

# Fungal Pathogen Genomics 2 – 6 June 2025 course timetable

Monday, 2<sup>nd</sup> of June

Time (BST)	Content
13:00 – 13:30	Meet & greet
13:30 – 15:45	<b>1. Database queries</b>  FungiDB site search - p. 6 Ensembl Fungi Molecular Interactions - p. 9 SGD Search Strategies – p. 23
15:45 – 16:00	Break
16:00 – 18:00	<b>2. Transcriptomics &amp; Proteomics</b>  Ensembl Fungi Track Hubs – p. 32 FungiDB Transcriptomic & Proteomic analysis– p. 40 SGD Expression tools - SPELL – p. 46
18:00 – 18:30	Q & A session

**Tuesday, 3<sup>rd</sup> of June**

Time (BST)	Content
13:00 – 13:30	Q & A session
13:30 – 15:45	<b>3. Enrichment analysis</b>  <a href="#">SGD GO Slim mapper</a> – p. 50 <a href="#">CGD GO Term finder</a> – p. 53 <a href="#">FungiDB GO enrichment</a> – p. 58
15:45 – 16:00	Break
16:00 – 17:00	Research Seminar: Prof Jason Stajich
17:00 – 18:00	<b>4. SNPs &amp; Variants</b>  <a href="#">SGD Variant viewer</a> – p. 68 <a href="#">FungiDB SNP analysis &amp; CNVs</a> – p. 72 <a href="#">Exploring variants in Ensembl Fungi</a> – p. 90 <a href="#">VEP</a> – p. 96
18:00 – 18:30	Q & A session

**Wednesday, 4<sup>th</sup> of June**

Time (BST)	Content
13:00 – 13:30	Q & A session
13:30 – 15:45	<b>5. Comparative Genomics &amp; Orthology and Evolutionary analysis &amp; cross-species inference</b>  <a href="#">Ensembl Fungi – WGA – p. 105</a> <a href="#">MycoCosm CAZy enzymes – p. 114</a> <a href="#">MycoCosm Synteny – p. 123</a> <a href="#">SGD predicting fungal biology – p. 128</a> <a href="#">Ensembl Fungi Evolutionary analysis (gene trees) – p. 136</a> <a href="#">FungiDB &amp; OrthoMCL: Orthology and Phyletic Patterns – p. 152</a>
15:45 – 16:00	Break
16:00 – 17:55	<b>6. Functional analysis: Pathways &amp; metabolites</b>  <a href="#">MycoCosm KEGG Browser &amp; Secondary metabolism clusters – p. 158</a> <a href="#">FungiDB pathways &amp; metabolites – p. 163</a>
17:55 – 18:00	<b>Introduction to Group Projects</b>
18:00 – 18:30	Q & A session

**Thursday 5<sup>th</sup> of June**

<b>Time (BST)</b>	<b>Content</b>
13:00 – 13:30	Q & A session
13:30 – 15:45	<b>7. My Data sets</b>  <a href="#">My workspace in FungiDB – p. 166</a> <a href="#">Exploring gene models – p. 180</a>
15:45 – 16:00	Break
16:00 – 18:00	<b>8. Group Projects</b>
18:00 – 18:30	Q & A session

**Friday, 6<sup>th</sup> of June**

Time (BST)	Content
13:00 – 13:30	Q & A session
13:30 – 15:45	Group Projects
15:45 – 16:00	<b>Break</b>
16:00 – 18:00	<b>Group Project Presentations</b>
18:00– 18:30	<b>Closing remarks</b>

# FungiDB Site Search

## Learning objectives:

- Incorporate keywords into the site search.
- Narrow site search results by categories, organisms, and other criteria.
- Export results to a search strategy.
- Locate genes using gene IDs.

The site search can be accessed from the header of the site and is available from every page. The site search queries the database for a term (e.g., text) or a specific ID and returns a list of pages and documents that contain the query term.

## Site search: text, term or gene id.

- Enter the word **kinase** in the site search window (at the top centre of the page). Click on the "enter" key on your keyboard or on the search icon as shown in the screenshot below.



- How many results with the word kinase did you get? Are all these records genes?
- Explore the filter panel on the left side of the page. Filter the results to view gene results only (hint: click on the word **Genes** in the “Filter results” section):

All results matching **kinase**

Export as a Search Strategy  
to download or mine your results ▶

1 - 20 of 394,386

Filter results

Genome  
Genes → 385,833  
Population biology  
Popset isolate sequences 7,994  
Metabolism  
Metabolic pathways  
Compounds 352  
Data access  
Data sets  
Searches 9  
About 4  
News 1

Hide zero counts

Data set - Analysis of the protein kinase A-regulated proteome of Cryptococcus neoformans

▶ Fields matched: Associated publications; Description; Name

Gene - CGB\_J0230W MAP kinase kinase kinase, MAP kinase kinase kinase, putative

Gene type: protein coding gene

Organism: Cryptococcus gattii WM276

▶ Fields matched: EC descriptions and numbers; GO terms; InterPro domains; Orthologs; PDB chains; Product description; Product descriptions (all)

Gene - A9K55\_006619 MAP kinase kinase kinase

Gene type: protein coding gene

Organism: Cordyceps militaris ATCC 34164

▶ Fields matched: EC descriptions and numbers; GO terms; InterPro domains; Orthologs; PDB chains; Product description; Product descriptions (all)

A screenshot of the search results page for "kinase". The page title is "All results matching kinase". At the top right is a blue button labeled "Export as a Search Strategy to download or mine your results". Below the title, it says "1 - 20 of 394,386". On the left, there is a "Filter results" sidebar with categories like Genome, Population biology, Metabolism, Data access, and About. The "Genes" category is selected and highlighted with a red arrow. The main content area shows a table of results. The first result is "Data set - Analysis of the protein kinase A-regulated proteome of Cryptococcus neoformans". It lists fields matched: Associated publications, Description, and Name. Below this is a "Gene" entry: "Gene - CGB\_J0230W MAP kinase kinase kinase, MAP kinase kinase kinase, putative". It details the Gene type as protein coding gene and the Organism as Cryptococcus gattii WM276. Another "Gene" entry follows: "Gene - A9K55\_006619 MAP kinase kinase kinase", also with protein coding gene type and Cordyceps militaris ATCC 34164 as the organism. Both entries have a note about EC descriptions and numbers, GO terms, InterPro domains, Orthologs, PDB chains, Product description, and Product descriptions (all).

Notice that clicking on the “Genes” category reveals additional filtering options (on the left) and activates the “Export as a Search Strategy” button on the top right, which is now shown in dark blue color. This is because the search strategy can be deployed on a single category only (e.g. Genes or Data sets, but not both).

- Select and apply the “Product descriptions (all)” filter.

Note: The applied filter can be easily cleared by clicking on “Clear filter” option as shown in the screenshot below.

- In the “Filter organisms” section, select to filter gene results by ***Malassezia restricta*** KCTC 27527. How many genes contain “kinase” in the product description field in this organism?

- Export the results to a search strategy.

To achieve this, click on the blue button called “Export as a search strategy...” at the top right-hand side of the results page.

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description
MRET_0047	MRET_0047-146..1	Malassezia restricta KCTC 27527	CP030251:95,680..97,545(–)	triose/dihydroxyacetone kinase/FAD-AMP lyase (cyclizing)
MRET_0094	MRET_0094-146..1	Malassezia restricta KCTC 27527	CP030251:170,464..171,498(+) adenosine kinase	
MRET_0099	MRET_0099-146..1	Malassezia restricta KCTC 27527	CP030251:179,095..181,227(+) aarF domain kinase	
MRET_0136	MRET_0136-146..1	Malassezia restricta KCTC 27527	CP030251:231,306..233,297(+) nucleoside-diphosphate kinase	
MRET_0167	MRET_0167-146..1	Malassezia restricta KCTC 27527	CP030251:270,552..272,270(+) pseudouridylate synthase/pseudouridine kinase	
MRET_0178	MRET_0178-146..1	Malassezia restricta KCTC 27527	CP030251:288,959..289,339(+) meiosis induction protein kinase IME2/SME1	
MRET_0205	MRET_0205-146..1	Malassezia restricta KCTC 27527	CP030251:331,297..333,045(+) tyrosine-protein kinase srms	
				type II pantothenate kinase

- Try running the same search but this time use a wild card (\*) (e.g., kinase\*).

When the wild card is combined with a word (**kinase\*** or **\*kinase**), the search will retrieve compound words ending or beginning with the word kinase (e.g. **\*kinase - phosphofructokinase**). The wild card (\*) can be used alone to retrieve all records available to the site search (see screenshot below).

All results matching \*

1 - 20 of 4,901,548

Hide zero counts

**Filter results**

Genome	1,885,291
Genes	162,441
Genomic sequences	
Organism	186
Organisms	
Transcriptomics	
ESTs	1,709,817
Population biology	
Popular isolate sequences	1,073,720
Metabolism	
Metabolic pathways	3,045
Compounds	61,998
Data access	
Data sets	381
Searches	435
Instructional	
Tutorials	15
Workshop exercises	1
About	
News	2
General info pages	16

**Filter fields**  
Select a result filter above

**Filter organisms**  
select all | clear all | expand all | collapse all  
 Type a taxonomic name

Hide zero counts

Compound - CHEBI:10000 Vismidine D

Compound - CHEBI:10001 Visnathin

Compound - CHEBI:10002 Visnagin

Compound - CHEBI:10003 ribostamycin sulfate

Definition: An aminoglycoside sulfate salt resulting from the reaction of ribostamycin with sulfuric acid.

Compound - CHEBI:10014 nalidixic acid

Definition: A monocarboxylic acid comprising 1,8-naphthyridin-4-one substituted by carboxylic acid, ethyl and methyl groups at positions 3, 1, and 7, respectively.

Compound - CHEBI:10014 vocamine

Compound - CHEBI:10015 vobasine

Definition: An indole alkaloid that is vobasan in which the bridgehead methyl group is substituted by a methoxycarbonyl group and an additional oxo substituent is present in the 3-position.

Compound - CHEBI:10016 volbutinine

Compound - CHEBI:10017 volmentol

Definition: A heptitol that is heptane-1,2,3,4,5,6,7-heptol that has R-configuration at positions 2, 3, and 6.

Compound - CHEBI:10018 volkenine

Definition: A cyanogenic glycoside that is (4R)-4-hydroxycyclopent-2-ene-1-carbonitrile attached to a beta-D-glucopyranosyloxy at position 1.

Compound - CHEBI:10019 vomericine

Compound - CHEBI:10022 vomitoxin

Compound - CHEBI:10023 voriconazole

Definition: A triazole-based antifungal agent used for the treatment of esophageal candidiasis, invasive pulmonary aspergillosis, and serious fungal infections caused by *Scedosporium apiospermum* and *Fusarium* spp. It is an inhibitor of cytochrome P450 2C9 (CYP2C9) and CYP3A4.

Compound - CHEBI:100241 ciprofloxacin

Definition: A quinolone that is quinolin-4(1H)-one bearing cyclopropyl, carboxylic acid, fluoro and piperazin-1-yl substituents at positions 1, 3, 6 and 7, respectively.

Export as a Search Strategy ► to download or mine your results

COMMUNITY CHAT

- The site search also works with gene ids. Run a site search for the following gene id: Afu2g13260

The gene id search will return the gene record card for [Afu2g13260](#).

Genes matching Afu2g13260

1 - 1 of 1

Hide zero counts

**Filter results**

Genome	1
Genes	

**Filter Gene fields**  
select all | clear all  
 External links  
 Gene ID  
 Names, IDs, and aliases  
 User comments

**Filter organisms**  
select all | clear all | expand all | collapse all  
 Type a taxonomic name  
 Fungi  
 Ascomycota

Hide zero counts

**Gene - Afu2g13260 Developmental regulator medA, putative**

Gene name or symbol: medA  
 Gene type: protein coding gene  
 Organism: *Aspergillus fumigatus* Af293

Fields matched: External links; Gene ID; Names, IDs, and aliases; User comments

**Gene - Afu2g13260 Developmental regulator medA, putative**

Gene name or symbol: medA  
 Gene type: protein coding gene  
 Organism: *Aspergillus fumigatus* Af293

Fields matched: External links; Gene ID; Names, IDs, and aliases; User comments

1 - 1 of 1

Export as a Search Strategy ► to download or mine your results

Clicking on the gene link in blue within the card will bring up the gene record page for this gene.

Clicking on the “Export as a Search Strategy” button will create a search strategy with a single gene ID. This may be useful if you are interested in cross-referencing different types of data for one gene.

Search strategy links:

**kinase** - <https://fungidb.org/fungidb/app/workspace/strategies/import/9c47e36cfa7790f6>

**kinase\*** - <https://fungidb.org/fungidb/app/workspace/strategies/import/eee9e7d2dfb3e7c1>

**Afu2g13260** -

<https://fungidb.org/fungidb/app/workspace/strategies/import/6fc6b7e52a15b76b>

## Exercise: Exploring host-pathogen interactions in Ensembl Fungi

Links to be clicked shown in blue, text to be entered shown in red.

*Zymoseptoria tritici* (also known as *Septoria tritici* and *Mycosphaerella graminicola*) is a fungal pathogen that causes septoria leaf blotch disease in *Triticum aestivum* (wheat). This fungus is considered a major threat to wheat production worldwide, and its ability to rapidly adapt to fungicides and host plants makes it a significant challenge for disease management.

You can explore molecular interactions of genes in Ensembl Fungi, ranging from pathogen-host interactions to symbiotic relationships across microbes and other Ensembl species.

**Step 1:** Find all genes involved in molecular interactions for *Zymoseptoria tritici*.

From the Ensembl Interactions REST API page <https://interactions.rest.ensembl.org>, search for all *Zymoseptoria tritici* genes that have a pathogenic interaction with *Triticum aestivum* (wheat).

Enter [https://interactions.rest.ensembl.org/interactions\\_by\\_prodbname/](https://interactions.rest.ensembl.org/interactions_by_prodbname/) into your browser and expand the GET /interaction documentation by clicking on **interaction\_list**. This opens a description and all available parameters for the endpoint. Click on **Try it out** to start your REST API request.

The screenshot shows the Ensembl Interactions REST API documentation for the GET /interaction endpoint. The URL is <https://interactions.rest.ensembl.org/interaction>. The page includes a navigation bar with 'interaction' and 'GET /interaction'. A callout box points to the 'interaction\_list' link in the top navigation bar. Another callout box points to the 'Object returned' section on the right. A large callout box points to the 'Build a curl query with these values' section in the main content area. A final callout box points to the 'Click on Try it out to expand search fields (Parameters) underneath' link at the bottom right.

Scroll down to the ‘Parameters’ section and fill in the query fields as follows:

species\_A: *Zymoseptoria tritici*

species\_B: *Triticum aestivum*

meta\_key: disease

## Database Queries: Ensembl Fungi (molecular interactions)

Click on **Execute** to submit your request.

The screenshot shows a 'Parameters' form with various input fields for querying molecular interactions. The fields include:

- interaction\_id (query) - interaction\_id
- interactor\_name (string) - interactor\_name
- ensembl\_gene (string) - ensembl\_gene
- species\_A (string) - Zymoseptoria tritici
- species\_B (string) - Triticum aestivum
- source\_db (string) - source db
- meta\_value (string) - meta\_value
- meta\_key (string) - disease

A callout box points to the 'species\_A' field with the text "Enter your parameters into the query fields". Another callout box points to the 'Execute' button at the bottom with the text "Click on Execute to submit your query".

Scroll down to 'Responses' to view your output.

The 'Responses' section displays the following information:

- Curl command:

```
curl -X 'GET' \
'https://interactions.rest.ensembl.org/interaction?species_A=Zymoseptoria%20tritici&species_B=Triticum%20aestivum&meta_key=disease' \
-H 'accept: application/json'
```
- Request URL:

```
https://interactions.rest.ensembl.org/interaction?species_A=Zymoseptoria%20tritici&species_B=Triticum%20aestivum&meta_key=disease
```
- Server response:

Code: 200 Response body

```
[{"interaction_id": 18174, "interactor_1": "uniprot:FBWR1", "interactor_2": "UNDETERMINED_PHI:4966_Triticum_aestivum", "ensembl_gene_1": "Mycgr3G53658", "ensembl_gene_2": "UNDETERMINED_PHI:4966_Triticum_aestivum", "species_1": "Zymoseptoria tritici", "species_2": "Triticum aestivum", "doi": "26032798", "source_db": "PHI-base"}, {"interaction_id": 18254, "interactor_1": "uniprot:ENK969", "interactor_2": "UNDETERMINED_PHI:2442_Triticum_aestivum", "ensembl_gene_1": "Mycgr3G53651", "ensembl_gene_2": "UNDETERMINED_PHI:2442_Triticum_aestivum", "species_1": "Zymoseptoria tritici", "species_2": "Triticum aestivum", "doi": "19522561", "source_db": "PHI-base"}, {"interaction_id": 18255, "interactor_1": "uniprot:ENK969", "interactor_2": "UNDETERMINED_PHI:2442_Triticum_aestivum", "ensembl_gene_1": "Mycgr3G53651", "ensembl_gene_2": "UNDETERMINED_PHI:2442_Triticum_aestivum", "species_1": "Zymoseptoria tritici", "species_2": "Triticum aestivum", "doi": "19522561", "source_db": "PHI-base"}]
```

You can copy the Request URL to obtain the results programmatically
- Response headers:

```
allow: GET, HEAD, OPTIONS  
content-length: 10111  
content-type: application/json  
date: Tue, 07 May 2024 16:49:19 GMT  
referrer-policy: same-origin
```
- Download options: "Download your results" and "Download"

Here, you can obtain the Curl script and request URL to access the same results programmatically.

Under 'Server response', you should get the following output:

```
zymoseptoria_tritici": ["Mycgr3G53658", "Mycgr3g88451",
"Ymcgr3G85040", "Mycgr3G40048", "Mycgr3G111221",
"Ymcgr3G103264", "Mycgr3G89160", "Mycgr3G80707",
"Mycgr3G65552", "Mycgr3g105487", "Mycgr3G70181",
"Ymcgr3G46840", "Mycgr3G93828", "Mycgr3G31676",
"Ymcgr3G51018", "Mycgr3G36951", "Mycgr3G77528",
"Ymcgr3G39611", "Mycgr3G96592", "Mycgr3G86705",
"Ymcgr3G107320", "Mycgr3G74194", "Mycgr3G87000",
"Ymcgr3G100355", "Mycgr3G92404", "Mycgr3G69942"]
```

**Step 2:** Let's find out more about the gene Mycgr3G65552 in the Ensembl Fungi browser.

**For the rest of the exercises, please use the archived version of EnsemblFungi 59**

On the [Ensembl Fungi 59 homepage](#), in the drop down menu “All genomes” select **Zymoseptoria tritici**. enter the gene ID **Mycgr3G65552** in the top right-hand corner and hit **Search**. Click on the gene ID **Mycgr3G65552** to open the ‘Gene’ tab.

The screenshot shows the Ensembl Fungi 59 Results Summary page. At the top, there is a navigation bar with links for BioMart, Downloads, Documentation, and Website help. A search bar on the right contains the query 'Mycgr3G65552'. Below the navigation bar, the species 'Zymoseptoria tritici (MG2)' is selected. On the left, a sidebar has buttons for New Search, Result Summary (which is highlighted in grey), Configure this page, Custom tracks, Export data, and Share this page. The main content area is titled 'Results Summary' and displays a message: 'Your search for Mycgr3G65552 returned 1 hits.' It notes that the search is limited to 10 results per category and term. Below this, a 'Gene or Gene Product' section shows one entry: '1. Gene: Mycgr3G65552 [Region in detail] Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F9WWD1]'. The entire interface is presented in a dark-themed browser window.

To find a list of species with which this particular *Z. tritici* gene has molecular interactions with, click on **Molecular interactions** in the left-hand panel.

## Database Queries: Ensembl Fungi (molecular interactions)

**Gene tab**

**Gene: Mycgr3G65552**

Description: Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F9WWD1]

Location: Chromosome 1: 1,786,483-1,788,643 reverse strand.

About this gene: MG2:ACPE01000001.1

Transcripts: This gene has 1 transcript ([splice variant](#), [279 orthologues](#) and [4 paralogues](#)).

Show transcript table

**Summary**

UniProtKB: This gene has proteins that correspond to the following UniProtKB identifiers: [F9WWD1](#)

Gene type: Protein coding

Annotation method: Protein coding genes annotated by the [JGI](#)

Go to [Region in Detail](#) for more tracks and navigation options (e.g. zooming)

**Molecular interactions**

Click on Molecular interactions

Genomic track showing protein coding genes across a 22.16 kb region. Labeled genes include 8164 > protein coding, Mycgr3T89063 > protein coding, Mycgr3T65551 > protein coding, ACPE01000001.1 >, and Mycgr3T28682 protein coding.

From this page, we can see that *Z. tritici* is known to interact with *T. aestivum*.

**Gene tab**

**Gene: Mycgr3G65552**

Description: Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F9WWD1]

Location: Chromosome 1: 1,786,483-1,788,643 reverse strand.

About this gene: MG2:ACPE01000001.1

Transcripts: This gene has 1 transcript ([splice variant](#), [279 orthologues](#) and [4 paralogues](#)).

Show transcript table

**Molecular interactions**

Click on Show metadata to view more details

This species

Species	Gene ID	Interactor	Identifier
Zymoseptoria tritici	Mycgr3G65552	protein	uniprot:F9WWD1

Interacts with

Species	Gene ID	Interactor	Identifier	Source DB
Triticum aestivum	UNDETERMINED	protein	UNDETERMINED	PHI-base

List of species and genes that Mycgr3G65552 interacts with

**About Us**

- About us
- Contact us
- Citing Ensembl Genomes

**Get help**

- Using this website
- Documentation
- Adding custom tracks

**Our sister sites**

- Ensembl
- Ensembl Bacteria
- Ensembl Plants

**Follow us**

- Blog
- Twitter

Can you find the wheat gene ID that Mycgr3G65552 interacts with? Look at the **Interacts with** table. The gene ID is ‘UNDETERMINED’. This means a molecular interaction has been experimentally verified between Mycgr3G65552 and wheat, but the former gene has not been identified yet.

## Database Queries: Ensembl Fungi (molecular interactions)

**Interacts with**

[Show metadata](#)

Species	Gene ID	Interactor	Identifier	Source DB
Triticum aestivum	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>

Can you find out what the phenotype for this interaction is? Click on [Show metadata](#) at the top right-hand corner of the ‘Interacts with’ table. Based on PHI-base, the interaction is associated with ‘loss of pathogenicity’.

**Interacts with**

[Show metadata](#)

Species	Gene ID	Interactor	Identifier	Source DB
Triticum aestivum	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
<b>Experimental evidence</b>	gene complementation			
<b>Interaction type</b>	interspecies interaction			
<b>Interaction phenotype</b>	PHIPO:0000010			
<b>Disease name</b>	PHIDO:0000331			
<b>Pathogen protein modification</b>	gene deletion: full			
<b>PHI-base high level term</b>	Loss of pathogenicity			
<b>Pathogen experimental strain</b>	IPO323			
<b>Host experimental strain</b>	cv. Riband			

**Step 3:** Next, let’s find all fungal orthologues. There are several ways of doing this. One way is to go to [Fungal Compara: Orthologues](#) in the left-hand panel.

The screenshot shows the Ensembl Fungi interface for the species *Zymoseptoria tritici* (MG2). The main content area displays the gene details for **Gene: Mycgr3G65552**. The sidebar on the left contains a navigation tree with various links such as Gene-based displays, Summary, Splice variants, Transcript comparison, Gene alleles, Sequence, Secondary Structure, Gene families, Literature, Fungal Compara, Genomic alignments, Phylogenetic tree, Gene gain/loss tree, Orthologues, Paralogues, Pan-taxonomic Compara, Gene Tree, Orthologues, Ontologies, GO: Molecular function, GO: Cellular component, GO: Biological process, Pft: Phylipse identifier, Phenotypes, Genetic Variation, Variant table, Variant image, Structural variants, Gene expression, Pathway, Molecular interactions, Regular, External references, Supporting evidence, and ID History.

