service-oriented resource created with the primary goal of providing fungal researchers with access to data and with tools and functionality that enable them to ask their own questions. While it is critical that all datasets are loaded carefully, and quality controlled before public release, FungiDB strives to enable data mining to drive novel testable hypotheses rather than only recapitulating published conclusions. FungiDB benefits tremendously from being part of EuPathDB since resources can be more efficiently allocated and functionality can be broadly applied across databases. Future developments in FungiDB will include improved data mining tools, expanded workspaces for primary data analysis and integration, dedicated curatorial effort for supported genomes, and integration of new datatypes and system-wide analysis tools. FungiDB will continue to rely on the guidance of the fungal research community through advisory groups and by direct communication with members of the community at scientific meetings or through emails to *help@fungidb.org*.

Acknowledgments: The authors want to thank the members of the EuPathDB research communities and scientific advisors for sharing knowledge and experience, and for making intellectual contributions to improve the services offered via the FungiDB project. We also want to thank data contributors for providing datasets prepublication and other fungal databases for collaborating with FungiDB to benefit the fungal community. Many thanks are also extended to all past and present EuPathDB staff members who contributed and continue to diligently work on numerous aspects of the quickly expanding project. FungiDB is supported in part by NIH HHSN272201400030C and the Wellcome Trust grant WT108443MA to CHF. The project establishment has also benefited in the past from the Burroughs Wellcome Fund (BWF) and the Alfred P. Sloan Foundation to Jason E. Stajich, University of California, Riverside, CA USA.

Author Contributions: Evelina Y. Basenko conceived, designed, and led manuscript production; Evelina Y. Basenko, Jane A. Pulman, Achchuthan Shanmugasundram, David Starns, Omar S. Harb, Kathryn Crouch, Cristina Aurrecoechea, Susanne Warrenfeltz made individual contributions to the manuscript; Evelina Y. Basenko, Jane A. Pulman, Achchuthan Shanmugasundram, Omar S. Harb and David Starns worked to produce the final draft; Christian J. Stoeckert Jr., Jessica C. Kissinger, David S. Roos, Christiane Hertz-Fowler are PIs of the project at corresponding locations.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Stajich, J.E.; Harris, T.; Brunk, B.P.; Brestelli, J.; Fischer, S.; Harb, O.S.; Kissinger, J.C.; Li, W.; Nayak, V.; Pinney, D.F.; et al. FungiDB: An integrated functional genomics database for fungi. *Nucleic Acids Res.* **2012**, *40*. [CrossRef] [PubMed]
2. Aurrecoechea, C.; Barreto, A.; Basenko, E.Y.; Brestelli, J.; Brunk, B.P.; Cade, S.; Crouch, K.; Doherty, R.; Falke, D.; Fischer, S.; et al. EuPathDB: The eukaryotic pathogen genomics database resource. *Nucleic Acids Res.* **2017**, *45*, D581–D591. [CrossRef] [PubMed]
3. Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2016**, *44*, D67–D72. [CrossRef] [PubMed]
4. Yates, A.; Akanni, W.; Amode, M.R.; Barrell, D.; Billis, K.; Carvalho-Silva, D.; Cummins, C.; Clapham, P.; Fitzgerald, S.; Gil, L.; et al. Ensembl 2016. *Nucleic Acids Res.* **2016**, *44*, D710–D716. [CrossRef] [PubMed]