The 'Orthologues' section is highlighted in the sidebar. It includes a 'Download orthologues' button and a summary table. The table has columns for Species set, Show details, With 1:1 orthologues, With 1:many orthologues, With many:many orthologues, and Without orthologues. The data is as follows:

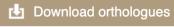
Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (1501 species)	<input checked="" type="checkbox"/>	139	91	11	1260
Acidomyces (2 species)	<input type="checkbox"/>	1	0	0	1
Agaricales (36 species)	<input type="checkbox"/>	0	19	0	17
Athellales (2 species)	<input type="checkbox"/>	0	1	1	0
Blastocladiidae (1 species)	<input type="checkbox"/>	0	0	0	1
Boletales (12 species)	<input type="checkbox"/>	0	9	0	3
Botryosphaerales (7 species)	<input type="checkbox"/>	2	0	0	5

## Database Queries: Ensembl Fungi (molecular interactions)

Can you find out if there are any orthologues in *Aspergillus fumigatus* with molecular interaction entries?

**Step 4:** You can hide the ‘Summary of orthologues of this gene’ table by clicking the **Hide** button. Enter *Aspergillus fumigatus* in the filter box on the top right-hand corner of the Orthologues table.

### Orthologues

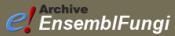
 Download orthologues

Summary of orthologues of this gene [Show !\[\]\(a5b836d7475c495160ae65b3a5563e89\_img.jpg\)](#)

Selected orthologues [Hide !\[\]\(986976fdffa5e534c8d04da003a09394\_img.jpg\)](#)

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Aspergillus fumigatus A1163	1-to-1	<a href="#">AFUB_089040</a>	54.37 %	42.25 %	n/a	n/a	Yes
		<a href="#">View Gene Tree</a>	DS499601:678,603-680,755:-1				
			<a href="#">View Sequence Alignments</a>				
Aspergillus fumigatus Af293	1-to-1	<a href="#">AFUA_7G02500</a>	54.37 %	42.25 %	n/a	n/a	Yes
		<a href="#">View Gene Tree</a>	7:680,932-683,084:-1				
			<a href="#">View Sequence Alignments</a>				

There are two orthologues in *A. fumigatus*. Click each of the gene IDs to find out which one has an entry under the **Molecular interactions** ‘Gene-based’ display. Molecular interactions are available for the second orthologue, [AFUA\\_7G02500](#).

 Archive BioMart | Downloads | Documentation | Website help

 Aspergillus fumigatus Af293 (ASM265v1) ▾

Location: 7:680,932-683,084 | Gene: AFUA\_7G02500 | Transcript: EAL84844

Search ...

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
  - Genomic
- Structure
  - Secondary Structure
  - Gene families
  - Literature
- Fungal Comparisons
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Pan-taxonomic Comparisons
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Molecular function
  - GO: Cellular component
  - GO: Biological process
  - PPI: Phibase identifier
- Phenotype
  - Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
  - Gene expression
  - Pathway
  - Molecular interactions
  - Regulation

**Gene: AFUA\_7G02500**

Description: polysaccharide synthase Cps1, putative

Location: Chromosome 7: 680,932-683,084 reverse strand. ASM265v1:CM000175.1

About this gene: This gene has 1 transcript ([splice variant](#)), 279 orthologues, 6 paralogues and is a member of 1 [Ensembl protein family](#).

Transcripts: [Show transcript table](#)

**Molecular interactions**: Cross-species interactions imported from PHI-base, HPIDB and PlasticDB with exact matches to proteins in Ensembl.

This species				Interacts with				Show metadata	
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB	
Aspergillus fumigatus	<a href="#">AFUA_7G02500</a>	protein	<a href="#">uniprot:Q4WAU2</a>	Mus musculus	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>	

Ensembl Fungi Archive release 59 - May 2024 © [EMBL-EBI](#)

---

**About Us**      **Get help**      **Our sister sites**      **Follow us**

[About us](#)      [Using this website](#)      [Ensembl](#)      [Blog](#)

[Contact us](#)      [Documentation](#)      [Ensembl Bacteria](#)      [Twitter](#)

What is the phenotype of the interaction for this orthologue with mice?

## Database Queries: Ensembl Fungi (molecular interactions)

Interacts with					<a href="#">Show metadata</a>
Species	Gene ID	Interactor	Identifier	Source DB	
Mus musculus	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>	
Several experiments exist for this interaction. Please click <a href="#">here</a> for more information					
Mus musculus	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>	
<b>Experimental evidence</b>	gene complementation				
<b>Interaction type</b>	interspecies interaction				
<b>Interaction phenotype</b>	PHIPO:0000015				
<b>Disease name</b>	PHIDO:0000020				
<b>Pathogen protein modification</b>	gene deletion: full				
<b>PHI-base high level term</b>	Reduced virulence				
<b>Pathogen experimental strain</b>	Af293				
<b>Host experimental strain</b>	C57BL/6				

The phenotype for the orthologue in mice is ‘reduced virulence’.

## Additional host-pathogen exercise 1 – Exploring GO terms and phenotypes

*Botrytis cinerea* is a necrotrophic fungus that infects a wide range of crops and ornamental plants, causing significant economic losses in agriculture and horticulture industries. It is known to cause botrytis bunch rot in various species. **Remember to use the Archived version of EnsemblFungi 59** to find out more information about molecular interactions in the species and answer the following questions:

- Using the [Ensembl Interactions REST API](#), can you retrieve all genes with molecular interaction information for *B. cinerea*?
- Open the ‘Molecular interactions’ page for the Bcin07g00720 gene in *B. cinerea*. What plant species does the gene interact with?
- Can you find the phenotype that is reported for each of the species the gene interacts with?
- Find all fungal orthologues. Is there any orthologue in *Magnaporthe oryzae* for Bcin07g00720? For which orthologue is molecular interaction information available?
- Which species does the *M. oryzae* orthologue interact with?
- Compare the molecular interaction phenotypes between the *B. cinerea* and *M. oryzae* orthologues. Can you find any common molecular functions that may explain this phenotype?

### Exercise 1 answers:

- Go to the Ensembl Interactions REST API and expand the [GET /interactions\\_by\\_prodid](#) endpoint documentation. Click on Try it out and then Execute.

The screenshot shows the 'interactions\_by\_prodid' endpoint documentation. The 'Parameters' section indicates 'No parameters'. At the bottom, there are 'Execute' and 'Clear' buttons.

In the ‘Response body’, search for **botrytis\_cinerea**. Alternatively, you can open the request URL [https://interactions.rest.ensembl.org/interactions\\_by\\_prodid](https://interactions.rest.ensembl.org/interactions_by_prodid) in your browser. You should get the following output:

```
"botrytis_cinerea": ["Bcin07g00720", "Bcin02g02570",
"Bcin12g04900", "Bcin16g00630", "Bcin02g06770",
"Bcin03g07190", "Bcin09g02390", "Bcin09g01800",
"Bcin07g03050", "Bcin08g05150", "Bcin10g01250"],
```

## Database Queries: Ensembl Fungi (molecular interactions)

```
"Bcin14g01870", "Bcin06g04870", "Bcin06g00240",
"Bcin06g03440", "Bcin03g07900", "Bcin03g06840",
"Bcin10g02530", "Bcin08g02990", "Bcin07g02610",
"Bcin03g08710", "Bcin10g05590", "Bcin16g01820",
"Bcin03g01540", "Bcin14g00650", "Bcin09g05460",
"Bcin10g02650", "Bcin02g02780", "Bcin05g03080",
"Bcin08g00160", "Bcin01g06010", "Bcin01g11360",
"Bcin15g00450", "Bcin03g04600", "Bcin09g01910",
"Bcin09g05050", "Bcin15g03580", "Bcin05g02590"]
```

- (b) Go to the [Ensembl Fungi homepage](#) and search for **Bcin07g00720**. In the results, click on the **Gene ID** to open the Gene tab.

The screenshot shows the Ensembl Fungi homepage with a search bar at the top. Below it, a sidebar on the left lists 'Search Ensembl Fungi' options like 'New Search', 'Configure this page', and 'Custom tracks'. The main content area displays 'Search results for 'Bcin07g00720'' with a summary table for the gene. The table includes columns for Description (n/a), Gene ID (Bcin07g00720), Species (Botrytis cinerea B05.10), and Location (7:260067-264879). At the bottom, it says 'Ensembl Fungi release 58 - January 2024 © EMBL-EBI'.

In the left-hand panel, click on **Molecular interactions** to open the page.

The screenshot shows the Ensembl Fungi gene page for Bpk3. The left sidebar has a 'Molecular interactions' section highlighted. The main content area shows a genomic track for the Bpk3 gene. A callout box points to the 'Molecular interactions' track with the instruction 'Click on Molecular interactions to open the page'. The track shows several protein coding regions across the genome coordinates 252,500 to 272,500.

## Database Queries: Ensembl Fungi (molecular interactions)

In the ‘Molecular interactions’ page, you can find all species the gene interacts with in the right-hand table. These include *Solanum lycopersicum* (tomato), *Vitis vinifera* (grape), *Cucumis sativus* (cucumber) and *Malus domestica* (apple).

**Molecular interactions** Cross-species interactions imported from PHI-base, HPIDB and PlasticDB with exact matches to proteins in Ensembl.

This species				Interacts with				Show metadata
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB
<i>Botrytis cinerea</i> B05.10	<a href="#">Bcin07g00720</a>	protein	<a href="#">uniprot:A6RYB8</a>	<i>Solanum lycopersicum</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<i>Vitis vinifera</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<i>Cucumis sativus</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<i>Malus domestica</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>

- (c) Click on **Show metadata** in the right-hand corner of the Interacts with table. You can find associated phenotypes under **PHI-base high level term**. The gene is associated with ‘Reduced virulence’ and ‘Loss of pathogenicity’.

**Molecular interactions** Cross-species interactions imported from PHI-base, HPIDB and PlasticDB with exact matches to proteins in Ensembl.

This species				Interacts with				Show metadata
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB
<i>Botrytis cinerea</i> B05.10	<a href="#">Bcin07g00720</a>	protein	<a href="#">uniprot:A6RYB8</a>	<i>Solanum lycopersicum</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	<a href="#">PHIPO:0000015</a>			
				<b>Disease name</b>	<a href="#">PHIDO:0000178</a>			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Reduced virulence			
				<b>Pathogen experimental strain</b>	B05.10			
				<i>Vitis vinifera</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	<a href="#">PHIPO:0000015</a>			
				<b>Disease name</b>	<a href="#">PHIDO:0000178</a>			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Reduced virulence			
				<b>Pathogen experimental strain</b>	B05.10			
				<i>Cucumis sativus</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	<a href="#">PHIPO:0000010</a>			
				<b>Disease name</b>	<a href="#">PHIDO:0000178</a>			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Loss of pathogenicity			
				<b>Pathogen experimental strain</b>	B05.10			
				<i>Malus domestica</i>	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	<a href="#">PHIPO:0000015</a>			
				<b>Disease name</b>	<a href="#">PHIDO:0000178</a>			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Reduced virulence			
				<b>Pathogen experimental strain</b>	B05.10			

- (d) To retrieve all fungal orthologues, go to **Fungal Compara: Orthologues** in the left-hand panel.

# Database Queries: Ensembl Fungi (molecular interactions)

**Gene-based displays**

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
- Secondary Structure
- Gene families
- Literature
- Fungal Comparisons
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues**
- Paralogues
- Pan-taxonomic Comparisons
- Gene Tree
- Orthologues
- Ontologies
- GO: Molecular function
- GO: Cellular component
- GO: Biological process
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- ID History

**Gene: Bpk3 Bcin07g00720**

**Gene Synonyms**  
Location

Chromosome 7: 260,067–264,879 forward strand.  
ASM83294v1:CP009811.1

About this gene  
This gene has 1 transcript (splice variant) and 313 orthologues.

Show transcript table

Download orthologues

**Summary of orthologues of this gene** Hide ⊖

Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'Configure this page' to choose a custom list of species.

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (1504 species)	<input checked="" type="checkbox"/>	279	16	0	<a href="#">1209</a>
Acidomyces (2 species)	<input type="checkbox"/>	1	0	0	<a href="#">1</a>
Agaricales (36 species)	<input type="checkbox"/>	4	0	0	<a href="#">32</a>
Athellales (2 species)	<input type="checkbox"/>	1	0	0	<a href="#">1</a>
Blastocladiiales (1 species)	<input type="checkbox"/>	0	1	0	0
Boletales (12 species)	<input type="checkbox"/>	3	0	0	<a href="#">9</a>

Scroll down to the Orthologues table and use the filter box in the top right-hand corner to search for *Magnaporthe oryzae*.

## Orthologues ?

[Download orthologues](#)

**Summary of orthologues of this gene** Show +

Selected orthologues Hide ⊖

Show All ▾ entries		Magnaporthe oryzae					
Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Magnaporthe oryzae	1-to-1	ATG1 ( <a href="#">MGG_06393</a> )	49.90 %	51.47 %	n/a	n/a	Yes
		<a href="#">View Gene Tree</a> 4:3,898,532-3,902,777:-1 <a href="#">View Sequence Alignments</a>					
Magnaporthe oryzae	1-to-1	M_BR32_EuGene_00042871	50.05 %	49.58 %	n/a	n/a	Yes
		<a href="#">View Gene Tree</a> BR32_scaffold00003:3,066,924-3,069,846:-1 <a href="#">View Sequence Alignments</a>					

Click on each of the orthologue gene IDs to open their respective gene tab and find out if the **Molecular interactions** Gene-based display is available. Molecular interaction information is available for the orthologue ATG1 ([MGG\\_06393](#)).

## Database Queries: Ensembl Fungi (molecular interactions)

**EnsemblFungi** ▾   HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog   [Login/Register](#)

**Magnaporthe oryzae (MG8) ▾**

Location: 4:3,898,532-3,902,777   Gene: ATG1   Transcript: MGG\_06393T0

**Gene-based displays**

- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
  - Gene families
  - Literature
- Fungal Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Molecular function
  - GO: Cellular component
  - GO: Biological process
  - PHI: Phibase identifier
  - Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
  - Gene expression
  - Pathway
  - Molecular interactions
  - Regulation
  - External references
  - Supporting evidence
- ID History
- Gene history

[Configure this page](#)

[Custom tracks](#)

**Gene: ATG1 MGG\_06393**

Description: Serine/threonine-protein kinase ATG1 [Source:UniProtKB/Swiss-Prot;Acc:Q52EB3]

Location: Chromosome 4: 3,898,532-3,902,777 reverse strand. MG8:CM001234.1

About this gene: This gene has 1 transcript ([splice variant](#)) and [313 orthologues](#).

Transcripts: [Show transcript table](#)

**Summary**

Name: ATG1 (UniProtKB Gene Name)

UniProtKB: This gene has proteins that correspond to the following UniProtKB identifiers: [Q52EB3](#)

Gene type: Protein coding

Annotation method: Protein coding genes annotation from the [Broad Institute](#).

The Molecular interactions link is available for ATG1 (MGG\_06393) in *Magnaporthe oryzae*

Zoom controls: Zoom in (e.g. zooming), Export image, Reset configuration, Reset track order, Drag/Select: Forward strand.

Genes: MGG\_06394T0 > protein coding, AACU030000115.1 >, < MGG\_06395T0 protein coding, < MGG\_06396T0 protein coding, < MGG\_06397T0 protein coding, < MGG\_06398T0 protein coding, < MGG\_06399T0 protein coding, < MGG\_06391T0 protein coding, < MGG\_06392T0 protein coding, < MGG\_06393T0 protein coding.

Contigs: MGG\_06394T0 > protein coding, AACU030000115.1 >, < MGG\_06395T0 protein coding, < MGG\_06396T0 protein coding, < MGG\_06397T0 protein coding, < MGG\_06398T0 protein coding, < MGG\_06391T0 protein coding, < MGG\_06392T0 protein coding, < MGG\_06393T0 protein coding.

- (e) Click on **Molecular interactions** in the left-hand panel. The ATG1 protein interacts with *Hordeum vulgare* (barley) and *Oryza sativa* (rice).

**Molecular interactions** Cross-species interactions imported from PHI-base, HPIDB and PlasticDB with exact matches to proteins in Ensembl.

This species: Magnaporthe oryzae 70-15

Species	Gene ID	Interactor	Identifier	Interacts with	Source DB																
Magnaporthe oryzae 70-15	<a href="#">MGG_06393</a>	protein	<a href="#">uniprot:Q52EB3</a>	<table border="1"> <thead> <tr> <th>Species</th> <th>Gene ID</th> <th>Interactor</th> <th>Identifier</th> </tr> </thead> <tbody> <tr> <td>Hordeum vulgare</td> <td>UNDETERMINED</td> <td>protein</td> <td>UNDETERMINED</td> </tr> <tr> <td>Oryza sativa</td> <td>UNDETERMINED</td> <td>protein</td> <td>UNDETERMINED</td> </tr> <tr> <td>Oryza sativa</td> <td>UNDETERMINED</td> <td>protein</td> <td>UNDETERMINED</td> </tr> </tbody> </table>	Species	Gene ID	Interactor	Identifier	Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED	Oryza sativa	UNDETERMINED	protein	UNDETERMINED	Oryza sativa	UNDETERMINED	protein	UNDETERMINED	<a href="#">Show metadata</a>
Species	Gene ID	Interactor	Identifier																		
Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED																		
Oryza sativa	UNDETERMINED	protein	UNDETERMINED																		
Oryza sativa	UNDETERMINED	protein	UNDETERMINED																		

- (f) Click on **Show metadata** to view the phenotypes associated with the molecular interactions. In *B. cinerea*, the phenotype are ‘Loss of pathogenicity’ and ‘Reduced virulence’

## Database Queries: Ensembl Fungi (molecular interactions)

### Molecular interactions Cross-species interactions imported from PHI-base, HPIDB and PlasticDB with exact matches to proteins in Ensembl.

This species				Interacts with				Show metadata
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB
Botrytis cinerea B05.10	<a href="#">Bcin07g00720</a>	protein	<a href="#">uniprot:A6RYB8</a>	Solanum lycopersicum	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	PHIPO:0000015			
				<b>Disease name</b>	PHIDO:0000178			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Reduced virulence			
				<b>Pathogen experimental strain</b>	B05.10			
				Vitis vinifera	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	PHIPO:0000015			
				<b>Disease name</b>	PHIDO:0000178			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Reduced virulence			
				<b>Pathogen experimental strain</b>	B05.10			
				Cucumis sativus	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	PHIPO:0000010			
				<b>Disease name</b>	PHIDO:0000178			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Loss of pathogenicity			
				<b>Pathogen experimental strain</b>	B05.10			
				Malus domestica	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	PHIPO:0000015			
				<b>Disease name</b>	PHIDO:0000178			
				<b>Pathogen protein modification</b>	gene mutation; gene complementation			
				<b>PHI-base high level term</b>	Reduced virulence			
				<b>Pathogen experimental strain</b>	B05.10			

In *M. oryzae* the phenotype is ‘Loss of pathogenicity’ only.

### Molecular interactions Cross-species interactions imported from PHI-base, HPIDB and PlasticDB with exact matches to proteins in Ensembl.

This species				Interacts with				Show metadata
Species	Gene ID	Interactor	Identifier	Species	Gene ID	Interactor	Identifier	Source DB
Magnaporthe oryzae 70-15	<a href="#">MGG_06393</a>	protein	<a href="#">uniprot:Q52EB3</a>	Hordeum vulgare	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				Several experiments exist for this interaction. Please click <a href="#">here</a> for more information				
				Oryza sativa	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				Several experiments exist for this interaction. Please click <a href="#">here</a> for more information				
				Oryza sativa	UNDETERMINED	protein	UNDETERMINED	<a href="#">PHI-base</a>
				<b>Interaction type</b>	interspecies interaction			
				<b>Interaction phenotype</b>	PHIPO:0000010			
				<b>Disease name</b>	PHIDO:0000315			
				<b>Pathogen protein modification</b>	gene deletion: full			
				<b>PHI-base high level term</b>	Loss of pathogenicity			
				<b>Pathogen experimental strain</b>	Guy11			
				<b>Host experimental strain</b>	cv. CO-39			

(g) Go to [Ontologies: GO: Molecular function](#) for both *B. cinerea* and *M. oryzae*. Comparing the GO terms for the two orthologues we can see that they have identical GO annotations:

- nucleotide binding
- protein kinase activity
- protein serine/threonine kinase activity

# Database Queries: Ensembl Fungi (molecular interactions)

- ATP binding
- kinase activity
- transferase activity
- protein serine kinase activity

**EnsemblFungi** ▾ HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

**Botrytis cinerea B05.10 (ASM83294v1) ▾**

Location: 7:260,067-264,879 Gene: Bpk3 Transcript: Bcin07g00720.1

**Gene-based displays**

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
- Secondary Structure
- Gene families
- Literature
- Fungal Compara
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues
- Paralogues
- Pan-taxonomic Compara
- Gene Tree
- Orthologues
- Ontologies
- GO: Cellular component
- GO: Molecular function**
- GO: Biological process
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- ID History
- Gene history

Configure this page

**Gene: Bpk3 Bcin07g00720**

**Gene Synonyms**

**Location**

Chromosome 7: 260,067-264,879 forward strand.  
ASM83294v1:CP009811.1

**About this gene**

This gene has 1 transcript (splice variant) and 313 orthologues.

**Transcripts**

Show transcript table

**GO: Molecular function**

Show/hide columns (1 hidden)						Filter
Accession	Term	Evidence	Annotation source	Transcript IDs		
GO:0000166	nucleotide binding	IEA	UniProt	Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0004672	protein kinase activity	IEA		Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0004674	protein serine/threonine kinase activity	IEA		Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0005524	ATP binding	IEA		Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0016301	kinase activity	IEA	UniProt	Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0016740	transferase activity	IEA	UniProt	Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0106310	protein serine kinase activity	IEA	RHEA	Bcin07g00720.1	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	

**EnsemblFungi** ▾ HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

**Magnaporthe oryzae (MG8) ▾**

Location: 4:3,898,532-3,902,777 Gene: ATG1 Transcript: MGG\_06393T0

**Gene-based displays**

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
- Secondary Structure
- Gene families
- Literature
- Fungal Compara
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues
- Paralogues
- Pan-taxonomic Compara
- Gene Tree
- Orthologues
- Ontologies
- GO: Cellular component
- GO: Molecular function**
- GO: Biological process
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- ID History
- Gene history

Configure this page

**Gene: ATG1 MGG\_06393T0**

**Description**

Serine/threonine-protein kinase ATG1 [Source:UniProtKB/Swiss-Prot;Acc:Q52EB3]

**Location**

Chromosome 4: 3,898,532-3,902,777 reverse strand.  
MG8:CM001234.1

**About this gene**

This gene has 1 transcript (splice variant) and 313 orthologues.

**Transcripts**

Show transcript table

**GO: Molecular function**

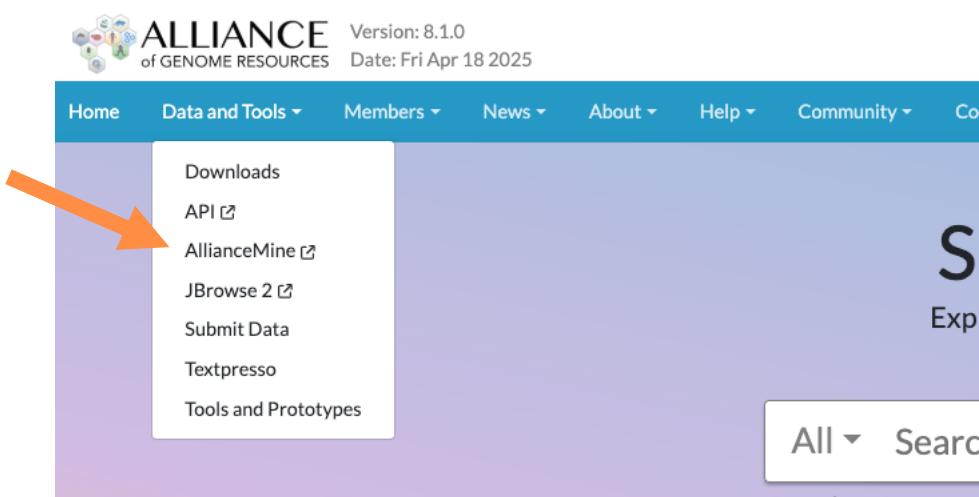
Show/hide columns (1 hidden)						Filter
Accession	Term	Evidence	Annotation source	Transcript IDs		
GO:0000166	nucleotide binding	IEA	UniProt	MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0004672	protein kinase activity	IEA	InterPro	MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0004674	protein serine/threonine kinase activity	IMP		MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0005524	ATP binding	IEA		MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0016301	kinase activity	IEA	UniProt	MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0016740	transferase activity	IEA	UniProt	MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	
GO:0106310	protein serine kinase activity	IEA	RHEA	MGG_06393T0	<ul style="list-style-type: none"> <li>• Search BioMart</li> <li>• View on karyotype</li> </ul>	

## Search Strategies in SGD

In addition to a faceted search tool, AllianceMine (within the Alliance of Genome Resources) is a means for SGD users to conduct more advanced queries. AllianceMine enables rapid retrieval and manipulation of curated biological data on *S. cerevisiae* genes and genomic features. By creating gene lists, users can retrieve data on multiple genes at once. Gene lists can then be continually modified, analyzed, and refined as desired, enabling you to answer complex biological questions such as, “How many plasma membrane proteins are required for viability?” or “Which kinases, if knocked out, increase chronological lifespan?”

In this exercise, we will use AllianceMine to search for as-yet undiscovered mitochondrial ribosomal proteins in yeast.

- Access AllianceMine from Alliance home page (<https://www.alliancegenome.org>); click on AllianceMine from the Data and Tools section at the top of the page.



### 1. Create a list of proteins that are known subunits of the mitochondrial ribosome (MTR):

The screenshot shows the 'Go by Most Popular Queries' section of the AllianceMine interface. It includes a navigation bar with tabs for GENOME, LITERATURE, DOWNLOADS, INTERACTIONS, PROTEINS, and PHENOTYPES. The GENOME tab is selected. Below the tabs is a list of popular queries:

- Chromosomal Region ➔ All genes
- Gene ➔ Non-Fungal and *S. cerevisiae* Homologs
- Gene ➔ Flanking features within a specific distance
- Feature Type ➔ Features of a selected feature Type
- All genes of a selected Feature Type ➔ Genes with introns
- Gene ➔ Chromosomal location
- Gene ➔ Genomic DNA
- Feature Type ➔ Genes
- Organism ➔ All genes
- Chromosome ➔ Genes of a selected Feature Type

- Click on "More queries here"

The screenshot shows a search interface with a navigation bar at the top. On the left is a 'Filter by text' section with a 'Search for keywords' input field. On the right is a 'Filter by category' section with tabs labeled 'All', 'Downloads', 'Interactions' (which is highlighted in red), 'Genome', 'Proteins', 'Function', 'Literature', 'Regulation', 'Homology', and 'Disease'. Below these tabs are smaller links for 'Phenotypes' and 'Expression'.

- And then select the **FUNCTION** tab and then **GO Slim Term => Gene**. Enter "mitochondrion" as your GO slim term. This will return many results. Go to the bottom of the results and click "**View ~39,000 rows.**"  
 [NOTE: if you get zero results, click the "Edit Query" button and use the trash can icons to remove 'F' and 'E' – the Annot Type filters. When these are removed, there are lots of results at the bottom of the page – you just need to wait for it to finish searching]

GO Slim Term ➔ Gene

Retrieve all genes that are annotated to the selected GO Slim term and children of that selected GO Slim Term. Only manually curated and high-throughput GO annotations are included.

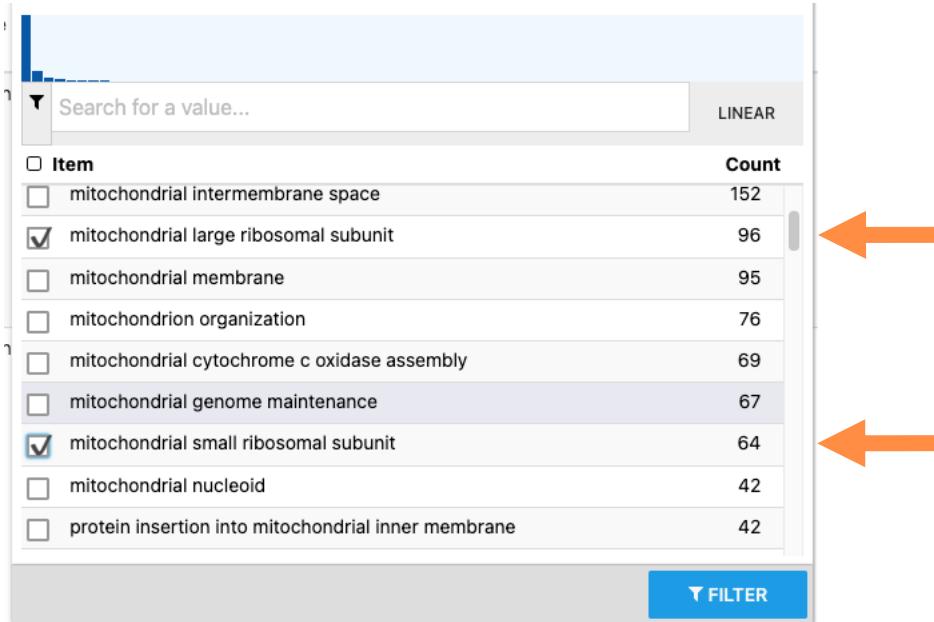
The screenshot shows a table titled 'Results Preview' with columns for Gene ID, Primary DBID, Systematic Name, Standard Name, Feature Type, Qualifier, GO Annotation > Ontology Term . Identifier, GO Annotation > Ontology Term . Namespace, GO Annotation > Ontology Term . Name, and Code > Code. The 'Gene > Primary DBID' column is highlighted with an orange arrow pointing to the input field. The table contains two rows of data:

Gene > Primary DBID	Gene > Systematic Name	Gene > Standard Name	Gene > Feature Type	Gene > Qualifier	GO Annotation > Ontology Term . Identifier	GO Annotation > Ontology Term . Namespace	GO Annotation > Ontology Term . Name	Code > Code
S000029023	YNCQ0027W	RPM1	ncRNA gene		GO:0005739	mitochondrion	cellular_component	IDA
S000029023	YNCQ0027W	RPM1	ncRNA gene		GO:0030678	mitochondrial ribonuclease P complex	cellular_component	IDA

- In the Query Results, first go to the **Gene Feature Type** column, click the filter icon and then select "**ORF**" from the drop-down menu and "Apply."

The screenshot shows a filter dialog for 'Gene Feature Type'. It has a green header bar with 'No active filters'. Below it is a dropdown menu with the value 'ORF' selected. At the bottom are 'ADD MORE' and 'APPLY' buttons.

- Next go to the "Ontology Term Name" column, hit the **graph icon**, and select the boxes for "mitochondrial large ribosomal subunit," "mitochondrial small ribosomal subunit," and "mitochondrial ribosome." Hit FILTER and you'll get 160 results.



Save this list by clicking the **Save List** button on the upper right of the table and selecting "**Genes (83)**" at the top of the pull-down. Give it the name "List 1 MTR genes" and save.

Save a list of 83 Genes

Name  
List 1 MTR genes

Optional attributes

Description  
Enter a description

CANCEL **SAVE**

## 2. Find proteins that genetically interact with MTR proteins:

- Scroll down to below the table and find the "Widgets" section, click "View All."

- The results table shows all genes/proteins with genetic or physical interactions with the MTR genes. There are over 17K of them.
- Go to the column for "Details Relationship Type" and filter for "Genetic." Hit Apply and you'll get a list of ~8500 rows.

### 3. Find MTR interactors that are uncharacterized:

- Go to the "Gene Standard Name" column and click the filter icon. For the purposes of this exercise, filter to include ONLY the gene **RML2**, which yields 615 rows.
- Go to the column "**Participant 2 Standard Name**" (these are the genes that interact with the MTR gene) and hit bottom arrow in the sort icon to get the "**No value**" items at the top of the list. This represents the potential uncharacterized MTR interactors.

Showing 1 to 250 of 718 rows

Rows per page: 250 ▾ Page 1 ►

Gene Systematic Name	Gene Standard Name	Organism Short Name	Details Name	Details Relationship Type	Details Role 1	Participant 2 Primary DBID	Participant 2 Standard Name	Experiment Name
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000035	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000035	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000132	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000639	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000686	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000923	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S00000923	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000001726	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000001902	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000001902	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000002852	NO VALUE	Costanzo M, et al. (2016)-27708008-Positive Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000003249	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000003354	NO VALUE	Costanzo M, et al. (2010)-20093466-Negative Genetic
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	S000003668	NO VALUE	Costanzo M, et al. (2016)-27708008-Negative Genetic

- Given the current state of bugginess in this tool (it's undergoing major revision), you'll do best to copy and paste the section of the list that includes just the rows that say "no value" in the Participant 2 column. Copy this section and paste it into a blank worksheet.
- In the worksheet, sort by **column G** (the "Participant 2 Primary DBID") and then do a quick de-dupe of the list to leave 18 potential uncharacterized interactors. Go back to the AllianceMine homepage and click the "**Analyse data**" button to create a new list. Paste your list of MTR interactors from your worksheet into the box. Save the list as **"List 3: Uncharacterized Interactors"**
- You might want to save this worksheet.



A	B	C	D	E	F	G	H	I	J
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000035	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000132	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000639	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000686	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000000923	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000001726	NO VALUE	Costanzo M. et al. (2016)-277080	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000001902	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000002852	NO VALUE	Costanzo M. et al. (2016)-277080	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003249	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S00000354	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003668	NO VALUE	Costanzo M. et al. (2016)-277080	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003699	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000003729	NO VALUE	Costanzo M. et al. (2016)-277080	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000004277	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000004991	NO VALUE	Costanzo M. et al. (2016)-277080	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000005039	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000005153	NO VALUE	Costanzo M. et al. (2010)-200934	
YEL050C	RML2	S. cerevisiae	NO VALUE	genetic	Bait	SGD:S000006009	NO VALUE	Costanzo M. et al. (2010)-200934	

- It would be a good idea to narrow down our candidates even more. For example, because the MTR is a mitochondrial complex, we would expect that deleting uncharacterized (but bona fide) subunits of the MTR would disrupt aerobic respiration. Let's refine our list of predicted MTR subunits by seeing which genes disrupt respiratory growth when deleted.

- Return to AllianceMine home page, click the "More Queries" option at the bottom and then choose the **PHENOTYPES** tab. **Then Retrieve all phenotypes for all genes.**"

The screenshot shows the AllianceMine search interface. At the top, there is a search bar labeled "Search for keywords" and a filter bar labeled "Filter by category" with tabs for All, Downloads, Interactions, Genome, Proteins, Function, Literature, Regulation, Homology, and Disease. The "Phenotypes" tab is highlighted with a dark gray background and white text. Below the filter bar, there is a message: "Retrieve all phenotypes for all genes." A blue "View >>" button is located below this message. To the right of the "View >>" button is a "Categories" section showing "Downloads" and "Phenotypes". Another orange arrow points to the "View >>" button. Further down the page, there is another section titled "Literature → Phenotype" with a "View >>" button and a "Categories" section showing "Downloads", "Literature", and "Phenotypes".

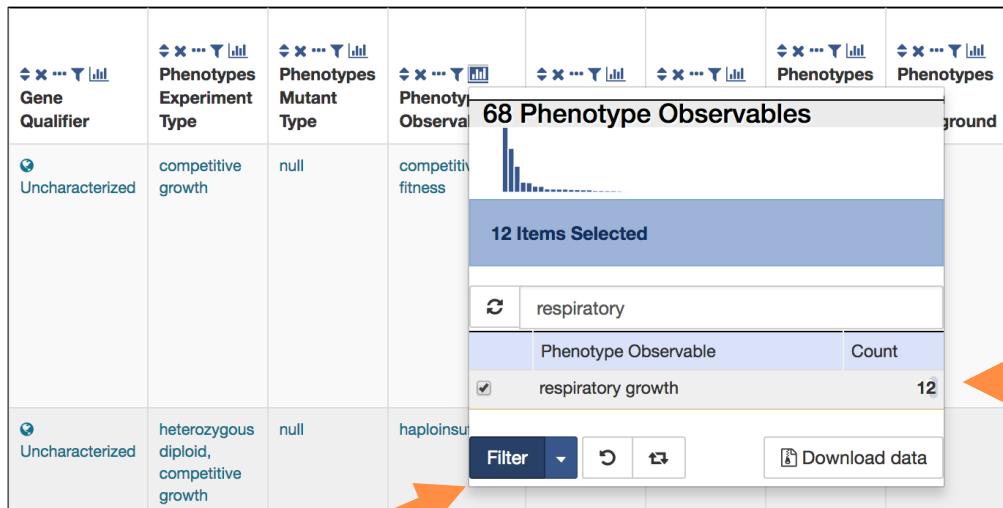
- You'll get a massive set of results but go ahead and open the table by clicking on the list of results at the bottom.
- In the Results table, find a column labeled **Phenotypes Observable**. Hover your mouse over the small icons above the column name and click on the bar graph icon.



- In the **Filter values** box, enter **respiratory** and scroll down the list to check the box next to **Respiratory growth**; hit **Filter**.

Showing rows 1 to 25 of 2,471 Rows per page: 25 page 1 ← → ← → ← →

Gene Primary DBID	Gene Standard Name	Gene Systematic Name	Gene Sgd Alias	Gene Qualifier	Phenotypes Experiment Type	Phenotypes Mutant Type	View column summary	Phenotypes Observable	Phenotypes Qualifier	Phenotypes Allele	Phenotypes Allele Comment	Phenotypes Strain Background	Phenotypes Chemical
S000000035	NO VALUE	YAL037W	NO VALUE	Uncharacterized	competitive growth	null	competitive fitness	increased	NO VALUE	NO VALUE	NO VALUE	S28c	
S000000035	NO VALUE	YAL037W	NO VALUE	Uncharacterized	heterozygous diploid, competitive growth	null	haploinsufficient	NO VALUE	NO VALUE	NO VALUE	NO VALUE	S28c	



- To filter the phenotypes for those where respiratory growth is impeded, find the **Phenotype Qualifiers** column and click the bar graph icon again. Select all items that refer to hindering respiratory growth: “decreased”, “decreased rate”, “absent”, etc. Then, hit **Filter**.
- Now save the gene list as "List 4: Genes with respiratory growth phenotypes"

Save a list of 1164 Genes

Name

Optional attributes

Description

CANCEL **SAVE**

- Go back to the Lists tab and click the boxes next to List 3 (uncharacterized interactors) and List 4 (genes with respiratory phenotypes). Choose "Intersect lists."

The screenshot shows a list management interface with the following details:

- Header:** Combine lists, Intersect lists (selected), Difference lists, Subtract lists.
- Page Settings:** Rows per page: 20, Page: 1 of 1.
- List Details:** DATE, TYPE, TAGS, and a refresh icon.
- Lists:**
  - List 3: Intersected Gene List (07 May 2024 15:01:09)**: [89] items, 17 mins ago, Gene type, edit icon.
  - List 4: Genes with respiratory phenotypes (Tue May 07 2024 11:19:32 GMT-0400 (Eastern Daylight Time))**: [1164] items, Just now, Gene type, edit icon.
  - List 1 MTR genes (Tue May 07 2024 10:53:51 GMT-0400 (Eastern Daylight Time))**: [91] items, 26 mins ago, Gene type, edit icon.
  - List 2: Genes that genetically interact with MTRs (Tue May 07 2024 10:58:04 GMT-0400 (Eastern Daylight Time))**: [90] items, 22 mins ago, Gene type, edit icon.

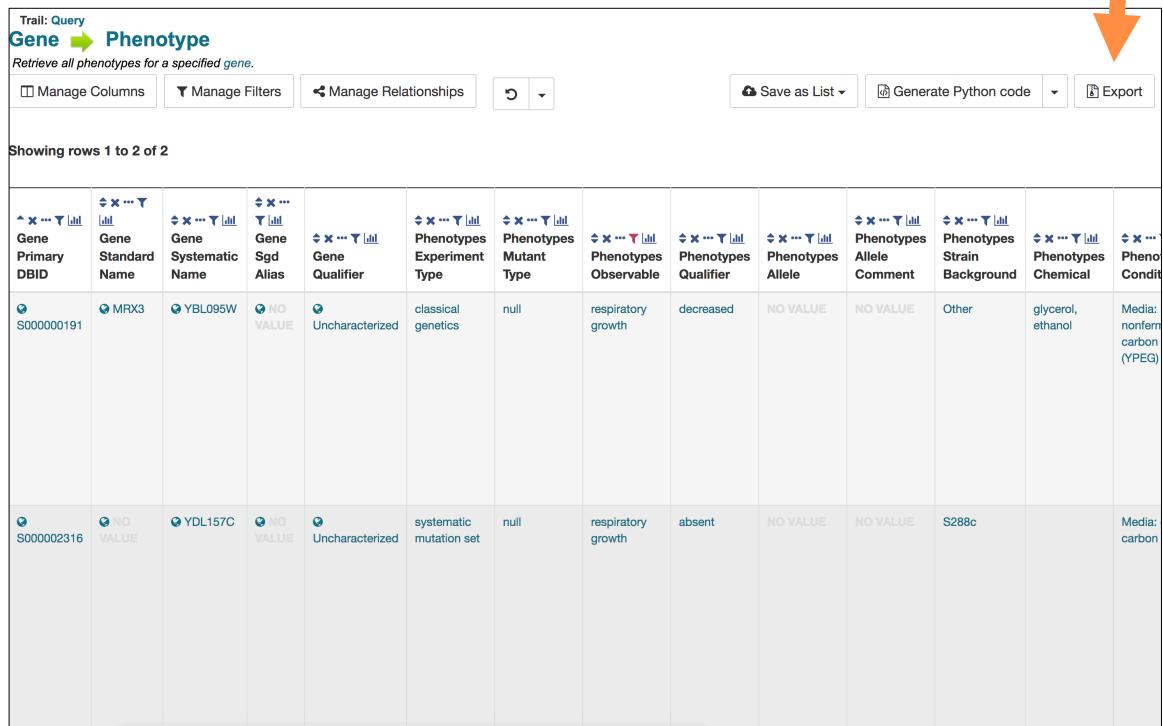
- Save your intersected list as " List 5: Uncharacterized MTR genes with respiratory phenotypes"

The dialog box displays the following information:

- Title:** List 3: Uncharacterized Interactors, Date: 2 mins ago, Type: Gene, Tags: -
- Title:** List 4: Genes with respiratory growth phenotypes, Date: 15 mins ago, Type: Gene, Tags: -
- New list:**
  - Title:** List 5: Uncharacterized MTR genes with respiratory phenotypes
  - Tags (Optional):** A dropdown menu.
  - Description (Optional):** An empty text area.
- Buttons:** CANCEL, SAVE NEW LIST.

- You should now have a list of two uncharacterized yeast genes whose products interact with mitochondrial ribosomes and mutations lead to respiratory growth defects. You can export the results into a .tsv file by clicking on the **Export** button,

and then on the “**Download file**” button in the resulting pop-up window.



The screenshot shows a table with 15 columns. The columns are labeled as follows:

- Gene Primary DBID
- Gene Standard Name
- Gene Systematic Name
- Gene Sgd Alias
- Gene Qualifier
- Phenotypes Experiment Type
- Phenotypes Mutant Type
- Phenotypes Observable
- Phenotypes Qualifier
- Phenotypes Allele
- Phenotypes Allele Comment
- Phenotypes Strain Background
- Phenotypes Phenotype Chemical
- Pheno Condit

Two rows of data are shown:

Gene Primary DBID	Gene Standard Name	Gene Systematic Name	Gene Sgd Alias	Gene Qualifier	Phenotypes Experiment Type	Phenotypes Mutant Type	Phenotypes Observable	Phenotypes Qualifier	Phenotypes Allele	Phenotypes Allele Comment	Phenotypes Strain Background	Phenotypes Phenotype Chemical	Pheno Condit
S000000191	MRX3	YBL095W	IO VALUE	Uncharacterized	classical genetics	null	respiratory growth	decreased	NO VALUE	NO VALUE	Other	glycerol, ethanol	Media: nonferm carbon (YPEG)
S000002316	IO VALUE	YDL157C	IO VALUE	Uncharacterized	systematic mutation set	null	respiratory growth	absent	NO VALUE	NO VALUE	S288c		Media: carbon

- The results of the above AllianceMine analysis suggest 2 genes that potentially encode undiscovered subunits of the mitochondrial ribosome. Although these genes are uncharacterized, more data may exist on their orthologs in other organisms. Using the exported Gene IDs, you can use FungiDB to survey the function of orthologs in Fungi and Oomycetes.

## Exercise: Attaching Track Hubs to Ensembl Fungi

Links to be clicked shown in blue, text to be entered shown in red.

There are a number of publicly available datasets that are available to add to views in Ensembl. You can find full lists of these at <https://trackhubregistry.org/>. We're going to search and add these files from within Ensembl.

Go to [fungi.ensembl.org](https://fungi.ensembl.org) on your browser and click the favourite species: *Magnaporthe oryzae (MG8)* then search for the region **1:5529955-5557959**

The screenshot shows the Ensembl Fungi search interface. At the top, it displays the species **Magnaporthe oryzae (MG8)**. Below the search bar, there is a placeholder text: "e.g. MGG\_01236 or 2:2789925-2792654 or WD repeat-containing protein".

This will take you directly to the Region in Detail page in the location tab. Click on the **Custom tracks** button found just below the 'Configure this page' button on the left. In the pop-up menu, click on **Track Hub Registry Search** on the left-hand navigation panel.

The screenshot shows the 'Track Hub Registry search' interface. It includes fields for **Assembly:** (set to **Botrytis cinerea B05.10**) and **Data type:** (set to **-- all --**). There is also a **Text search:** field and a **Search** button. A note at the bottom says: "Hint: Leave 'text search' empty to show all track hubs for this genome". Three callout boxes provide instructions: one pointing to the 'Track Hub Registry search' button, another pointing to the 'Filter public tracks by data type' dropdown, and a larger one pointing to the 'Search' button with the text "Search with default options (without a keyboard) to see all available track hubs for this genome".

Just click **Search** with no options selected.

## Transcriptomics & proteomics: Ensembl Fungi (RNAseq data)

The screenshot shows the Ensembl Fungi search results page. At the top, there are tabs for 'Configure Region Image', 'Configure Overview Image', 'Configure Chromosome Image', and 'Personal Data'. Below these are three buttons: 'Custom tracks', 'Track Hub Registry Search' (which is highlighted in brown), and 'Manage Configurations'. The main content area is titled 'Search Results' and shows a search for 'Magnaporthe oryzae MG8'. It displays two results:

- Magnaporthe oryzae poly(A) sites**  
Description: Magnaporthe oryzae poly(A) sites  
Data type: genomics  
Number of tracks: 1  
An 'Attach this hub' button is located to the right.
- RepetDB Magnaporthe oryzae 70-15 (rice blast fungus) (MG8)**  
Description: Repeat region consensus copies annotations created by TEannot (from the REPET package). Go to [RepetDB](#) for more info.  
Data type: genomics  
Number of tracks: 7  
An 'Attach this hub' button is located to the right.

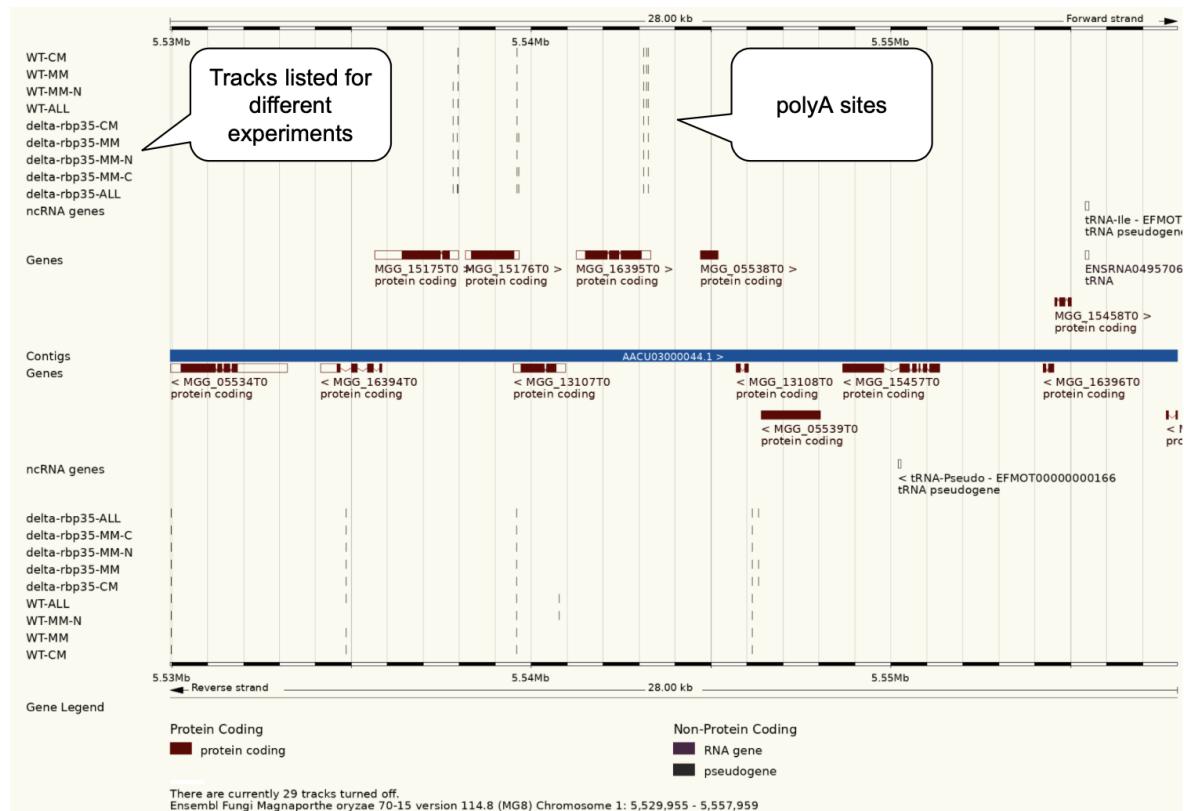
A note at the top right says: 'Can't see the track hub you're interested in? We only search for hubs compatible with assemblies used on this website - please [search the registry directly](#) for data on other assemblies. Alternatively, you can [manually attach any hub](#) for which you know the URL.'