5. Nordberg, H.; Cantor, M.; Dusheyko, S.; Hua, S.; Poliakov, A.; Shabalov, I.; Smirnova, T.; Grigoriev, I.V.; Dubchak, I. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **2014**, *42*. [CrossRef] [PubMed]
6. Cerqueira, G.C.; Arnaud, M.B.; Inglis, D.O.; Skrzypek, M.S.; Binkley, G.; Simison, M.; Miyasato, S.R.; Binkley, J.; Orvis, J.; Shah, P.; et al. The *Aspergillus* Genome Database: Multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res.* **2014**, *42*, D705–D710.
7. Skrzypek, M.S.; Binkley, J.; Binkley, G.; Miyasato, S.R.; Simison, M.; Sherlock, G. The *Candida* Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.* **2017**, *45*, D592–D596. [CrossRef] [PubMed]
8. Cherry, J.M.; Hong, E.L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; Chan, E.T.; Christie, K.R.; Costanzo, M.C.; Dwight, S.S.; Engel, S.R.; et al. *Saccharomyces* Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*. [CrossRef] [PubMed]
9. McDowall, M.D.; Harris, M.A.; Lock, A.; Rutherford, K.; Staines, D.M.; Bähler, J.Ü.; Kersey, P.J.; Oliver, S.G.; Wood, V. PomBase 2015: Updates to the fission yeast database. *Nucleic Acids Res.* **2015**, *43*, D656–D661. [CrossRef] [PubMed]
10. Cuomo, C.A.; Birren, B.W. The fungal genome initiative and lessons learned from genome sequencing. *Methods Enzymol.* **2010**, *470*, 833–855. [CrossRef] [PubMed]
11. Ruiz-Roldán, C.; Pareja-Jaime, Y.; González-Reyes, J.A.; G-Roncero, M.I. The transcription factor Con7-1 is a master regulator of morphogenesis and virulence in *Fusarium oxysporum*. *Mol. Plant Microbe Interact.* **2015**, *28*, 55–68. [CrossRef] [PubMed]
12. Maglott, D.; Ostell, J.; Pruitt, K.D.; Tatusova, T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **2011**, *39*. [CrossRef] [PubMed]
13. McCluskey, K.; Wiest, A.; Plamann, M. The fungal genetics stock center: A repository for 50 years of fungal genetics research. *J. Biosci.* **2010**, *35*, 119–126. [CrossRef] [PubMed]
14. Apweiler, R. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef] [PubMed]
15. Winnenburg, R. PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res.* **2006**, *34*, D459–D464. [CrossRef] [PubMed]
16. Huerta-Cepas, J.; Capella-Gutiérrez, S.; Pryszcz, L.P.; Marcet-Houben, M.; Gabaldón, T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **2014**, *42*. [CrossRef] [PubMed]
17. Baldwin, T.T.; Basenko, E.; Harb, O.; Brown, N.A.; Urban, M.; Hammond-Kosack, K.E.; Bregitzer, P.P. Sharing mutants and experimental information prepublication using FgMutantDb (https://scabusa.org/FgMutantDb). *Fungal Genet. Biol.* **2018**.
18. Perkins, D.D.; Radford, A.; Sachs, M.S. *The Neurospora Compendium*; Academic Press: Cambridge, MA, USA, 2000; ISBN 0-12-550751-8.
19. Dewey, C.N. Aligning multiple whole genomes with mercator and MAVID. *Methods Mol. Biol.* **2007**, *395*, 221–235. [CrossRef] [PubMed]
20. Coradetti, S.T.; Xiong, Y.; Glass, N.L. Analysis of a conserved cellulase transcriptional regulator reveals inducer-independent production of cellulolytic enzymes in *Neurospora crassa*. *Microbiologyopen* **2013**, *2*, 595–609. [CrossRef] [PubMed]
21. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [CrossRef] [PubMed]
22. Caspi, R.; Billington, R.; Ferrer, L.; Foerster, H.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L.A.; et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2016**, *44*, D471–D480. [CrossRef] [PubMed]
23. Suh, M.J.; Fedorova, N.D.; Cagas, S.E.; Hastings, S.; Fleischmann, R.D.; Peterson, S.N.; Perlin, D.S.; Nierman, W.C.; Pieper, R.; Momany, M. Development stage-specific proteomic profiling uncovers small, lineage specific proteins most abundant in the *Aspergillus fumigatus* conidial proteome. *Proteome Sci.* **2012**, *10*. [CrossRef] [PubMed]