There are two available track hubs for this assembly, one of which is currently unavailable (e! genomes 61)

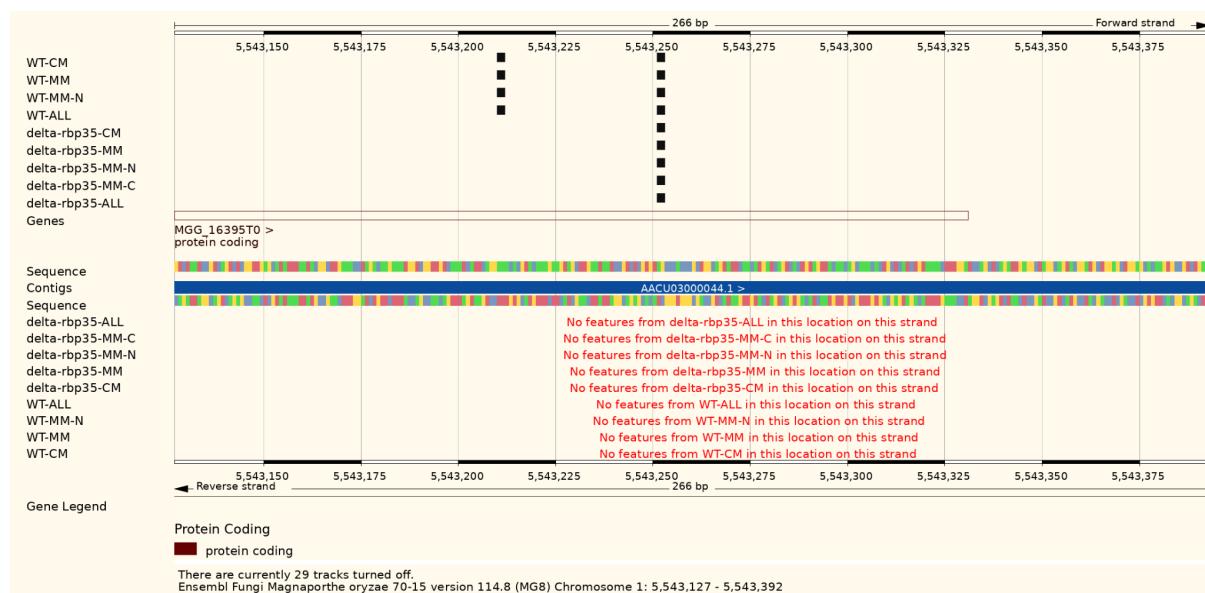
Choose the ‘Magnaporthe oryzae poly(A) sites’ by clicking on the [Attach this hub](#) button on the right. These are polyadenylation sites which are necessary for virulence - <https://www.nature.com/articles/sdata2018271>

The track hub should now load and appear on the most-detailed image at the bottom of the ‘Region in detail’ page.

## Transcriptomics & proteomics: Ensembl Fungi (RNAseq data)



If you zoom in further, you can see a more detailed representation of the data:



- (a) Go to [www.trackhubregistry.org](http://www.trackhubregistry.org) on your browser and search for **Magnaporthe oryzae poly(A) sites**. Can you jump to Ensembl Fungi directly from the Track Hub Registry page? (this may not work in e! genomes 61)

# The Track Hub Registry

A global centralised collection of publicly accessible track hubs

The goal of the Track Hub Registry is to allow third parties to advertise [track hubs](#), and to make it easier for researchers around the world to discover and use track hubs containing different types of genomic research data.

SRP062592 🔍

The screenshot shows the main interface of the Track Hub Registry. At the top, there's a navigation bar with links for "The Track Hub Registry", "Submit data", "Documentation", "About", "Help", "Data Preservation", a search bar ("Search by keywords: hg19, 🔎"), and user options ("Register", "Login"). Below the navigation, a breadcrumb trail indicates the current page: "Home / MagnaportheOryzaePolyAsites - GCA\_000002495.2". The main content area is divided into two main sections: "General Info" and "Hub".

**General Info:**

- Remote data tracks: 1
- Data Type: genomics
- File type(s):
- bigbed: 1
- Source URL: [View](#)
- Last updated: Mon Dec 12 2022 18:42:02 GMT+0000
- Status: All is Well
- [View in Genome Browser](#)

**Hub:**

- Name:** MagnaportheOryzaePolyAsites
- Short Label:** Magnaporthe oryzae poly(A) sites
- Long Label:** Magnaporthe oryzae poly(A) sites
- Assembly Hub:** X
- Public URL:** [View](#)

**Species:**

- Taxonomy**: 242507
- Scientific name:** Magnaporthe oryzae 70-15
- Common name:**

If you have your own files, or know a file you want to attach that is not present on the TrackHub registry, you can also attach these. There are two ways to do this, either by URL or by file upload.

Larger files, such as BAM files generated by NGS, need to be attached as remote files by URL. There are some BAM files for *Schizosaccharomyces pombe* available at:  
[ftp://ftp.ensemblgenomes.org/pub/misc\\_data/bam/fungi/Spom/](ftp://ftp.ensemblgenomes.org/pub/misc_data/bam/fungi/Spom/)

Let's take a look at that URL.

**NOTE:** Many internet browsers have recently dropped support for FTP, including the latest Firefox and Google Chrome versions. Firefox v87.0 still contains built-in FTP implementation. If you struggle to open the FTP site, try the HTTP version:  
[https://ftp.ebi.ac.uk/ensemblgenomes/pub/misc\\_data/bam/fungi/Spom/](https://ftp.ebi.ac.uk/ensemblgenomes/pub/misc_data/bam/fungi/Spom/)

## Index of /ensemblgenomes/pub/misc\_data/bam/fungi/Spom

	<a href="#">Name</a>	<a href="#">Last modified</a>	<a href="#">Size</a>	<a href="#">Description</a>
	<a href="#">Parent Directory</a>		-	
	<a href="#">Spom_all_61G9EAAXX_and_61G9UAAXX.+sorted.bam</a>	2014-11-26 15:06	3.3G	
	<a href="#">Spom_all_61G9EAAXX_and_61G9UAAXX.+sorted.bam.bai</a>	2014-11-26 15:06	36K	
	<a href="#">Spom_all_61G9EAAXX_and_61G9UAAXX.-sorted.bam</a>	2014-11-26 15:04	3.8G	
	<a href="#">Spom_all_61G9EAAXX_and_61G9UAAXX.-sorted.bam.bai</a>	2014-11-26 15:04	37K	

Here you can see two BAM files (file names ending in ‘.bam’) with corresponding index files (file names ending in ‘.bam.bai’). We’re interested in the files

[Spom\\_all\\_61G9EAAXX\\_and\\_61G9UAAXX.+sorted.bam](#) and

[Spom\\_all\\_61G9EAAXX\\_and\\_61G9UAAXX.+sorted.bam.bai](#). These files are the BAM file and the index file respectively. When attaching a BAM file to Ensembl, there must be an index file in the same folder.

From the Ensembl Fungi homepage, click on [Schizosaccharomyces pombe](#) (ASM294v2), then on [Display your data in Ensembl Fungi](#).

A menu will appear:

## Transcriptomics & proteomics: Ensembl Fungi (RNAseq data)

The screenshot shows the 'Add a custom track' form. At the top, there's a note about track hubs and indexed files not working with certain cloud services. Below that, there's a field to 'Give your track a name'. The 'Species' field is set to 'Schizosaccharomyces pombe' and 'Assembly: ASM294v2'. The 'Data' field contains a URL: [http://ftp.ensemblgenomes.org/pub/misc\\_data/bam/fungi/Spom/Spom\\_all\\_61G9EAAXX\\_and\\_61G9UAAXX.+.sorted.bam](http://ftp.ensemblgenomes.org/pub/misc_data/bam/fungi/Spom/Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam). There are two ways to upload data: 'Paste data or URL here' (with a tooltip) and 'Upload a file from your local machine (max. 20MB)' (with a tooltip). A dropdown menu for 'Data format' has 'BAM' selected. A tooltip for this dropdown says 'Ensembl automatically recognises the file extension when given'. At the bottom, there's a 'Help on supported formats, display types, etc.' link and a 'Add data' button.

The interface detects file extensions if you upload or attach a file. If you want to upload a file, just click on [Choose file](#), select the file from your local machine and it should automatically detect the file type you have submitted.

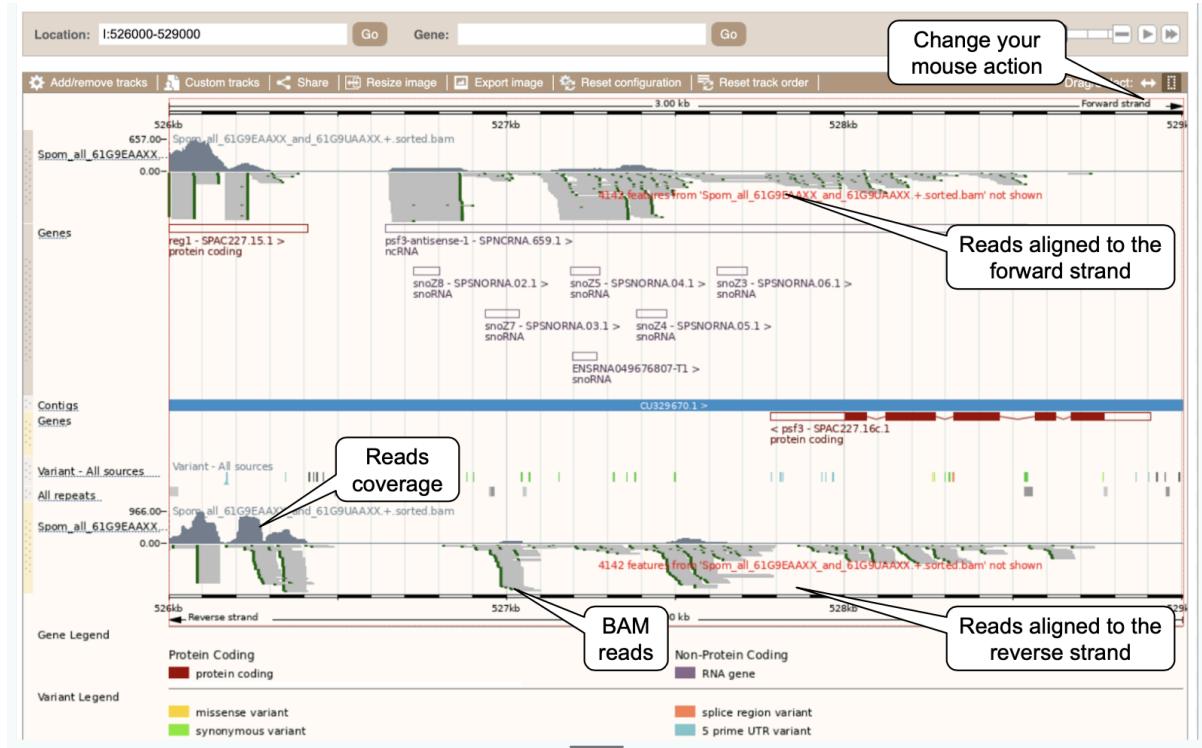
If you have a URL, like the one we located earlier, paste the URL of the BAM file itself into the 'Data' field

[http://ftp.ensemblgenomes.org/pub/misc\\_data/bam/fungi/Spom/Spom\\_all\\_61G9EAAXX\\_and\\_61G9UAAXX.+.sorted.bam](http://ftp.ensemblgenomes.org/pub/misc_data/bam/fungi/Spom/Spom_all_61G9EAAXX_and_61G9UAAXX.+.sorted.bam)

Since this is a file, the interface is able to detect the '.bam' file extension and automatically labels the format as **BAM**. Click on [Add data](#) and close the menu. It may take a while to load as there is a lot of data (Firefox tends to be fast). Once the data has been uploaded, you'll get a thank you message. Close the window and jump to a [Location](#) tab to see this data. Let's go to [I:526000-529000](#).

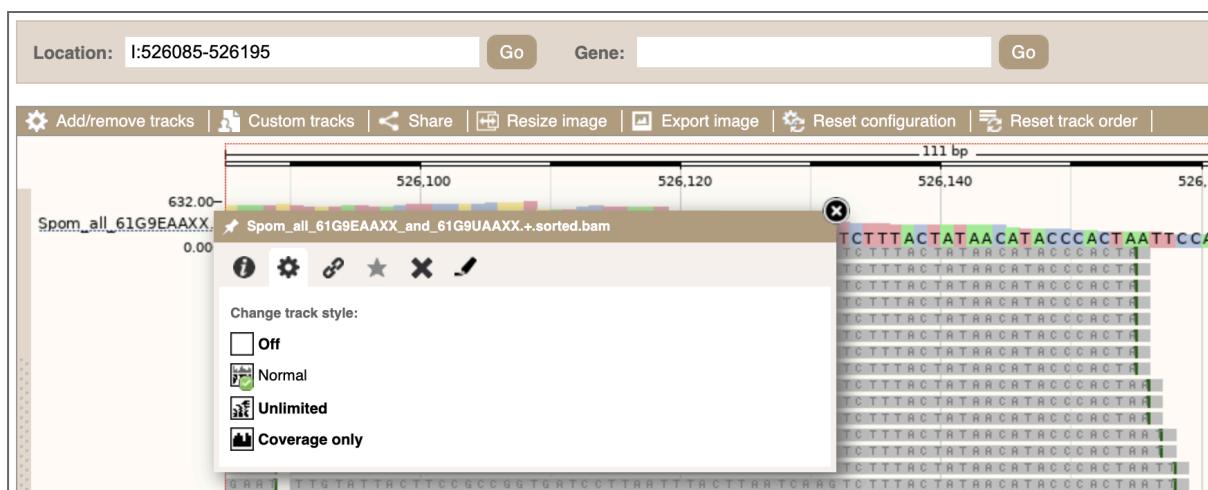
The screenshot shows the Ensembl Fungi search results for the query 'I:526000-529000'. The results page has a header with the Ensembl Fungi logo and links for HMMER, BLAST, BioMart, Tools, Downloads, Help & Docs, and Blog. Below that, it shows the species 'Schizosaccharomyces pombe (ASM294v2)'. The main area is a search bar containing the query 'I:526000-529000' with a 'Go' button. Below the search bar, there's a note: 'e.g. [SPAC2F7.03c](#) or [I:521420-541420](#) or [nucleolus](#)'.

## Transcriptomics & proteomics: Ensembl Fungi (RNAseq data)

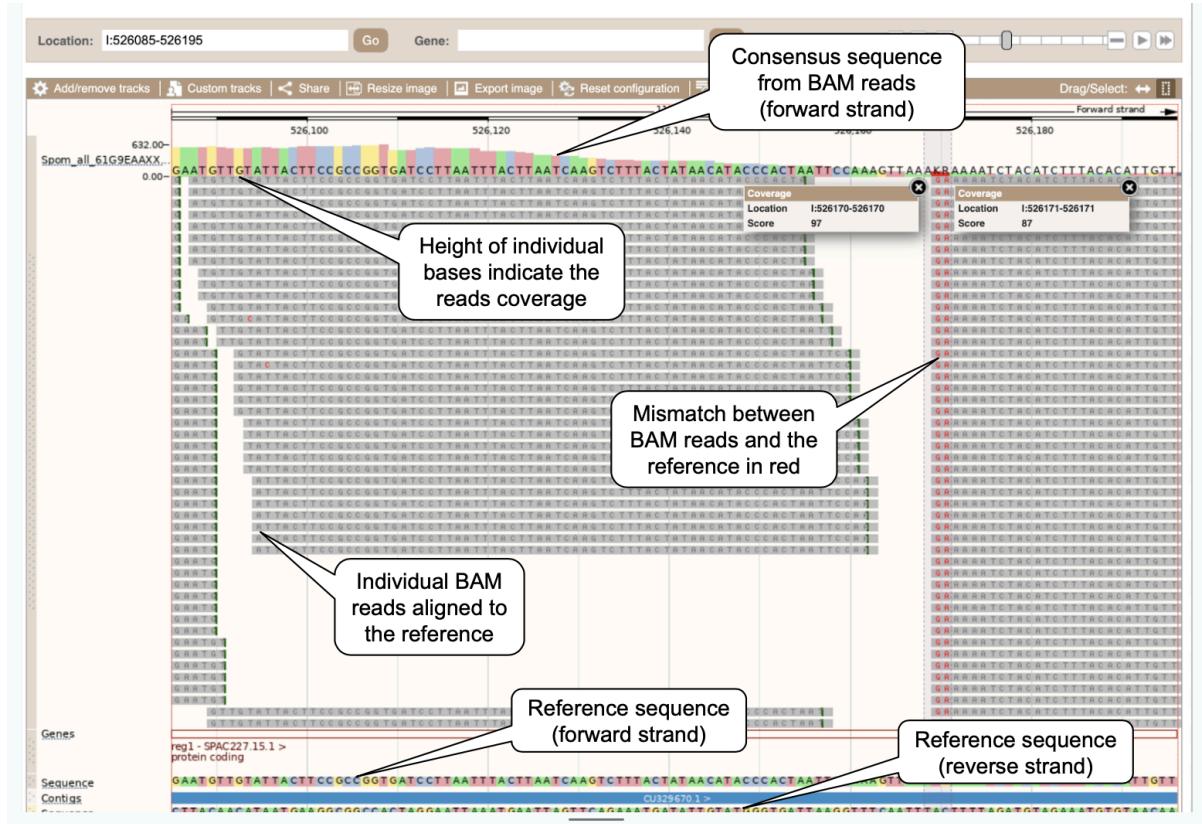


Newly added BAM file track split into forward and reverse stranded reads. You can zoom in to see the sequence itself. Drag out boxes in the view to zoom in, until you see a sequence of individual reads, or jump to a 110 bp region: [I:526085-526195](#).

- (b) Change the track style of the newly added track to **Unlimited** (showing all reads). Can you spot a site called differently from the reference in our sample? What is its genomic position? What is the read coverage at this position on the forward strand? Would you consider it a real variant or an artefact?



## Transcriptomics & proteomics: Ensembl Fungi (RNAseq data)



## Advanced Search Strategies.

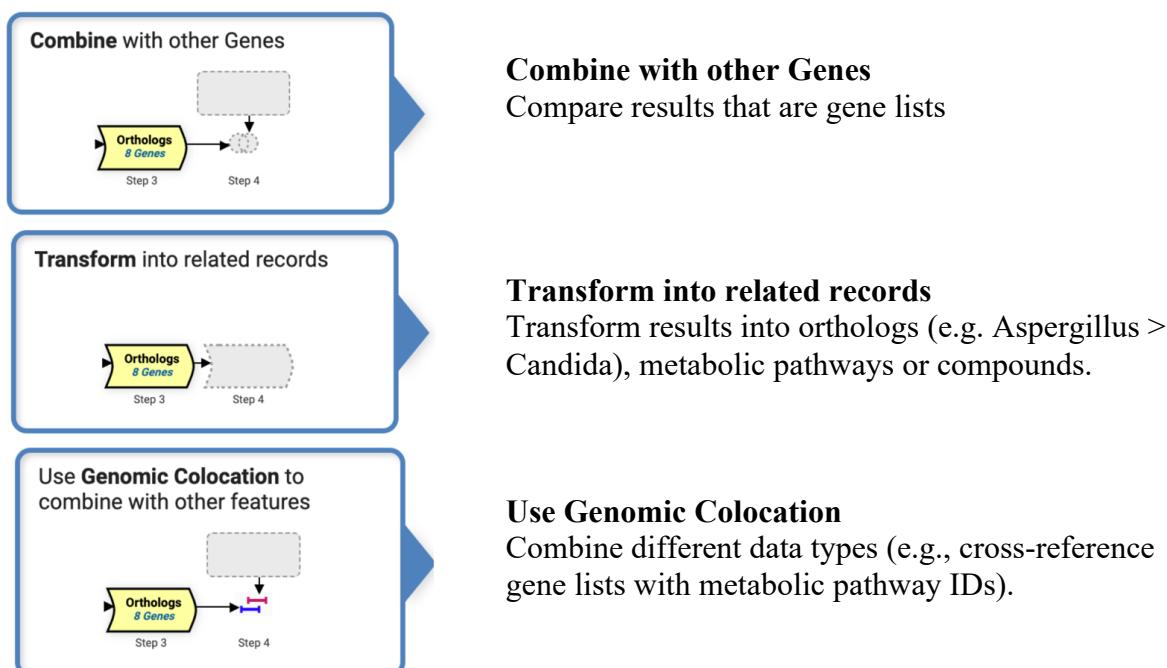
### Learning objectives

- Conduct various data searches and develop an advanced search strategy in FungiDB.

The strategy system offers a unique system of structured searches that can be combined to create multi-step *in-silico* experiments. As seen above, searches can be deployed from the site search, or the ‘Search For...’ menu on the home page, and from the ‘Searches’ dropdown menu in the header of every page.

Searches listed under the “Genes” category will return a list of gene IDs, while searches listed under the ‘SNPs’ or ‘Metabolic Pathways’ will return records relevant to SNPs data (e.g., sequences) and metabolic pathways, respectively.

When creating multi-step search strategy, the search strategy steps can be combined via three methods:



Within the search strategy, each step is connected via the system of Boolean operators that can intersect, unite, or subtract similar records (e.g., gene lists) and cross-references different types of data via the genomic colocation option.

Steps within the strategy can also be concealed using "ignore step" Boolean operators, enabling rapid modifications to the strategy without necessitating step deletion.

Revise as a boolean operation

1 INTERSECT 2     1 UNION 2     1 MINUS 2     2 MINUS 1

---

Revise as a span operation

1 RELATIVE TO 2, using genomic colocation

---

Ignore one of the inputs

IGNORE 2     IGNORE 1

---

Revise

## Advanced Search Strategies. Transcriptomics.

In this section, you will learn how to use transcriptomics and epitope data to create an *in silico* experiment.

There are different ways to search through transcriptomics datasets. The following search schemas can be used to explore the datasets in various ways:

Legend:  C Coexpression  S Similarity  DE Differential Expression  FC Fold Change  MC MetaCycle  P Percentile  SA SenseAntisense

- **Coexpression.** Search for genes which have positive or negative correlations with a set of genes.
- **Similarity.** Search for genes which have a similar profile for an experiment.
- **Differential Expression (DE).** We will use this search today. It uses DESeq2 analysis results. You can choose the directionality and magnitude of the difference by setting both fold change and adjusted p values. For example, selecting up-regulated genes with a fold difference of 2 and an adjusted p-value cutoff of 0.1 will only show results where the comparator is twice that of the reference with an adjusted p-value of 0.1 or less.
- **Fold change (FC).** Find genes with changes in gene expression when statistical analysis is not available (e.g. no replicates). After selecting samples, you have the option to take the average, minimum, or maximum expression value within each group. If choosing only one sample from a group, the selected 'operation' will not affect your results. Time-series experiments will offer an extra parameter called "Global min/max" which allows you to filter your results further. Finally, you can choose the directionality and the magnitude of the difference (e.g., up/down regulated, fold difference of 2, etc.)
- **MetaCycle.** This search is applied to circadian datasets. For each study/experiment, you can choose either ARSER (Yang and Su 2010) or JTK\_Cycle (Hughes et al. 2010), which are methods for detecting rhythmic signals. The search will return the corresponding period, amplitude, and p-value.
- **Percentile (P).** For each Experiment and Sample, genes are ranked by expression level (e.g., search for low/high gene expression levels).
- **Sense/antisense (SA).** This search is applied to stranded datasets. You can find genes that exhibit simultaneous changes in sense and antisense transcripts in the Comparison sample relative to the Reference Sample. For example, you could look for genes showing increasing antisense transcripts and decreasing sense transcripts, as might occur when antisense transcription suppresses sense transcription. The search will perform all pairwise comparisons between the Comparison and Reference samples.