24. Chen, Y.; Toffaletti, D.L.; Tenor, J.L.; Litvintseva, A.P.; Fang, C.; Mitchell, T.G.; McDonald, T.R.; Nielsen, K.; Boulware, D.R.; Bicanic, T.; et al. The *Cryptococcus neoformans* transcriptome at the site of human meningitis. *MBio* **2014**, *5*. [CrossRef] [PubMed]
25. Brown, G.D.; Denning, D.W.; Gow, N.A.R.; Levitz, S.M.; Netea, M.G.; White, T.C. Hidden killers: Human fungal infections. *Sci. Transl. Med.* **2012**, 4. [CrossRef] [PubMed]
26. Molloy, S.F.; Chiller, T.; Greene, G.S.; Burry, J.; Govender, N.P.; Kanyama, C.; Mfinanga, S.; Lesikari, S.; Mapoure, Y.N.; Kouanfack, C.; et al. Cryptococcal meningitis: A neglected NTD? *PLoS Negl. Trop. Dis.* **2017**, *11*. [CrossRef] [PubMed]
27. 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. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef] [PubMed]
28. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Res.* **2017**, *45*, D331–D338. [CrossRef]
29. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **1995**, *57*, 289–300.
30. Simes, R.J. An improved bonferroni procedure for multiple tests of significance. *Biometrika* **1986**, *73*, 751–754. [CrossRef]
31. Kolde, R.; Vilo, J. GOsummaries: an R Package for visual functional annotation of experimental data. *F1000Research* **2015**. [CrossRef] [PubMed]
32. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* **2011**, *6*. [CrossRef] [PubMed]
33. De Sousa Lima, P.; Casaletti, L.; Bailão, A.M.; de Vasconcelos, A.T.R.; da Rocha Fernandes, G.; de Almeida Soares, C.M. Transcriptional and proteomic responses to carbon starvation in *Paracoccidioides*. *PLoS Negl. Trop. Dis.* **2014**, *8*. [CrossRef] [PubMed]
34. Sharpton, T.J.; Stajich, J.E.; Rounsley, S.D.; Gardner, M.J.; Wortman, J.R.; Jordar, V.S.; Maiti, R.; Kodira, C.D.; Neafsey, D.E.; Zeng, Q.; et al. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res.* **2009**, *19*, 1722–1731. [CrossRef] [PubMed]
35. Noble, J.A.; Nelson, R.G.; Fufaa, G.D.; Kang, P.; Shafir, S.C.; Galgiani, J.N. Effect of geography on the analysis of coccidioidomycosis-associated deaths, United States. *Emerg. Infect. Dis.* **2016**, *22*, 1821–1823. [CrossRef] [PubMed]
36. Engelthaler, D.M.; Roe, C.C.; Hepp, C.M.; Teixeira, M.; Driebe, E.M.; Schupp, J.M.; Gade, L.; Waddell, V.; Komatsu, K.; Arathoon, E.; et al. Local population structure and patterns of Western Hemisphere dispersal for *Coccidioides* spp., the fungal cause of Valley fever. *MBio* **2016**, *7*. [CrossRef] [PubMed]
37. Godfrey, D.; Böhlenius, H.; Pedersen, C.; Zhang, Z.; Emmersen, J.; Thordal-Christensen, H. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genom.* **2010**, *11*. [CrossRef] [PubMed]
38. Afgan, E.; Baker, D.; van den Beek, M.; Blankenberg, D.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Eberhard, C.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **2016**, *44*, W3–W10. [CrossRef] [PubMed]
39. Madduri, R.K.; Sulakhe, D.; Lacinski, L.; Liu, B.; Rodriguez, A.; Chard, K.; Dave, U.J.; Foster, I.T. Experiences building Globus Genomics: A next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. *Concurr. Comput. Pract. Exp.* **2014**, *26*, 2266–2279. [CrossRef] [PubMed]
40. Open Wheat Blast. Available online: http://s620715531.websitehome.co.uk/owb/ (accessed on 7 March 2018).
41. Islam, M.T.; Croll, D.; Gladieux, P.; Soanes, D.M.; Persoons, A.; Bhattacharjee, P.; Hossain, M.S.; Gupta, D.R.; Rahman, M.M.; Mahboob, M.G.; et al. Emergence of wheat blast in Bangladesh was caused by a South American lineage of *Magnaporthe oryzae*. *BMC Biol.* **2016**, *14*. [CrossRef] [PubMed]
42. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Institute: Cambridge, UK, 2017.
43. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
44. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25*, 1105–1111. [CrossRef] [PubMed]
45. Anders, S.; Pyl, P.T.; Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169. [CrossRef] [PubMed]

46. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*. [CrossRef] [PubMed]
47. Mathioni, S.M.; Beló, A.; Rizzo, C.J.; Dean, R.A.; Donofrio, N.M. Transcriptome profiling of the rice blast fungus during invasive plant infection and *in vitro* stresses. *BMC Genom.* **2011**, *12*. [CrossRef] [PubMed]
48. Zhou, P.; Emmert, D.; Zhang, P. Using Chado to Store Genome Annotation Data. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2002; ISBN 9780471250951.