In this exercise, we will identify *Aspergillus fumigatus* Af293 genes that:

- A. Are up-regulated when *Aspergillus* is exposed to human airway epithelial cells,
- B. Are known to be immune-reactive.

Here is a step-by-step guide on how to create this *in-silico* experiment:

**A. Deploy the “RNA-Seq Evidence” search to identify genes that are up-regulated when *Aspergillus* is exposed to human airway epithelial cells.**

1. Select the search from the “Search for...” panel (shown below) or the “Searches” menu at the top of the page.
2. Identify the “Response to caspofungin (Valero et al. 2020)” dataset and click on the DE (Differential expression) button.
- Tip: Use the filter box to quickly find relevant search results.
3. Choose the parameters of your search:
  - i. Reference sample: **WT\_CT**
  - ii. Comparator sample: **WT\_CSP**
  - iii. Direction: **up-regulated**
  - iv. Fold difference: **2**
 Leave other parameters at default.
4. Click on the “Get Answer” button.

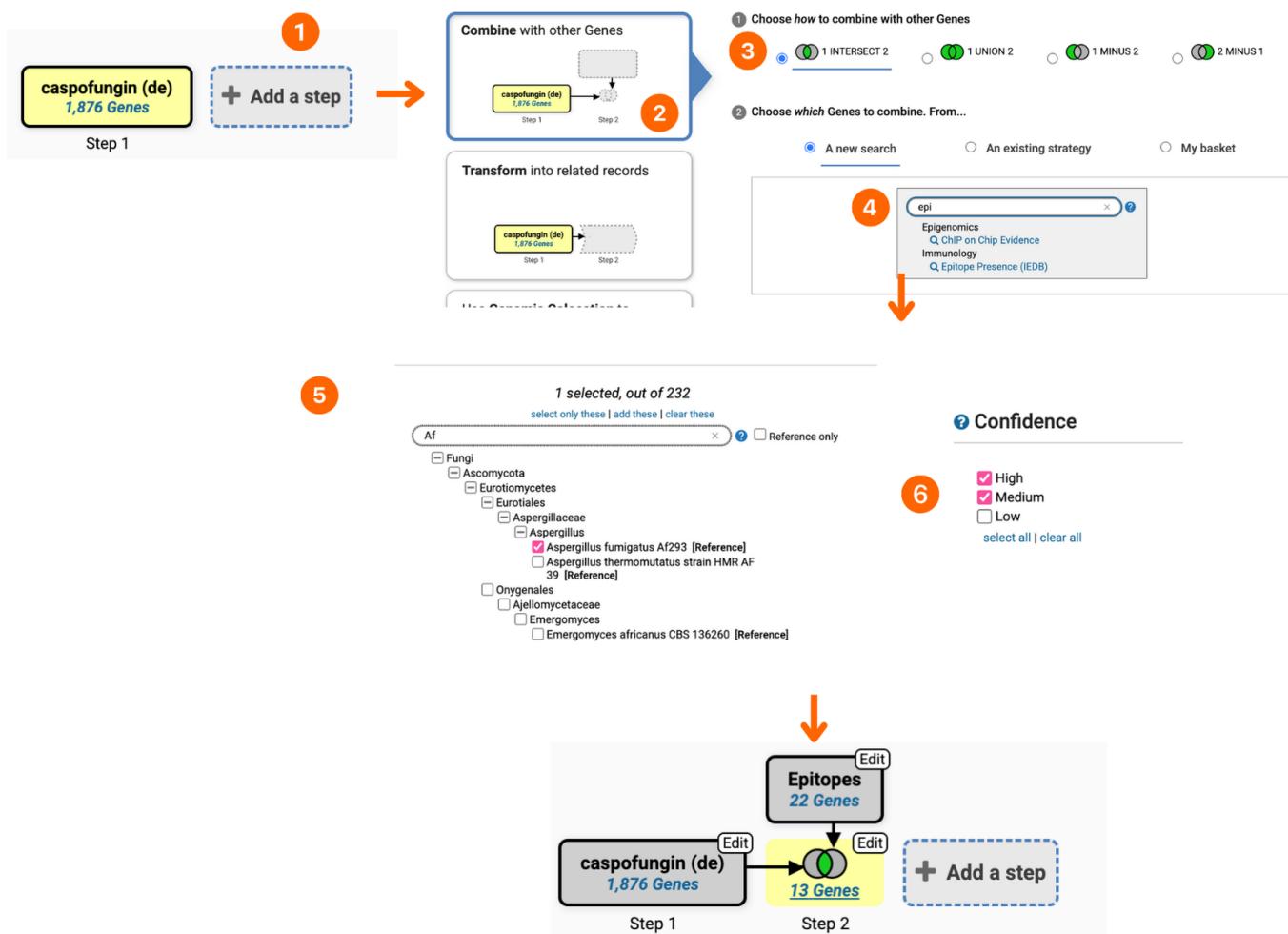
The screenshot shows the RNA-Seq Evidence search interface with the following steps highlighted:

- Step 1:** The "Search for..." panel on the left shows a search term "mif". A red circle labeled "1" is over the "RNA-Seq Evidence" option under the "Gene models" section.
- Step 2:** An arrow points from the "Search for..." panel to the "Identify Genes based on RNA-Seq Evidence" search results page. The results show a dataset for "Response to caspofungin (Valero et al. 2020)" with 2 results. A red circle labeled "2" is over the "DE" button in the "Choose a Search" section.
- Step 3:** An arrow points from the search results page to the "Experiment" configuration page. A red circle labeled "3" is over the "Reference Sample" section where "WT\_CT" is selected.
- Step 4:** An arrow points from the "Experiment" page to the "Get Answer" button. A red circle labeled "4" is over the "Get Answer" button, which is highlighted in yellow. To the right, a yellow box highlights the search results: "caspofungin (de) 1,876 Genes".

## B. Identify genes with immune reactivity.

The immune system recognizes epitopes, which can be used for vaccine development. To identify genes that have annotated epitope records:

1. Click on the “Add a step” button.
2. Make sure to select the “Combine with other Genes” option.
3. Select the “2 INTERSECT 3” Boolean operator (if not selected by default).
4. Filter available searches for “epitope” to identify and deploy the “Epitope Presence (IEDB)” search.
5. Set organism to *Aspergillus fumigatus* Af293.
6. Set Confidence to “high” and “Medium” and click on the “Run Step” button.



Well done! You have created an in-silico experiment using three different types of data – RNA-Seq, SNPs, and epitope data.

Search strategy link:

<https://fungidb.org/fungidb/app/workspace/strategies/import/2c7df1a9e21d4d00>

## Advanced Search Strategies. Proteomics

In this section, you will learn how to query quantitative mass spec data.

Fungal extracellular vesicles (EVs) have been implicated in host-pathogen and pathogen-pathogen communication. In this exercise, we will find proteins abundant in EVs when compared to whole cell lysate (WCL) in biofilm conditions.

- **Identify proteins more abundant in EVs than whole cell lysate (WCL).**

1. Navigate to the “Quantitative Mass Spec. Evidence” search.
2. Filter for “albicans” and click the “FC” button for the Dawson et al. 2020 dataset.
3. Look for up-regulated genes.
4. With a Fold change  $\geq 1$ .
5. Set Reference strain to DAY286 biofilm WCL mean.
6. Set Comparison Sample to DAY286 biofilm EV mean

Identify Genes based on Quantitative Mass Spec. Evidence

Legend: DC Direct Comparison FC Fold Change

Filter Data Sets: albicans 1 result (filtered from a total of 11)

Organism: Candida albicans SC5314 Data Set: Extracellular vesicle and whole cell lysate proteomes for DAY226 yeast/biofilm, ATCC90028 and ATCC10231 strains. (Dawson et al. 2020)

For the Experiment

Extracellular vesicle and whole cell lysate proteomes for DAY226 yeast/biofilm, ATCC90028 and ATCC10231 strains.

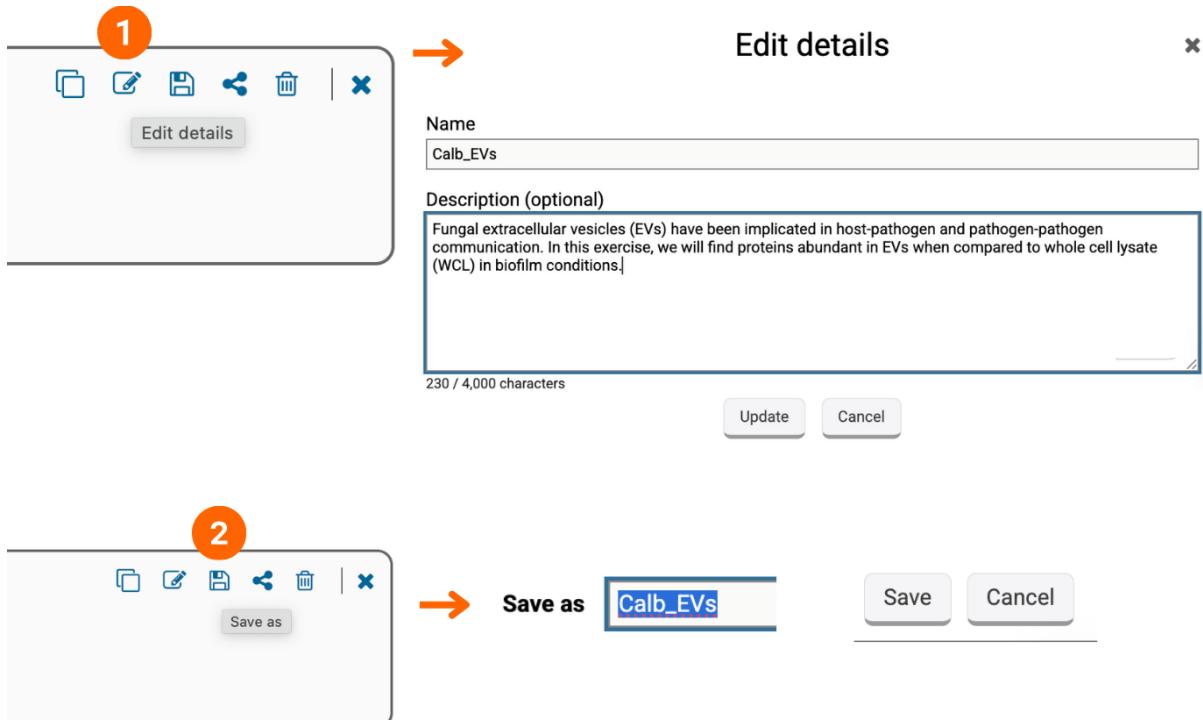
return protein coding Genes  
that are up-regulated  
with a Fold change  $\geq 1$   
between each gene's average expression value  
in the following Reference Samples

5. Reference Samples: DAY286 biofilm WCL mean (checked), DAY286 yeast EV mean, ATCC90028 yeast EV mean, ATCC90028 yeast WCL mean, ATCC10231 yeast EV mean, ATCC10231 yeast WCL mean  
select all | clear all

6. Comparison Samples: DAY286 yeast WCL mean, DAY286 biofilm EV mean (checked), DAY286 biofilm WCL mean, ATCC90028 yeast EV mean, ATCC90028 yeast WCL mean, ATCC10231 yeast EV mean  
select all | clear all

Calibcans\_EVs (fc)  
324 Genes

- Learn how to incorporate notes into your search strategy and save it.



Strategy URL:

Calb\_EVs: <https://fungidb.org/fungidb/app/workspace/strategies/import/c971467cff5062fa>

# Using SPELL to Analyze Expression Datasets & Coexpressed Genes at SGD

SPELL (Serial Pattern of Expression Levels Locator) is a query-driven search engine for large gene expression microarray compendia. Given a small set of query genes, SPELL identifies which datasets are most informative for these genes, then within those datasets additional genes are identified with expression profiles most similar to the query set.

Use SPELL to find out which genes are coexpressed with genes involved in glycolysis.

## Compile a list of genes involved in glycolysis.

- On the SGD home page ([www.yeastgenome.org](http://www.yeastgenome.org)), enter glycolysis into the search box and hit Enter.

The SGD home page features a search bar at the top with the query 'glycolysis'. Below the search bar, there's a section titled 'About SGD' containing general information about glycolysis. An orange arrow points to the search bar.

- On the Results page, click on the **Genes** category.

The SGD Results page for 'glycolysis' shows a sidebar with categories: References, Genes (selected), Biological Processes, Downloads, Molecular Functions, Cellular Components, and Chemicals. The main content area displays 644 results for 'glycolysis'. An orange arrow points to the 'Genes' link in the sidebar.

- Scroll down the page and find the **Biological Process** category on the left hand menu. Hit Show more and select **glycolytic process (direct)**.
- To download the list of genes, click on **Wrapped** and then on **Download**.

The SGD Results page for 'glycolysis' shows the 'Biological Process' category expanded. The sidebar includes 'Genes / Genomic Features' (selected) and 'Feature Type'. The main content area shows 15 results for 'glycolytic process (direct)'. At the bottom right, there are 'List' and 'Wrapped' buttons, with an orange arrow pointing to the 'Wrapped' button.

- The **Analyze** button, directly to the right of Download, enables you to import your search results directly into SPELL (among other tools at SGD). However, for the sake of demonstration, in this exercise we are instead going to enter our gene list into SPELL manually.

## Import your gene list into SPELL and run a query:

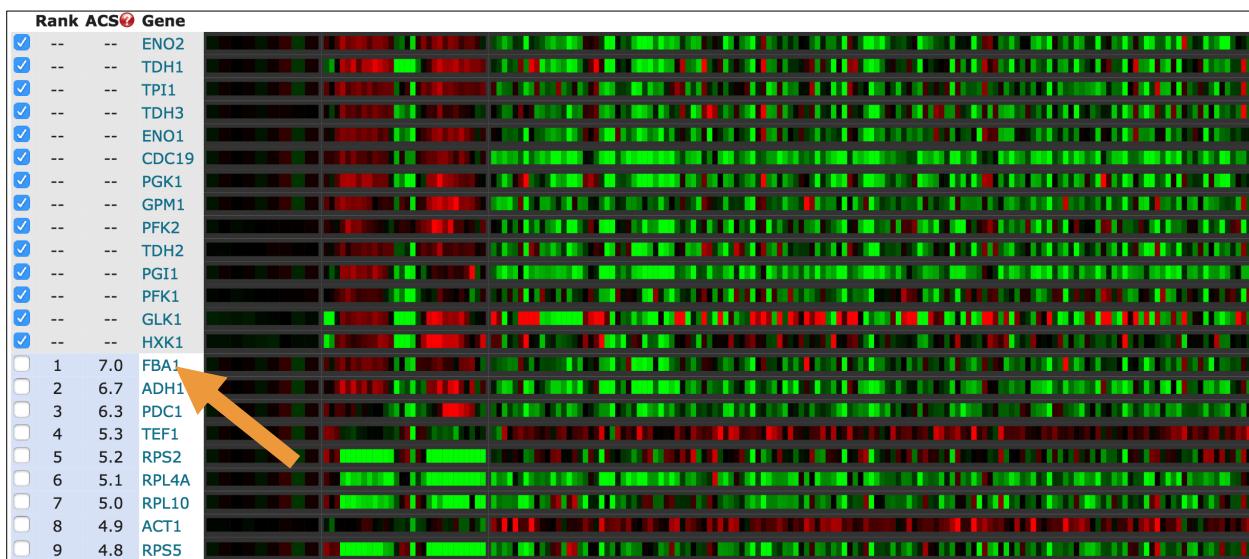
- To access SPELL, go to the SGD home page at [www.yeastgenome.org](http://www.yeastgenome.org), open the **Function** tab on top of the page and click on **Expression**. Or, if you are already on a Locus Summary page, open the Expression tab and click on the SPELL link under the histogram.

The screenshot shows the SGD home page with the 'Expression' menu item highlighted in orange. The 'Expression' menu contains links to Gene Ontology, Biochemical Pathways, Phenotypes, Interactions, YeastGFP, and Resources. To the right of the menu, a box titled 'About SGD' provides a brief overview of the database. Below the menu, there is a 'Meetings' section and a 'New & Noteworthy' section. On the far right, there is a 'Tweets' sidebar.

- On the SPELL page, copy and paste the list of glycolysis genes you downloaded in step 1 into the Gene Name(s) box. For the sake of demonstration, remove **FBA1** from your list before hitting Search. This is to test if SPELL can properly identify missing members of glycolysis based on coexpression.

The screenshot shows the SPELL search interface for *S. cerevisiae*. The search bar at the top contains the query 'SPELL (Yeast)'. Below the search bar, the 'Gene Name(s)' input field contains the text 'GPM1 PGK1 PFK1 PFK2 ENO2 ENO1 CDC'. An orange arrow points to the 'Search' button next to the input field. Below the input field, there is a dropdown menu labeled '# Results: 20' with an orange arrow pointing to it. At the bottom of the interface, there is a purple bar with the text '+ Options for Filtering Results by Dataset Tags'.

- Scroll down the list of genes on the left. Genes with checked boxes are from our query; the remaining genes are "hits", ordered from top to bottom according to their ranks. The rank reflects the correlation of expression of that gene with the query gene(s), given the relevance weight of that expression dataset. Thus, genes that show the highest degree of coexpression with the query genes in the most relevant datasets receive the highest rank.



- Notice that the glycolysis gene we deleted earlier, FBA1, is indeed the highest-ranking gene!
- Examine other genes enriched for this query set. You can click on their names to be taken to their respective summary pages at SGD. Does it make sense for any of these genes to be highly coexpressed with members of glycolysis?
- Click on + Additional Display Options to change the default mapping method and color scheme to blue/yellow. Directly above this section are options to change the number of genes and datasets shown in your results.

# of Result Genes to Show: 20 Datasets to view: From 1 to 10

+ Additional Display Options

Mapping method	Color scheme
For single channel data: Per-gene log <sub>2</sub> fold change	Red/Green
For dual channel data: Reported log <sub>2</sub> fold change	Red/Green

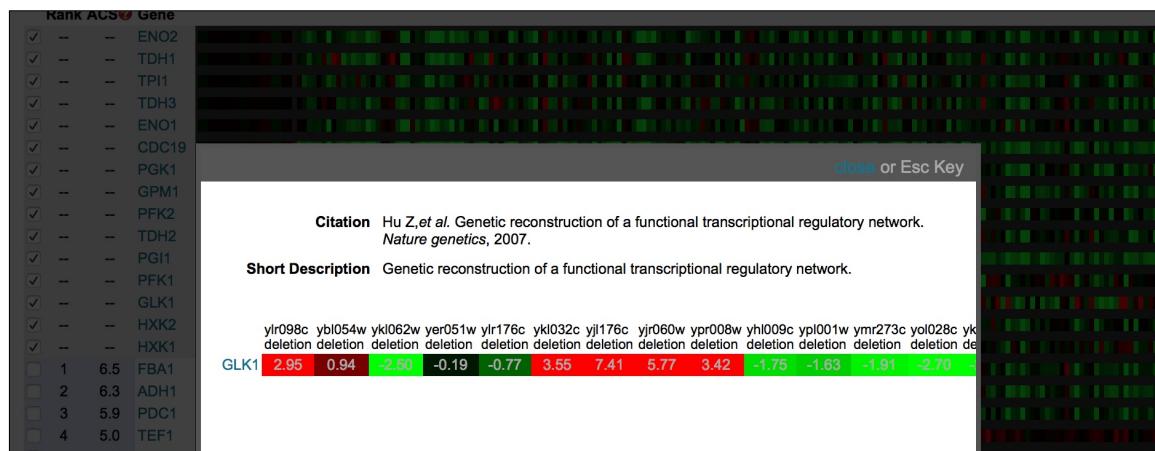
- To select only datasets with particular tags, click on + Options for Filtering Results.

Dataset Tags

Select: all none previous query toggle

<input type="checkbox"/> amino acid metabolism	<input type="checkbox"/> evolution	<input type="checkbox"/> organelles, biogenesis, structure, and function	<input type="checkbox"/> RNA catabolism
<input type="checkbox"/> amino acid utilization	<input type="checkbox"/> fermentation	<input type="checkbox"/> osmotic stress	<input type="checkbox"/> signaling
<input type="checkbox"/> carbon utilization	<input type="checkbox"/> filamentous growth	<input type="checkbox"/> oxidative stress	<input type="checkbox"/> sporulation
<input type="checkbox"/> cell aging	<input type="checkbox"/> flocculation	<input type="checkbox"/> oxygen level alteration	<input type="checkbox"/> starvation
<input type="checkbox"/> cell cycle regulation	<input type="checkbox"/> genetic interaction	<input type="checkbox"/> phosphorus utilization	<input type="checkbox"/> stationary phase entry
<input type="checkbox"/> cell morphogenesis	<input type="checkbox"/> genome variation	<input type="checkbox"/> ploidy	<input type="checkbox"/> stationary phase maintenance
<input type="checkbox"/> cell wall organization	<input type="checkbox"/> heat shock	<input type="checkbox"/> protein dephosphorylation	<input type="checkbox"/> stress
<input type="checkbox"/> cellular ion homeostasis	<input type="checkbox"/> histone modification	<input type="checkbox"/> protein glycosylation	<input type="checkbox"/> sulfur utilization
<input type="checkbox"/> chemical stimulus	<input type="checkbox"/> lipid metabolism	<input type="checkbox"/> protein modification	<input type="checkbox"/> synthetic biology
<input type="checkbox"/> chromatin organization	<input type="checkbox"/> mating	<input type="checkbox"/> protein phosphorylation	<input type="checkbox"/> transcription
<input type="checkbox"/> cofactor metabolism	<input type="checkbox"/> metabolism	<input type="checkbox"/> protein trafficking, localization and degradation	<input type="checkbox"/> transcriptional regulation
<input type="checkbox"/> diauxic shift	<input type="checkbox"/> metal or metalloid ion stress	<input type="checkbox"/> proteolysis	<input type="checkbox"/> translational regulation
<input type="checkbox"/> disease	<input type="checkbox"/> mitotic cell cycle	<input type="checkbox"/> QTLs	<input type="checkbox"/> ubiquitin or ULP modification
<input type="checkbox"/> DNA damage stimulus	<input type="checkbox"/> mRNA processing	<input type="checkbox"/> radiation	
<input type="checkbox"/> DNA replication, recombination and repair	<input type="checkbox"/> nitrogen utilization	<input type="checkbox"/> respiration	
<input type="checkbox"/> environmental-sensing	<input type="checkbox"/> nutrient utilization	<input type="checkbox"/> response to unfolded protein	

- Click on any patch in the heat map to open a page with information about its parent dataset.



- SPELL also runs a **Gene Ontology (GO) enrichment** for the results of your query. GO enrichments can tell you which gene ontology terms (in this case, biological process terms) are significantly associated with your set of genes. You can scroll down to the bottom of the page to view it.

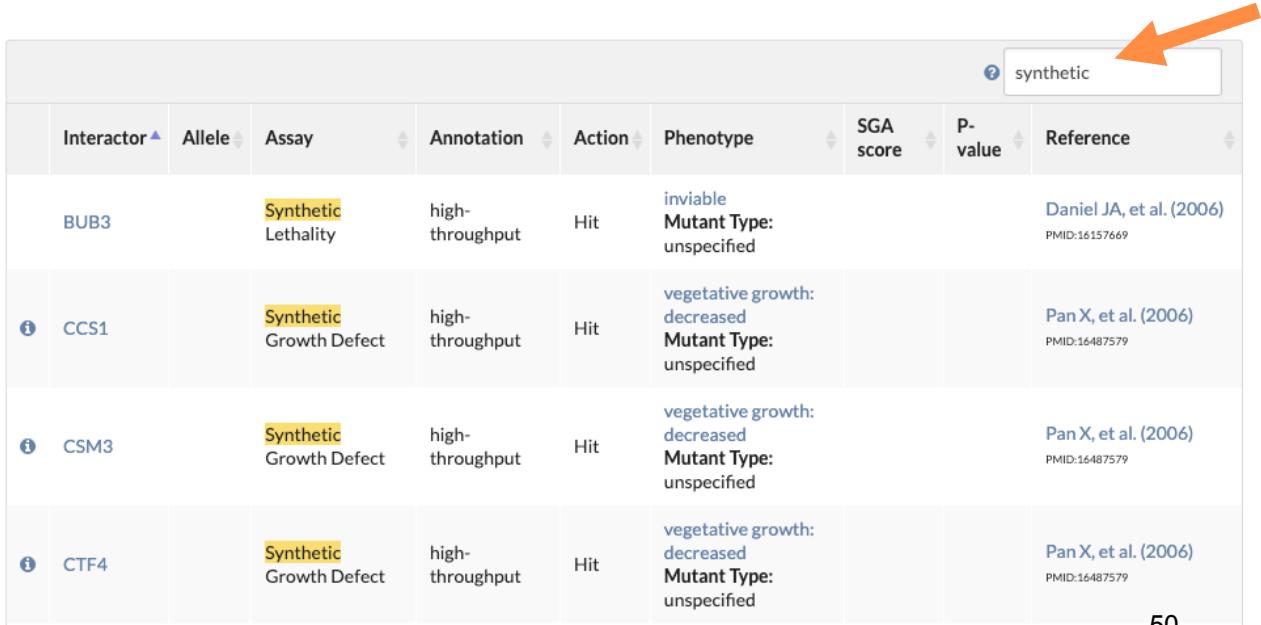
GO Term Enrichment?				Annotated Genes			
GOTerm	P-val	% query	% genome				
glucose catabolic process (biological_process)	1.33e-29	19 of 35	52 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
hexose catabolic process (biological_process)	2.39e-28	19 of 35	59 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
monosaccharide catabolic process (biological_process)	2.91e-27	19 of 35	66 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
glycolysis (biological_process)	4.79e-27	16 of 35	32 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, CDC19, PGK1, TDH2			
glucose metabolic process (biological_process)	1.66e-23	19 of 35	99 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
single-organism carbohydrate catabolic process (biological_process)	3.62e-22	19 of 35	115 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
hexose metabolic process (biological_process)	4.32e-22	19 of 35	116 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
monosaccharide metabolic process (biological_process)	1.42e-21	19 of 35	123 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
carbohydrate catabolic process (biological_process)	1.97e-21	19 of 35	125 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
generation of precursor metabolites and energy (biological_process)	7.97e-18	19 of 35	190 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
single-organism carbohydrate metabolic process (biological_process)	1.60e-13	19 of 35	319 of 6381	FBA1, TDH3, ENO1, HXK1, HXK2, PFK2, GLK1, GPM1, PFK1, TPI1, TDH1, PGI1, ENO2, PDC1, ADH1, PDC5, CDC19, PGK1, TDH2			
gluconeogenesis (biological_process)	3.72e-13	10 of 35	33 of 6381	FBA1, TDH3, ENO1, GPM1, TPI1, TDH1, PGI1, ENO2, PGK1, TDH2			
hexose biosynthetic process (biological_process)	5.25e-13	10 of 35	34 of 6381	FBA1, TDH3, ENO1, GPM1, TPI1, TDH1, PGI1, ENO2, PGK1, TDH2			
monosaccharide biosynthetic process (biological_process)	7.33e-13	10 of 35	35 of 6381	FBA1, TDH3, ENO1, GPM1, TPI1, TDH1, PGI1, ENO2, PGK1, TDH2			

# Using SGD GO Slim Mapper and Interaction Data to Predict Gene Function

The Gene Ontology (GO) is structured in a hierarchy, such that granular terms (“perinuclear space”) are connected and further down the hierarchy than their related broader terms (“nucleus”). However, for many purposes, such as reporting the upregulated cellular functions of a transcriptomics experiment, is very useful to focus on the broad, high-level part of the GO. For example, if you were interested in which of your upregulated genes are involved in DNA replication, it would be useful to map genes that have been annotated to specific terms (e.g. “synthesis of RNA primer involved in nuclear cell cycle DNA replication”) to more general terms (e.g. “DNA replication”).

The **Gene Ontology (GO) Slim Mapper** at SGD maps granular GO annotations of a group of genes to more general terms and/or bins them into broad categories, i.e., “**GO Slim**” terms. Using GO Slim Mapper, predict what biological processes an uncharacterized gene may be involved in based on its genetic interactions.

- From the SGD home page ([www.yeastgenome.org](http://www.yeastgenome.org)), go to the Locus Summary page for the uncharacterized gene **YLR287C**.
- Select **Interactions** tab at the top and then scroll to the **Genetic Interactions** section on the page. Here, we are interested in finding genes that have a genetic interaction with YLR287C, as the function of these genes may provide hints about the function of YLR287C.
- Enter “synthetic” in filter box on the **Genetic Interactions** table. This will filter the table for genes that, when knocked out in combination with YLR287C, elicit some sort of synthetic growth defect, haploinsufficiency, lethality, etc. These harsh phenotypes may suggest clues about related functions to YLR287C.



Interactor	Allele	Assay	Annotation	Action	Phenotype	SGA score	P-value	Reference
BUB3		Synthetic Lethality	high-throughput	Hit	inviable Mutant Type: unspecified			Daniel JA, et al. (2006) PMID:16157669
CCS1		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Pan X, et al. (2006) PMID:16487579
CSM3		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Pan X, et al. (2006) PMID:16487579
CTF4		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified			Pan X, et al. (2006) PMID:16487579

- Find and click on the **Analyze** button at the bottom of the Annotation table. This will import the table you filtered to a page where you can send the genes to other SGD tools.
- On the next page that lists the YLR287C interactors, select **GO Slim Mapper**.

Tools

---

<b>GO Term Finder</b> Find common GO annotations between genes.	<b>GO Slim Mapper</b> Sort genes into broad categories.	<b>SPELL</b> View expression data.	<b>YeastMine</b> Conduct advanced analysis.
--	--	---------------------------------------	--

---

Genes

Gene Name	Description
BUB3	Kinetochore checkpoint WD40 repeat protein; localizes to kinetochores during prophase and metaphase, delays anaphase in the presence of unattached kinetochores; forms complexes with Mad1p-Bub1p and with Cdc20p, binds Mad2p and Mad3p; functions at kinetochore to activate APC/C-Cdc20p for normal mitotic progression

- The GO Slim Mapper has three steps (plus one optional step) in which you can specify your query. The Query Set (Your Input) box has been preloaded in memory with the list of genes you imported from the table.

**Query Set (Your Input)**

Your gene list has been saved in the memory. Please pick a GO Slim Set, refine the Slim Terms, and Submit the form. 

Enter Gene/ORF names (separated by a return or a space):

Note: If you have a big gene list (> 100), save it as a file and upload it below.  
**OR** Upload a file of Gene/ORF names (.txt or .tab format):  
 No file selected.

**Specify your Slim Terms**

Choose a GO Set:

 Yeast GO-Slim: process

Refine your list of GO Slim Terms:

Select or unselect multiple datasets by pressing the Control (PC) or Command (Mac) key while clicking. Selecting a category label selects all datasets in that category.

**SELECT ALL Terms from Yeast GO-Slim: process**

DNA recombination ; GO:0006310  
 DNA repair ; GO:0006281  
 DNA replication ; GO:0006260  
 DNA-templated transcription, elongation ; GO:0006354

- Choose a **GO Set** by selecting **Yeast GO-Slim: Process** from the pull-down.
- Highlight **SELECT ALL Terms from Yeast GO-Slim: Process**.
- Click the **Submit Form** button to use the default settings or go further down to customize your query.

- Results appear in a table with four columns:
  - GO Slim terms picked by GO Slim Mapper
  - Genes from your list that are annotated to that term, hyperlinked to their Locus Summary pages.
  - GO Term Usage in Gene List (cluster frequency), the number and percentage of genes in your list annotated to each term.
  - Genome frequency of use, the number and percentage of all genes in the genome annotated to each term.
- You can also download the results in a tab-delimited file.

**Search Results**

Save Options: [HTML Table](#) | [Plain Text](#) | [Tab-delimited](#) | [Your Input List of Genes](#) | [Your GO Slim List](#)

GO version 2023-04-01

GO Terms from the biological process Ontology			
GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
DNA replication (GO:0006260)	YMR048W, YNL273W, YOR080W, YPR135W	4 of 13 genes, 30.77%	140 of 6489 annotated genes, 2.16%
regulation of DNA metabolic process (GO:0051052)	YLR233C, YMR048W, YNL273W, YOR080W	4 of 13 genes, 30.77%	108 of 6489 annotated genes, 1.66%
mitotic cell cycle (GO:0000278)	YGL086W, YJL030W, YMR048W, YOR026W	4 of 13 genes, 30.77%	317 of 6489 annotated genes, 4.89%
protein modification by small protein conjugation or removal (GO:0070647)	YDR510W, YJL030W, YOR026W, YOR080W	4 of 13 genes, 30.77%	137 of 6489 annotated genes, 2.11%
regulation of cell cycle (GO:0051726)	YGL086W, YJL030W, YNL273W, YOR026W	4 of 13 genes, 30.77%	247 of 6489 annotated genes, 3.81%
chromosome segregation (GO:0007059)	YGL086W, YJL030W, YMR048W	3 of 13 genes, 23.08%	162 of 6489 annotated genes, 2.50%

- Based on the results, what biological processes might YLR287C be involved in?

## GO Enrichment, Phenotype Data at CGD

The Gene Ontology (GO) provides a common language to describe aspects of a gene product's biology. GO Terms are standardized phrases, arranged in a hierarchy, that describe a gene product's **molecular function** ("protein kinase activity"), **biological process** ("gluconeogenesis"), and **cellular component** ("cytoplasm"). Together, molecular function, biological process, and cellular component are the three ontologies of GO that describe a gene product's function, the processes that function is involved in, and the location where the function is performed.

**GO Term Finder** takes a list of genes and identifies what GO terms are significant for the list. It is a powerful way to interpret the results of omics experiments or any situation where determining common functions and roles are important. For example, GO Term Finder can take a list of upregulated genes from an RNA-Seq experiment and determine what biological processes are significant for the set of genes, providing an idea of what processes are being upregulated in the cell.

In this exercise, we will attempt to uncover what processes are important for hygromycin B tolerance in *C. albicans*. To do so, we will use the CGD GO Term Finder to find shared biological processes for a set of genes whose mutation lowers resistance to hygromycin B.

- From the CGD home page ([www.candidagenome.org](http://www.candidagenome.org)), go to the Locus Summary page for the hygromycin B-sensitivity gene PMT6. Enter **PMT6** into the **search our site** box and click **GO**. On the next page, under ***Candida albicans* Search Results**, click on hyperlinked **1 Gene names (gene name/alias/ORF name)**.

**CGD Quick Search Result**

[Go to Advanced Search Page](#)

Below are the search results for your query, **pmt6**. If you would like to broaden your search, you may use one or more wildcard characters (\*) to indicate the location(s) where any text will be tolerated in your search term.

**General Search Results for : pmt6**

- 0 Gene Ontology terms (GO terms, synonyms)
- 0 Colleagues (by last name)
- 0 Authors (by last name, first initial)
- 0 PubMed ID
- 0 Gene Ontology ID
- 0 External ID

***Candida albicans* Search Results for : pmt6**

- 1 **Gene names (gene name/alias/ORF name)** 
- 0 Biochemical pathways
- 2 **General Descriptions**
- 0 **Phenotypes [Expanded Phenotype Search]**
- 2 **Ortholog or Best Hit**

***Candida glabrata* Search Results for : pmt6**

- 0 **Gene names (gene name/alias/ORF name)**

- From the PMT6 Locus Summary page, find other genes involved in hygromycin B sensitivity: scroll down to the **Mutant Phenotype** section and click on **resistance to Hygromycin B: decreased**

Mutant Phenotype		View all <a href="#">PMT6 Phenotype details and references</a>
<b>Classical genetics</b>		
heterozygous null	<ul style="list-style-type: none"> <li>▪ hyphal growth: decreased</li> <li>▪ hyphal growth: normal</li> <li>▪ resistance to Hygromycin B: decreased</li> <li>▪ viable</li> </ul>	
homozygous null	<ul style="list-style-type: none"> <li>▪ adhesion: decreased</li> <li>▪ biofilm formation: decreased</li> <li>▪ hyphal growth: absent</li> <li>▪ hyphal growth: decreased</li> <li>▪ hyphal growth: normal</li> <li>▪ chitinase distribution: normal</li> <li>▪ Als1p modification: normal</li> <li>▪ resistance to Hygromycin B: decreased</li> <li>▪ resistance to Calcofluor White: normal</li> <li>▪ resistance to Congo red: normal</li> </ul>	

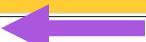
- On the **Phenotype Search Results** page, click on **Jump to: Analyze Gene List** above the table on the right (or simply scroll down to the bottom of the page). Click on **GO Term Finder** link.

Results: 1 - 30 of 42 records  
1 2

Jump to: top | [Results Table](#)

Analyze gene list: further analyze the gene list displayed above or download information for this list			
<b>Further Analysis:</b> <a href="#">GO Term Finder</a> Find common features of genes in list	<b>GO Slim Mapper</b> Sort genes into broad categories	<a href="#">View GO Annotation Summary</a> View all GO terms used to describe genes in list	
<b>Download:</b> <a href="#">Download All Search Results</a> Download data for the entire gene list in a tab-delimited file	<a href="#">Batch Download</a> Download selected information for entire gene list. Available information types include Sequence, Coordinates, Chromosomal Feature information, GO annotations, Phenotypes, and Ortholog or Best Hit.		

- With your own list of genes, you can access GO Term Finder from any CGD page by opening **GO** menu in the banner on top and clicking on **GO Term Finder**. Or you use this URL: <http://www.candidagenome.org/cgi-bin/GO/goTermFinder>
- The **CGD Gene Ontology Term Finder** has five steps (two optional) to specify your query. First, make sure that **Candida albicans** is selected as your species.
- Your input genes should be already entered. Alternatively, copy and paste your own list of genes into the text box (note: the more genes processed, the longer it takes). Choose **Process** as the ontology. Click the **Search** button to use the default settings.

**Step 1: Choose Species**  
Please select a species for genes in Query and Background sets :  

**Step 2: Query Set (Your Input)**

Enter Gene/ORF names:  
(separated by a return or a space)  



**OR** Upload a file of Gene/ORF names:  
 no file selected

**Step 3: Choose Ontology (Choose from only one of the 3 ontologies at a time)**

Process  
 Function  
 Component

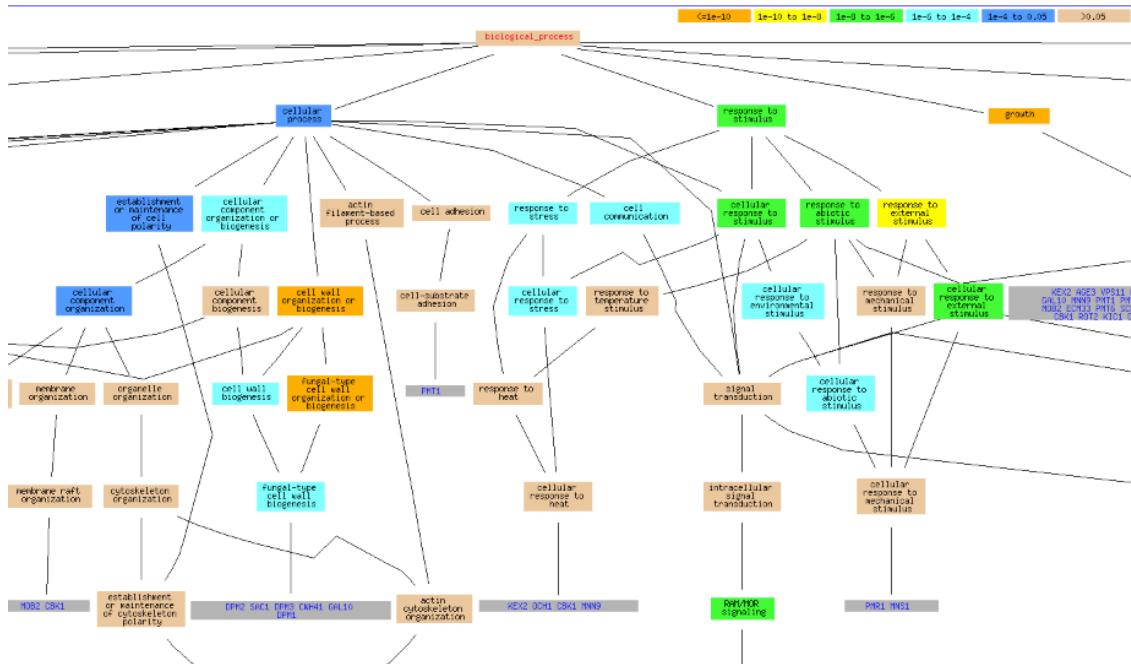
Search using default settings or use Step 4 and/or Step 5 below to customize your options.  
 
  


You can further customize your query in the next steps down the page:

- Optional Step 4 allows submitting a custom background set; use default set, all *C. albicans* genes in CGD
- Step 4 also allows restricting the search to specific feature types; use default settings
- Optional Step 5 allows selection of annotation methods, sources and evidence; leave all options checked

<b>Optional Step 4: Specify your background set of genes using the options below.</b>		
<input type="checkbox"/> Use default background set (all features in the database)	<b>OR</b> Enter Gene/ORF names: (separated by a return or a space)	<b>OR</b> Upload a file of Gene/ORF names: <input type="button" value="Choose File"/> no file selected
<b>Customize the gene list in the default or your specific background set (OPTIONAL)</b>		
Feature type <small>Default includes all feature types listed here</small> <div style="display: flex; justify-content: space-between;"> <input checked="" type="checkbox"/> ORF <input checked="" type="checkbox"/> allele <input checked="" type="checkbox"/> ncRNA <input checked="" type="checkbox"/> not in systematic sequence <input checked="" type="checkbox"/> pseudogene <input checked="" type="checkbox"/> rRNA <input checked="" type="checkbox"/> snRNA <input checked="" type="checkbox"/> snoRNA <input checked="" type="checkbox"/> tRNA         </div> <input type="button" value="Search"/> <input type="button" value="Clear All"/>		
<b>Optional Step 5: Refine the Annotations used for calculation</b>		
<small>You can use this option with Step 4. All Annotation Types are included by default.</small>		
<b>Select by Annotation Method</b>	Manually curated: <input checked="" type="radio"/> yes <input type="radio"/> no High-throughput: <input checked="" type="radio"/> yes <input type="radio"/> no Computational: <input checked="" type="radio"/> yes <input type="radio"/> no	
<b>Select by Annotation Source</b>	<input checked="" type="checkbox"/> CGD	
<b>Select by Evidence Codes:</b>	<input checked="" type="checkbox"/> IC <input checked="" type="checkbox"/> IDA <input checked="" type="checkbox"/> IEA <input checked="" type="checkbox"/> IEP <input checked="" type="checkbox"/> IGC <input checked="" type="checkbox"/> IGI <input checked="" type="checkbox"/> IMP <input checked="" type="checkbox"/> IPI <input checked="" type="checkbox"/> ISA <input checked="" type="checkbox"/> ISM <input checked="" type="checkbox"/> ISO <input checked="" type="checkbox"/> ISS <input checked="" type="checkbox"/> NAS <input checked="" type="checkbox"/> ND <input checked="" type="checkbox"/> RCA <input checked="" type="checkbox"/> TAS	
<input type="button" value="Search"/> <input type="button" value="Clear All"/>		

- Click **Search**. The input is checked and any genes that are not recognized as valid for the selected *Candida* species are rejected; click on **Proceed** in the following window.
- The results page displays the significant shared GO terms (or their parents) in both graphic and table form, within the set of genes associated with hygromycin B sensitivity entered on the previous page:
  - The graph shows the GO tree that includes terms used directly or indirectly in annotations for the genes in your list.
  - The terms are color-coded to indicate their statistical significance (p-value score), where the terms in orange have the highest likelihood of sharing meaningful relationships for the genes in your list.
  - Genes associated with the GO terms are shown in gray boxes, with links to their respective Locus Summary pages.



- The table below the graph lists each significant GO term, the number of times the GO term is used to annotate genes in the list, and the number of times that the term is used to annotate genes in the background set (all genes in *C. albicans* genome)

Terms from the Process Ontology					
Gene Ontology term	Cluster frequency	Background frequency	Corrected P-value	False discovery rate	Genes annotated to the term
cell wall organization or biogenesis   AmiGO	27 out of 41 genes, 65.9%	242 out of 6473 background genes, 3.7%	6.92e-27	0.00%	CAS4, CBK1, CWH41, DPM1, DPM2, DPM3, ECM33, GAL10, HRD1, HYM1, KIC1, MNN9, MNS1, MOB2, OCH1, PMR1, PMT1, PMT4, RHB1, ROT2, SAC1, SAP10, SAP9, SEC20, SFP1, SOG2, UBC7
fungal-type cell wall organization or biogenesis   AmiGO	25 out of 41 genes, 61.0%	213 out of 6473 background genes, 3.3%	6.17e-25	0.00%	CAS4, CBK1, CWH41, DPM1, DPM2, DPM3, ECM33, GAL10, HRD1, HYM1, KIC1, MNN9, MNS1, MOB2, PMR1, PMT1, PMT4, RHB1, ROT2, SAC1, SAP10, SAP9, SEC20, SOG2, UBC7
glycoprotein metabolic process   AmiGO	18 out of 41 genes, 43.9%	130 out of 6473 background genes, 2.0%	5.31e-18	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, MNS1, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, ROT2, SAC1, VRG4
macromolecule glycosylation   AmiGO	16 out of 41 genes, 39.0%	117 out of 6473 background genes, 1.8%	1.46e-15	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
protein glycosylation   AmiGO	16 out of 41 genes, 39.0%	117 out of 6473 background genes, 1.8%	1.46e-15	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
glycosylation   AmiGO	16 out of 41 genes, 39.0%	118 out of 6473 background genes, 1.8%	1.69e-15	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
glycoprotein biosynthetic process   AmiGO	16 out of 41 genes, 39.0%	121 out of 6473 background genes, 1.9%	2.57e-15	0.00%	CWH41, DPM1, DPM2, DPM3, GAL10, MNN14, MNN9, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, SAC1, VRG4
fungal-type cell wall organization   AmiGO	17 out of 41 genes, 41.5%	155 out of 6473 background genes, 2.4%	4.88e-15	0.00%	CAS4, CBK1, ECM33, HRD1, HYM1, KIC1, MNN9, MOB2, PMR1, PMT1, PMT4, RHB1, SAP10, SAP9, SEC20, SOG2, UBC7
external encapsulating structure organization   AmiGO	17 out of 41 genes, 41.5%	156 out of 6473 background genes, 2.4%	5.45e-15	0.00%	CAS4, CBK1, ECM33, HRD1, HYM1, KIC1, MNN9, MOB2, PMR1, PMT1, PMT4, RHB1, SAP10, SAP9, SEC20, SOG2, UBC7
cell wall organization   AmiGO	17 out of 41 genes, 41.5%	156 out of 6473 background genes, 2.4%	5.45e-15	0.00%	CAS4, CBK1, ECM33, HRD1, HYM1, KIC1, MNN9, MOB2, PMR1, PMT1, PMT4, RHB1, SAP10, SAP9, SEC20, SOG2, UBC7
filamentous growth   AmiGO	26 out of 41 genes, 63.4%	626 out of 6473 background genes, 9.7%	1.84e-14	0.00%	AGE3, CAS4, CBK1, CWH41, ECM33, GAL10, HYM1, KEX2, KIC1, MNN9, MNS1, MOB2, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, ROT2, SAC1, SAP9, SCH9, SOG2, VPS11, VRG4
growth   AmiGO	26 out of 41 genes, 63.4%	633 out of 6473 background genes, 9.8%	2.43e-14	0.00%	AGE3, CAS4, CBK1, CWH41, ECM33, GAL10, HYM1, KEX2, KIC1, MNN9, MNS1, MOB2, OCH1, PMR1, PMT1, PMT2, PMT4, PMT6, RHB1, ROT2, SAC1, SAP9, SCH9, SOG2, VPS11, VRG4

- Additional columns list the p-value, the false discovery rate (FDR), and a list of all the genes annotated, either directly or indirectly, to the term. FDR is an estimate of the percent chance that a particular GO term might actually be a false positive. It represents the fraction of the nodes with p-values as good or better than the node with this FDR that would be expected to be false positives.
- Explore the table. Based on the results, what biological processes are important for resisting the antibiotic action of hygromycin B in *C. albicans* cells?

## FungiDB: Performing GO Enrichment analysis

### Learning objectives:

- Perform a GO enrichment analysis
- Create a complex search strategy using both FungiDB and SGD

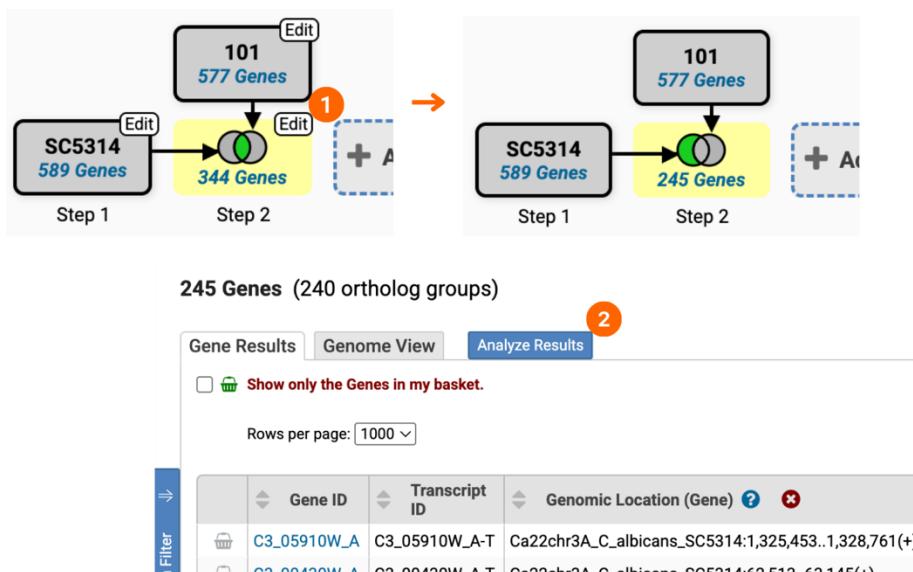
For this exercise, we will query *Candida albicans* transcriptomics by Kirchner et al. in 2019. The study focuses on the oropharyngeal candidiasis experimental model in mice, which was used to examine *C. albicans*' interaction with the host at mucosal surfaces in vivo. The study involved two strains of *C. albicans*: SC5314, a virulent lab strain, and the persistent strain 101. A persistent strain can resist medical treatment, often leading to chronic or recurrent infections.

- Using the strategy linked below, perform an enrichment analysis on *C. albicans* genes up-regulated in SC5314 cells only when they are exposed to mucosal surfaces.

The following strategy was created to identify genes up-regulated in both SC5314 and 101 strains in response to mucosal surfaces:

<https://fungidb.org/fungidb/app/workspace/strategies/import/25c2bed9444c4423>

1. Click on the Boolean operator in Step 2 and modify it to identify genes that are upregulated in SC5314 only.
2. Click on the “Analyze Results” tab for enrichment analysis options.
3. Deploy GO enrichment analysis by clicking the “Gene Ontology Enrichment” button.
4. Perform GO enrichment analysis (Biological Process) using default selection criteria.



The enrichment analysis tools include Gene Ontology, Metabolic Pathway, and Word Enrichment tools.

The three types of analysis apply Fisher's Exact test to evaluate ontology terms, overrepresented 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. P-values corrected for multiple testing are provided using the Benjamini-Hochberg false discovery rate and Bonferroni methods.

GO enrichment analysis can be performed on the following ontology groups:

- Molecular function,
- Biological processes,
- Cellular component.

Other parameters limit users' analysis to 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 curation. The default P-value is set to 0.05 but can be adjusted manually.

When the GO Slim option is chosen, the genes of interest and the background are limited to GO terms that are part of the generic GO Slim subset.

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
G0:0042273	ribosomal large subunit biogenesis	558	67	12.0	3.03	4.20	1.08e-17	1.68e-14	1.68e-14
G0:0000470	maturation of LSU-rRNA	440	55	12.5	3.16	4.20	3.31e-15	2.59e-12	5.17e-12
G0:0000463	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	432	53	12.3	3.10	4.07	2.62e-14	1.37e-11	4.10e-11

The results table includes several additional statistical measurements:

- **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 whether the odds of the GO term appearing in the list of interest are the same as those for the background list.
- **P-value** - Assumptions under a null hypothesis, the probability of getting a result that is equal to or greater than what was observed.
- **Benjamini-Hochberg 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.

The GO enrichment table can be opened in Revigo, viewed as a word cloud (produced via the GO Summaries R package) or downloaded.

Notice that 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 (linked in blue).

## 5. Examine GO enrichment analysis results. What kinds of GO terms are enriched?

Note: you can sort genes in your results using the sort options within a column:

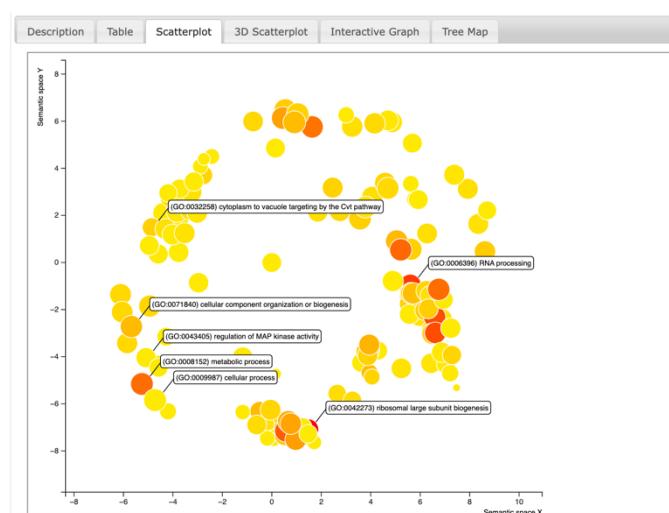
Genes in your result with this term	Percent of bkgd genes in your result
202	7.2
184	4.3
181	4.5

## 6. Visualize the Revigo results by clicking the Revigo button above the results table and leaving other parameters at default. Click the Start Revigo button below the results set and then select Scatterplot.

The bubble colour corresponds to the user-provided p-value.

The bubble size represents the frequency of the GO term in the underlying database.

The table tab provides a detailed overview of the GO terms, P-values and also parent GO terms used to describe a group of related GO terms



More information available here:

<http://geneontology.org/docs/ontology-relations/>

### **Optional exercise. Creating queries across FungiDB and SGD.**

Use case: During a genetic screen in *Lomentospora prolificans*, you identified several exciting genes, including jhhlp\_004726, a hypothetical protein. Use FungiDB and SGD records to learn more about this gene.

#### **1. Navigate to jhhlp\_004726 in FungiDB and examine available records.**

[https://fungidb.org/fungidb/app/record/gene/jhhlp\\_004726](https://fungidb.org/fungidb/app/record/gene/jhhlp_004726)

- Run an InterPro search and a GPI anchor prediction tool. What did you learn about this protein?

Hint: InterPro and GPI search tools can be found in the gene record page's Protein features and properties section.

#### **2. Export orthologs of this gene and carry over *S. cerevisiae* gene IDs into SGD.**

- Click on the Download gene link at the top of the gene record page and select the option to export orthologs, as shown below.

The screenshot shows the FungiDB gene record page for jhhlp\_004726. At the top, there are three buttons: 'Add to basket' (with a plus icon), 'Add to favorites' (with a star icon), and 'Download Gene' (with a download icon). Below these is the gene ID 'jhhlp\_004726 hypothetical protein'. A large orange arrow points from the 'Download Gene' button down to the 'Choose Tables' section. In this section, there is a dropdown menu set to 'ortholo' and two checkboxes: 'Orthology and synteny' (which is checked) and 'Orthologs and Paralogs within VEuPathDB' (which is also checked). Other options in the 'Choose Tables' section include 'select only these', 'add these', and 'clear these'.

The exported text file can be opened with Excel.

- Sort genes on the [Organism].
- Copy GeneIDs for *S. cerevisiae* (e.g., YDR144C).
- Navigate to the SGD gene lists search to create a new upload.
- Paste *S. cerevisiae* orthologs for jhhlp\_004726 in the form:  
<https://www.yeastgenome.org/locus/YDR144C>.



Create a new list

Select the type of list to create and then enter your identifiers or upload them from a file.

**i** • Separate identifiers by a comma, space, tab or new line  
• Qualify any identifiers that contain whitespace with double quotes like so: "even skipped"

List type  
Gene

Organism  
*S. cerevisiae*

Identifiers are case sensitive

YDR144C  
YGL259W  
YLR121C  
YIL015W  
YLR120C

- Give your list a name such as 'Yeast orthologs 1' and save it.
- Click on the GeneIDs to examine *S. cerevisiae* genes and answer the following questions:
  - What is the function of MKC7 ([YDR144C](#)) in *S. cerevisiae*?
  - Does it encode a protein with enzymatic activity?
  - Where in the cell does the protein execute its function? What biological process?

Hint: see the **GO Annotation** section under the ‘Data’ on the locus page.

### 3. Find known genetic interactions for MKC7.

Functional relationships between genes and pathways can sometimes be revealed by examining genetic interactions between two or more genes. Genes are described as having a genetic interaction if the simultaneous mutation of both genes produces an unexpected phenotype, given the phenotypes of the single mutants.

- In SGD, find the MKC7 locus page and navigate to the **Interactions** section on the left, listed in the Quick Links panel near the top. The interactions are divided into physical and genetic interactions, as shown in the tables below the summary.
- Filter the **Genetic Interactions** table on “synthetic.” This will show only the genetic interactions that produce some sort of synthetic growth defect, haploinsufficiency, or lethality.

SUMMARY | SEQUENCE | PROTEIN | GENE ONTOLOGY | PHENOTYPE | **INTERACTIONS** | REGULATION | EXPRESSION | LITERATURE | HOMOLOGY

### MKC7 / YDR144C Interactions ⓘ

**Summary:** The mkc7 null mutant is viable; the null mutant of paralog *yps1* is viable; the mkc7 *yps1* double mutant has osmoremedial heat sensitivity, increased sensitivity to caffeine, congo red, caspofungin, calcofluor white, growth at low pH and a secretion defect; a mkc7 *yps1* *yps3* triple mutant has severe osmoremedial heat sensitivity and decreased tolerance to high salt.

Source: All physical and genetic interaction annotations listed in SGD are curated by BioGRID.

Analyze

Physical    Genetic    Intersection    All

### Genetic Interactions ⓘ

Genetic Interactions 121 entries for 102 genes

Interactor	Allele	Assay	Annotation	Action	Phenotype	SGA score	P-value	Reference
ACT1	Synthetic Haploinsufficiency	high-throughput	Hit					Haarer B, et al. (2007) PMID:17167106
GIM5	Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased Mutant Type: unspecified				Tong AH, et al. (2004) PMID:14764870

- Click on the **Download** button, which is located under the results table, and save this gene list. *Rename the file to synthetic.txt*.

*Note: Rename the file to synthetic.txt so that we can find it easily later.*

- Click on the **Analyze** button, then on **GO Term Finder**.
- Run a **process** enrichment for the MKC7 genetic interaction genes.

*Hint: GO Term Finder finds common Gene Ontology (GO) annotations between genes. To run a Biological Process enrichment, select the Process button as shown below, then submit the form. More ways to customize your GO Term Finder query can be found in the GO Term Finder exercise.*

**Step 2. Choose Ontology**

**Pick an ontology aspect:**

Process       Function       Component

Search using default settings or use Step 3 and/or Step 4 below to customize your options.

- Scroll down the results page to see the table of enriched biological processes. What kind of processes are associated with the genes we analyzed? What do these results suggest about MKC7's functional relationships in the cell?
- Click on any of the genes shown for a biological process of interest to visit the gene's page on SGD. Use the gene page to uncover how the respective gene is involved in the biological process you were interested in.

Result Table

Terms from the Process Ontology of gene\_association.sgd with p-value <= 0.01

Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	FDR	False Positives	Genes annotated to the term
tubulin complex assembly	3 of 9 genes, 33.3%	10 of 7166 genes, 0.1%	1.96e-05	0.00%	0.00	YML094W, YLR200W, YGR078C
protein folding	4 of 9 genes, 44.4%	121 of 7166 genes, 1.7%	0.00109	0.00%	0.00	YML094W, YLR200W, YKL117W, YGR078C
peptide pheromone maturation	2 of 9 genes, 22.2%	9 of 7166 genes, 0.1%	0.00603	0.67%	0.02	YNL238W, YLR120C
chaperone-mediated protein complex assembly	2 of 9 genes, 22.2%	9 of 7166 genes, 0.1%	0.00603	0.50%	0.02	YKL117W, YLR200W
fungal-type cell wall organization	4 of 9 genes, 44.4%	205 of 7166 genes, 2.9%	0.00878	0.40%	0.02	YHR079C, YLR120C, YLR121C, YFL039C

Now, let's go back to the file of MKC7 "synthetic" genetic interactors we downloaded earlier and find the orthologs of these genes in *Lomentospora prolificans*.

- Open this file in Excel and copy the Gene IDs in the **Interactor Systematic Name** column (not including the header)

Interactor	Interactor Sys	Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR144C	ACT1	YFL039C	Genetic	Synthetic Ha	high-through
MKC7	YDR144C	GIM5	YML094W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	IRE1	YHR079C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	KEX2	YNL238W	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	PAC10	YGR078C	Genetic	Synthetic Let	high-through
MKC7	YDR144C	SBA1	YKL117W	Genetic	Synthetic Let	high-through
MKC7	YDR144C	YKE2	YLR200W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	YPS3	YLR121C	Genetic	Synthetic Let	manually cur

- Revisit FungiDB and initiate the List of IDs search query  
The query can be deployed from the “Searches” menu at the top of the “Search for Genes” section on the main page.

The screenshot shows the FungiDB search interface. At the top, there is a navigation bar with links: Searches, Tools, My Workspace, Data, About, Help, Contact. Below the navigation bar is a search bar containing the word "gene". Underneath the search bar, there is a section titled "Genes" with the sub-section "Annotation, curation and identifiers". This section includes links for "List of IDs" and "User Comments". Below this, there is another section titled "Epigenomics" with a link "Find a gene by its ID".

- Paste the list of Gene IDs with the “synthetic” genetic interactions with MKC7 into the FungiDB query and click the **Get Answer** button.

The screenshot shows the "Identify Genes based on List of IDs" search form. At the top, there are buttons for "Configure Search", "Learn More", and "View Data Sets Used". Below these is a "Reset values to default" button. The main input area is labeled "Gene ID input set" with a red arrow pointing to it. There are two options: "Enter a list of IDs or text:" and "Upload a text file:". The "Enter a list of IDs or text:" option is selected, and a text box contains the following gene IDs:  
YNL238W  
YGR078C  
YLR177W  
YLR200W  
YLR120C  
YLR121C

Below this, there are other options: "Upload from a URL:", "Copy from My Basket:", and "Copy from My Strategy:". At the bottom right of the form is a "Get Answer" button with a red arrow pointing to it.

**9 Genes (8 ortholog groups)** [Revise this search](#)

[Gene Results](#) [Genome View](#) [Analyze Results](#)

Rows per page: 1000 [Download](#) [Send to...](#) [Add Columns](#)

	Gene ID	Transcript ID	Gene Name or Symbol	Organism	Genomic Location (Gene)	Product Description
<a href="#">YFL039C</a>	YFL039C-t26_1	ACT1	<i>Saccharomyces cerevisiae</i> S288C	BK006940:53,260..54,696(-)	actin	
<a href="#">YML094W</a>	YML094W-t26_1	GIM5	<i>Saccharomyces cerevisiae</i> S288C	BK006946:82,275..82,849(+)	Gim5p	
<a href="#">YHR079C</a>	YHR079C-t26_1	IRE1	<i>Saccharomyces cerevisiae</i> S288C	BK006934:258,244..261,591(-)	bifunctional endoribonuclease/protein	

- Find orthologs in *Lomentospora prolificans*.

Click the “Add a step” button to **Transform** the list into related records. Select the option to transform into **orthologs**, then use the search bar to filter on *Lomentospora prolificans* and **Run Step**.

**Add a step to your search strategy**

**Transform 9 Genes into...**

**Orthologs**

**Add a step to your search strategy**

Your Genes from Step 1 will be converted into Orthologs

**Organism**

Note: You must select at least 1 values for this parameter.  
1 selected, out of 163

add these | clear these | select only these

Lom

Fungi

Ascomycota

Sordariomycetes

Microascales

Lomentospora prolificans JH-5317

**Syntenic Orthologs Only?**

no

**Run Step**

**Gene Results** | **Genome View** | **Analyze Results**

Rows per page: 20

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
jhhlp_002587	jhhlp_002587- t41_1	Lomentospora prolificans JHH-5317	NLAX01000008:3,258,120..3,260,362(-)	hypothetical protein	YFL039C	OG6_100127	0	239
jhhlp_004481	jhhlp_004481- t41_1	Lomentospora prolificans JHH-5317	NLAX01000010:4,766,898..4,769,585(+)	hypothetical protein	YNL238W	OG6_100362	0	167
jhhlp_004364	jhhlp_004364- t41_1	Lomentospora prolificans JHH-5317	NLAX01000010:4,180,492..4,181,475(-)	hypothetical protein	YKL117W	OG6_101574	0	157
	jhhlp_007003	Lomentospora						

How many interacting *S. cerevisiae* genes have a hypothetical protein ortholog in *Lomentospora prolificans*? Can you find jhhlp\_004726 amongst these genes?

### Strategy URL:

<https://fungidb.org/fungidb/app/workspace/strategies/import/c0978bdb48a8392d>

Glycosylphosphatidylinositol (GPI)-anchored proteins are involved in cell wall integrity and cell-cell interactions, and perturbations in GPI biosynthesis lead to hypersensitivity to host defences. Given the accumulated biological information we uncovered at SGD and FungiDB, summarize your predictions about the hypothetical *L. prolificans* protein jhhlp\_004726.

- What is the likely jhhlp\_004726 ortholog in *S. cerevisiae*?
  - Is this gene a GPI protein in yeast?
- Do you have sufficient information to think the hypothetical gene in *L. prolificans* may be a putative GPI-anchor protein?
- How many “synthetic” genetic interactors exist in SGD for MKC7 in yeast?
  - What GO terms were enriched in biological processes associated with MKC7 interactors in *S. cerevisiae*?
  - How many orthologs of these genes are found in *L. prolificans*?
  - Why do you think the number of genes varies between *S. cerevisiae* and *L. prolificans*?

### Additional resources:

More info on Fischer's exact test:

<https://www.biostathandbook.com/fishers.html>

Some more info about Odds ratios:

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/>

False discovery rates and P value correction:

<http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/>

## SGD Variant Viewer

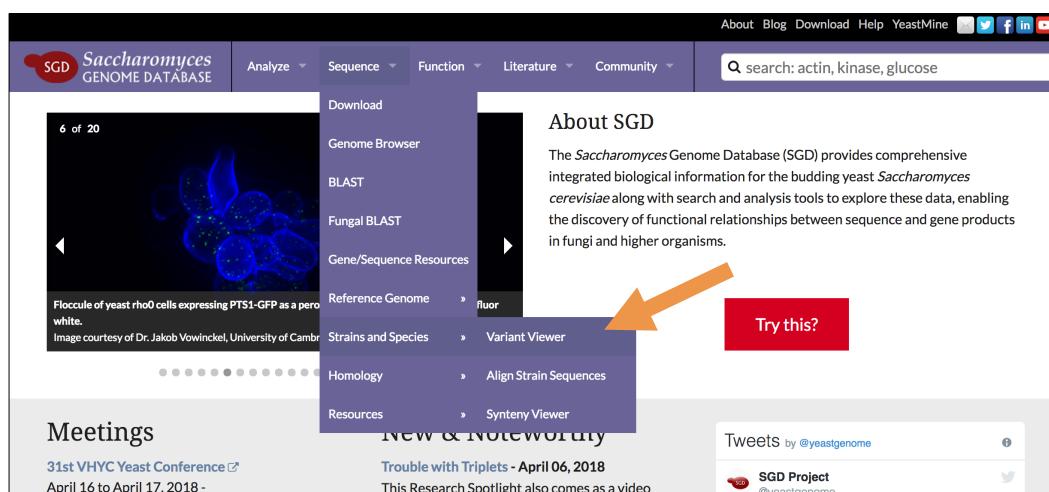
SGD's Variant Viewer (<https://yeastgenome.org/variant-viewer>) is an open-source web application that compares nucleotide and amino acid sequence differences between 12 common *S. cerevisiae* laboratory strains. For a given open reading frame, Variant Viewer breaks down the position and nature of any strain-specific sequence differences relative to the reference strain S288C. When used at a multi-gene level, it also provides a matrix of alignment scores that enables quick identification of genes with higher or lower variation.

Variant Viewer can be used to probe the genetic differences between *S. cerevisiae* strains that give rise to their unique phenotypes. For example, while haploid S288C cells exhibit an axial budding pattern, diploid cells exhibit a bipolar budding pattern. On the other hand, strain W303 shows bipolar bud site selection in both haploid and diploid cells.

In this exercise, we will use Variant Viewer to find out what genetic differences between Sigma1278b and S288C explain why they differ in their ability to form pseudohyphae.

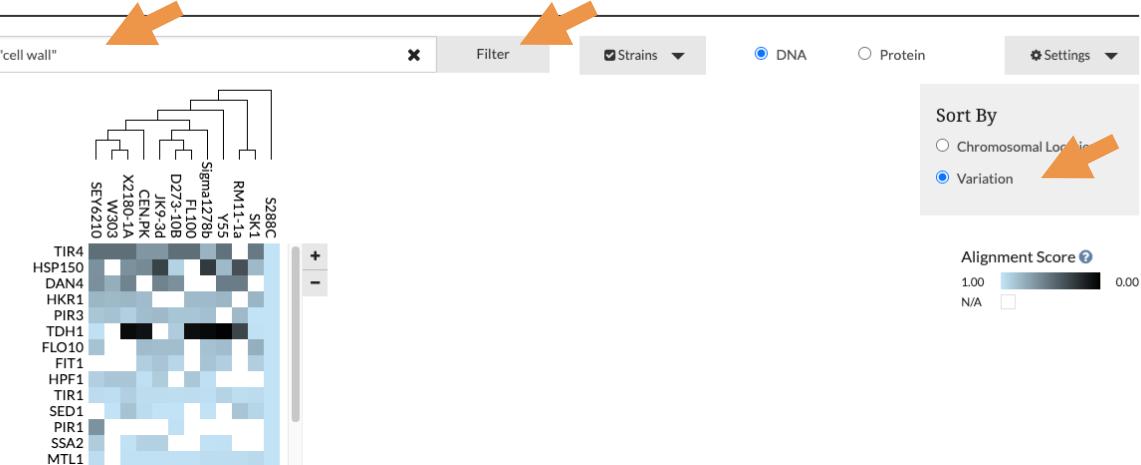
### S288C vs. Sigma1278b: Cell Wall

- Open the SGD home page ([www.yeastgenome.org](http://www.yeastgenome.org)), open the Sequence tab on top of the page, then select Strains and Species followed by Variant Viewer from the pull-down menus. Or just type in the URL: [yeastgenome.org/variant-viewer](https://yeastgenome.org/variant-viewer)



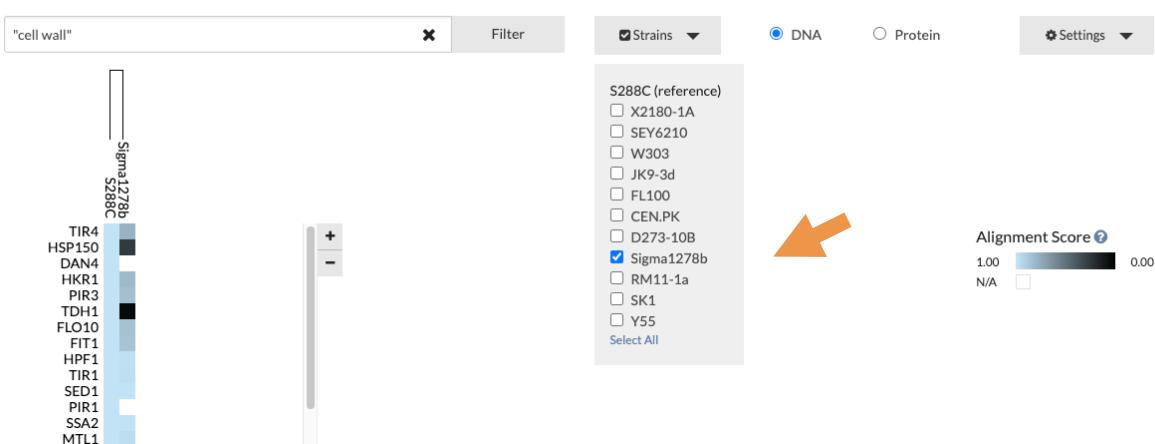
- The **Filter** box accepts one or more genes, as well as Gene Ontology (GO) terms. Because we are interested in genes involved in cell wall development, enter the GO term “**“cell wall”** (don't forget the quotation marks) and select **variation** in the settings pull-down, then click **Filter**.

## Variant Viewer



- The **matrix**, shown on the left, will have changed to only include the genes that localize to cell walls.
  - This matrix enables you to visualize high-level differences in multiple genes relative to strain S288C. Each square in the matrix corresponds to one of the twelve strains in Variant Viewer, shown at the top, and to an open reading frame, shown on the left.
  - The color of each square indicates how similar the sequence is relative to strain S288C. As indicated on the Alignment Score figure on the right, lighter shades of blue indicate high sequence similarity whereas darker shades indicate more dissimilarity. Note that if the square is white, it means a comparison could not be made.
- Next, we will want to make the matrix display only info for the strains we are interested in (S288C and Sigma1278b). Open the **Strains** pull-down menu, press Deselect All, then re-select Sigma1278b.

## Variant Viewer



- Click on **PIR3** (O-glycosylated covalently bound cell wall protein) and in the sequence window select **Protein**. Scroll with your mouse along the green bar of sequence to see what the changes between strains are due to. Find the deletion beginning at Chr X1144715 and compare the protein sequences below.



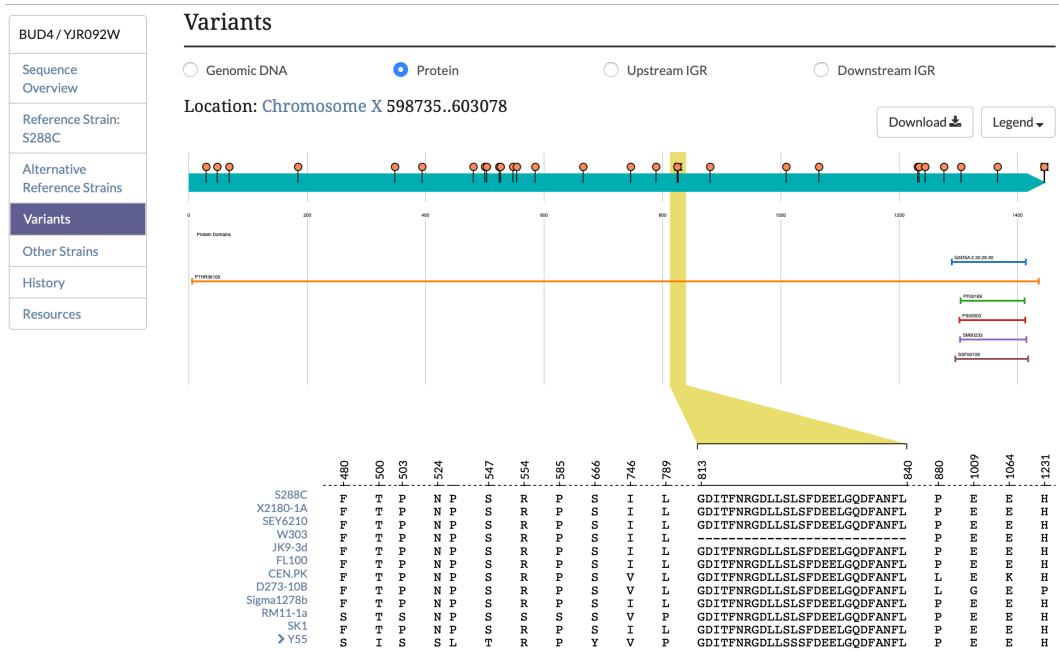
- Now that we have identified that a deleted section of protein in a cell wall protein of Sigma1278b, we have a clue as to why this strain behaves differently from S288C. To examine PIR3 more closely, click the name in the upper left of the page to go to the locus summary page. From the PIR3 Locus Summary page, you can see in the Description that this protein is known to vary between strains.
- In the list of references below, you'll find papers referring to the role of this cell wall protein (and its relations) in heat shock, response to toxins, and cell wall integrity. The differences in this protein between strains might contribute to variations in behavior, such as differences in pseudohyphal growth for Sigma1278b relative to S288C

## References ⓘ 9

- Toh-e A, et al. (1993)** Three yeast genes, PIR1, PIR2 and PIR3, containing internal tandem repeats, are related to each other, and PIR1 and PIR2 are required for tolerance to heat shock. *Yeast* 9(5):481-94 PMID: 8322511  
SGD Paper DOI full text PubMed
- Yun DJ, et al. (1997)** Stress proteins on the yeast cell surface determine resistance to osmotin, a plant antifungal protein. *Proc Natl Acad Sci U S A* 94(13):7082-7 PMID: 9192695  
SGD Paper DOI full text PMC full text PubMed
- Doolin MT, et al. (2001)** Overlapping and distinct roles of the duplicated yeast transcription factors Ace2p and Swi5p. *Mol Microbiol* 40(2):422-32 PMID: 11309124  
SGD Paper DOI full text PMC full text PubMed
- Porter SE, et al. (2002)** The yeast pafl-rRNA polymerase II complex is required for full expression of a subset of cell cycle-regulated genes. *Eukaryot Cell* 1(5):830-42 PMID: 12455700  
SGD Paper DOI full text PMC full text PubMed
- Jung US and Levin DE (1999)** Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Mol Microbiol* 34(5):1049-57 PMID: 10594829  
SGD Paper DOI full text PubMed

## Variant Viewer: Sequence Tab

- Variant Viewer is also embedded in the Sequence tab of every gene page, with the data for the gene already pre-loaded from the results of the Variant Viewer search. This allows you to look at the variant information for a gene without starting from the tool's entry page.



# FungiDB: SNPs and Population Genetics

## Learning Objective:

- Investigate SNP datasets using the following searches:
  - o SNP characteristics,
  - o SNPs between groups of isolates,
- Identify aneuploidy with the copy number variations search.

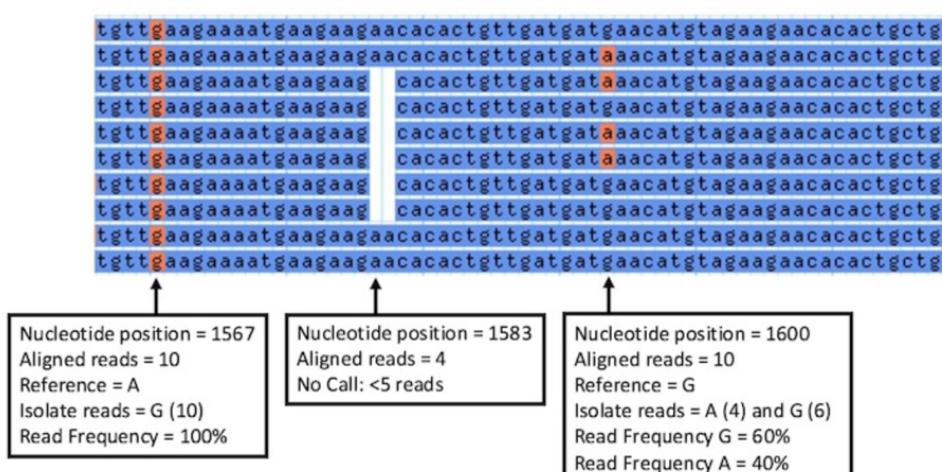
Single nucleotide polymorphisms (SNPs) are genetic variations that may or may not have an impact on the function of a gene. Most SNPs do not affect gene function. However, some SNPs that lead to a change in the amino acid or a premature stop codon (nonsense) can directly affect protein function. SNPs that do not occur within genes are non-coding, but they may still influence processes such as splicing, mRNA stability and transcription. SNPs are useful for identifying similarities and differences between isolates or groups of isolates. They can also be used to identify genes that are under evolutionary pressure, either to remain unchanged (purifying selection) or to change (diversifying or balancing selection).

## Read Frequency Threshold:

The parameter lets you control the quality of the sequencing data and the confidence of the SNP calls. Read Frequency Threshold is a particularly important parameter when dealing with diploid (or aneuploid) organisms since a read frequency of ~50% is expected for heterozygous SNPs.

Each isolate's sequencing reads are aligned to a reference genome and then each nucleotide position with 5 or more aligned reads is examined. A base call is made if the aligned reads meet your Read Frequency Threshold. For example, *Isolate X* has 10 aligned reads at nucleotide position 1600. If 6 reads are G and 4 reads are A, the read frequency is 60% for the G call and 40% for A. Running this search with the Read Frequency Threshold set to 80% will prevent a base call and consequently exclude *Isolate X* when returning SNPs for nucleotide position 1600. Running the search with the Read Frequency Threshold set to 60% will bring back a G for this isolate and a 40% threshold will return two calls (both G and A) at this position.

## Isolate X aligned sequencing reads



### Minor allele frequency:

The minor allele frequency refers to the least common base call for a single nucleotide position across all isolates. The default setting for this parameter is 0% and returns all SNPs - instances where at least one isolate has a base call that differs from reference. Increase the Minor allele frequency to ensure that SNPs returned by the search are shared by a larger percentage of isolates in your group.

### Isolate consensus sequences aligned to reference genome.

reference	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTCTACTG
303.1	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTCTACTG
309.1	TGATAATNCT GGTTTTGTA CTCCACTTCC CAGTGCTTCA TTTCTACTG
RV_3600	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTCTACTG
RV_3606	TGATAATNCT GGTTTTGTA CTCCACTTCC CAGTGCTTCA TTTCTACTG
RV_3610	TGATGATTCCT GGTTTTGTA CTCCACTTCC CAGTGCTTCA TTTCTACTG
SenT119.09	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTCTACTG
SenT123.09	TGATRATTCCT GGTTTTGTA CTCCACTTCC CAGTGCTTCA TTTCTACTG
SenT140.08	TGGTGATACT GGTTTTGTA CTCCACTTCC CGGTGCTTCA TTTCTACTG
SenT142.09	TGGTGATACT GGTTTTGTA CTCCACTTCC CAGTGCTTCA TTTCTACTG
SenT175.08	TGGTGATACT GGTTTTGTA CTCCACTTCT CGGTGCTTCA TTTCTACTG

Reference = G  
6 isolate seq = G  
4 isolate seq = A  
% with base call = 100  
Minor allele = A  
Minor allele freq = 40% (4/10)

Reference = A  
6 isolate seq = A  
2 isolate seq = T  
2 isolate seq = N (no call)  
% with base call = 80  
Minor allele = T  
Minor allele freq = 25% (2/8)

Reference = G  
5 isolate seq = G  
5 isolate seq = A  
% with base call = 100  
Minor allele = G or A  
Minor allele freq = 50% (5/10)

### Percent isolates with a base call:

Sometimes an isolate does not have a base call at a certain nucleotide position because the Read Frequency Threshold was not met or because there were less than 5 aligned sequencing reads for that nucleotide position. In this case, an SNP can be returned by the search based on a subset of your isolate group. The 'Percent isolates with a base call' parameter defines the fraction of isolates that must have a base call before an SNP is returned for that nucleotide position. The default setting for this parameter is 80%, or 8 out of 10 isolates in your group must have a base call for an SNP to be returned by the search. The higher this parameter, the more likely the SNP is to be high quality as regions difficult to align or difficult to sequence will tend to have a lower percentage of calls since the coverage and/or quality will be lower in that region.

#### A. Identify putative nuclear effectors with at least 1 non-synonymous SNP in *Pyricularia oryzae*.

In this exercise, you will analyze a collection of *P. oryzae* isolates obtained from rice plants in different parts of Africa to identify genes with at least one non-synonymous SNP that also exhibit characteristics of nuclear effectors.

*Pyricularia oryzae* (also known as *Magnaporthe oryzae*) is a pathogen that affects rice crops, causing a severe disease known as rice blast. During the infection process, *P. oryzae* and other plant pathogens use various effectors to manipulate the plant's immune system. Nuclear effectors are a type of effector that contain both a signal peptide and a DNA-binding domain (e.g. IPR007219 or IPR009071).

- **Identify genes with at least 1 non-synonymous SNP.**
  1. Deploy the “SNP characteristics” search.
  2. Select *Pyricularia oryzae* 70-50 from the organism tree.
  3. In the Data Set section, select the datasets where isolates were collected in Zambia and other African fields.
  4. Set the “SNP Class” parameter to “Non-Synonymous”.
  5. Choose to identify genes with at least 1 non-synonymous SNPs and click on the “Get Answer” button.

**1**

**2**

**3**

**4**

**5**

**Identify Genes based on SNP Characteristics**

**Organism**  
1 selected  
Pyricula oryzae 70-15 [Reference]

**Set of Samples**  
83 Set of Samples Total  
41 of 83 Set of Samples selected  
Data Set  
Find a variable  
Sample type, Fungal organism, Data Set, Fungal strain, Sample collection, Geographic location, Sample source  
Keep checked values at top  
83 (100%) of 83 Set of Samples have data for this variable

Data Set	Remaining Set of Samples	Set of Samples	Distribution	%
Pyricularia oryzae 70-15 Genome Sequence and Annotation	1 (1%)	1 (1%)	1	(100%)
SNP calls on WGS of GE10A2 and GE12B isolates	2 (2%)	2 (2%)	2	(100%)
SNP calls on WGS of Magnaporthe field-isolates	13 (16%)	13 (16%)	13	(100%)
SNP calls on WGS of Pyricularia oryzae isolated from Bangladesh in 2016 and 2017	23 (28%)	23 (28%)	23	(100%)
SNP calls on WGS of Pyricularia oryzae isolates from different hosts	3 (4%)	3 (4%)	3	(100%)
SNP calls on WGS of Pyricularia oryzae isolates from Zambia	13 (16%)	13 (16%)	13	(100%)
SNP calls on WGS data of Pyricularia oryzae isolates from Africa	28 (34%)	28 (34%)	28	(100%)

**Read frequency threshold**  
80%

**Minor allele frequency >=**  
0

**Percent isolates with a base call >=**  
20

**SNP Class**  
Non-Synonymous

**Number of SNPs of above class >=**  
1

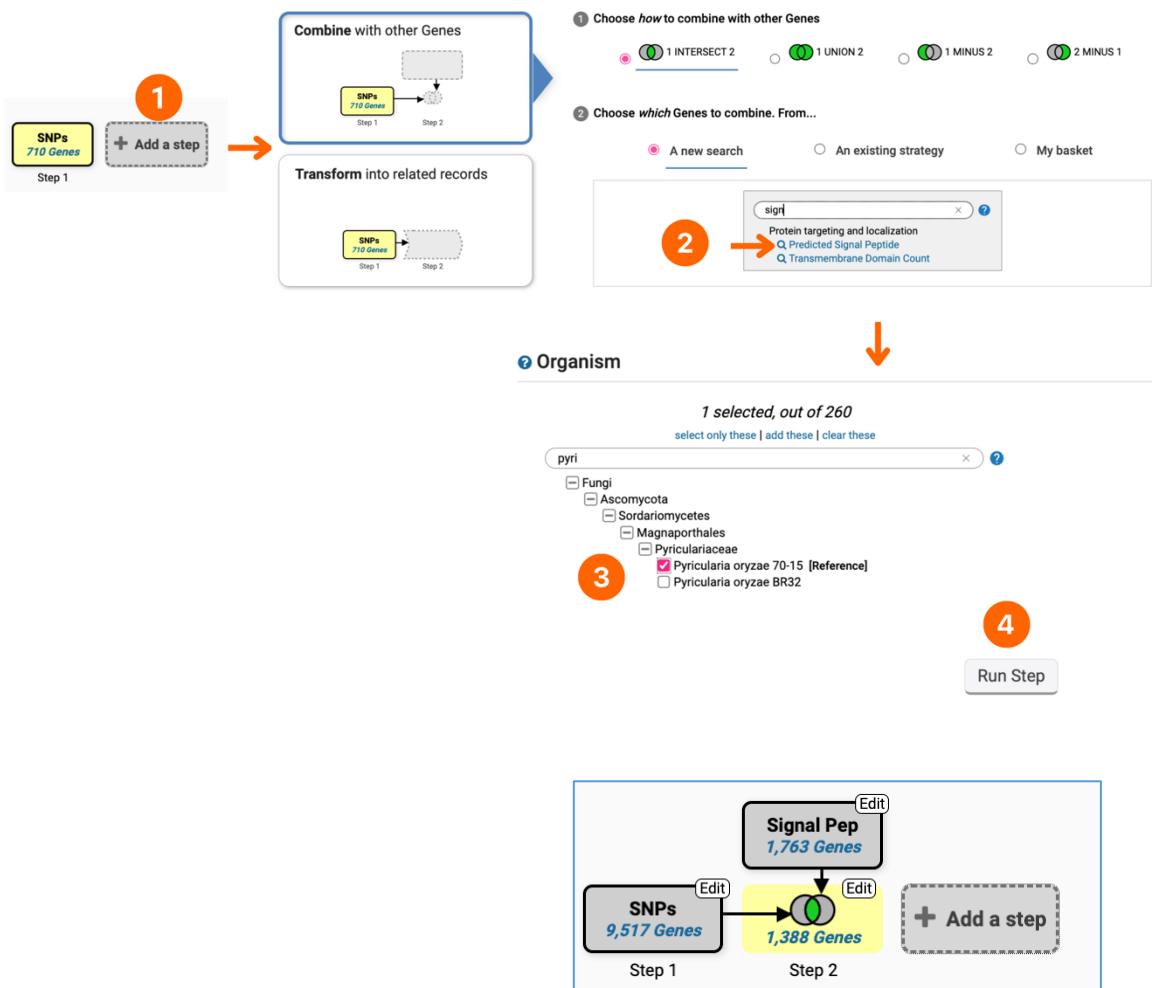
**Step 1**

**Edit**  
**SNPs 9,517 Genes**

**Add a step**

- **Identify putative nuclear effectors.**

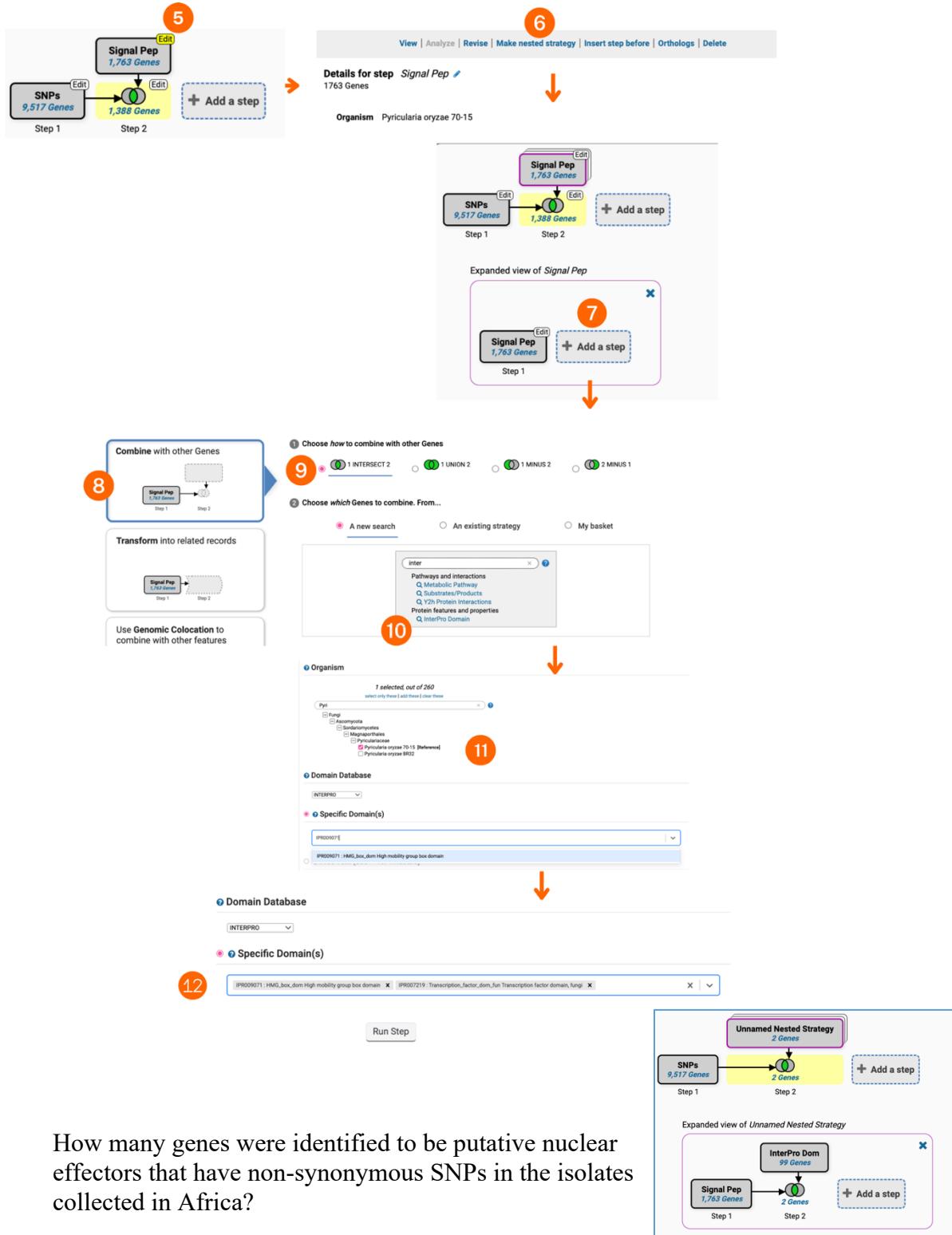
1. Click on the “Add a Step” button.
2. Use the “Combine with Other Genes” option to deploy the “Predicted Signal Peptide” search.
3. Set the genome to *Pyricularia oryzae* 70-50.
4. Click on the “Run Step” button.



The strategy returns genes with at least 1 SNP and a predicted signal peptide domain. How can we identify genes with at least 1 SNP and a predicted signal peptide domain AND a DNA-binding domain (IPR007219 or IPR009071)? (Hint: you can do this with a nested strategy as described below).

5. Hover over the “Signal Pep” search box and click on the “Edit” option.
6. Select the “Make nested strategy” option at the top.
7. Click on the “Add a Step” button within the “Expanded view of Signal Pep” (nested) strategy (screenshot below).
8. Select the “Combine with other Genes” search.
9. Set the Boolean operator to “1 intersect 2”.
10. Deploy the “InterPro Domain” search.

- Set the genome to *Pyricularia oryzae* 70-50 and set the “Domain database” to InterPro and enter and select the following DNA binding domains from the dropdown menu: IPR007219, IPR009071.
- Click on the “Run Step” once both domains are selected.



## B. Explore SNP records in JBrowse

Coccidioidomycosis, also known as Valley fever, is a disease caused by two closely related species of fungi – *Coccidioides immitis* (*C. immitis*) and *Coccidioides posadasii* (*C. posadasii*). The disease is associated with high morbidity and mortality that affect tens of thousands of people every year. These two fungal species are found in several regions in the Western Hemisphere, but recent studies suggest that their geographic range is expanding.

Start by clicking on the strategy linked below. This strategy uses the “Differences Between Two Groups of Isolates” search, which distinguishes SNPs between two groups of *C. posadasii* str Silveira isolates (A and B) collected in the Caribbean (group A) and western hemispheres (group B):

<https://fungidb.org/fungidb/app/workspace/strategies/import/7dc0e63eb2679bd9>

SNP Id	Location	Gene ID	Position in protein	Set A Major Allele	Set A Major Pct	Set A Major Product	Set B Major Allele	Set B Major Pct	Set B Major Product
NGS_SNP.GL636538.9073	GL636538: 9,073	N/A	N/A	C	100	-	G	90	-
NGS_SNP.GL636538.8514	GL636538: 8,514	N/A	N/A	G	100	-	C	100	-
NGS_SNP.GL636538.3960	GL636538: 3,960	N/A	N/A	C	100	-	T	95.7	-
NGS_SNP.GL636537.6464	GL636537: 6,464	N/A	N/A	A	100	-	G	100	-

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

SNP Id	Location	Gene ID	Position in protein	Set A Major Allele	Set A Major Pct	Set A Major Product	Set B Major Allele	Set B Major Pct	Set B Major Product
NGS_SNP.GL636538.9073	GL636538: 9,073	N/A	N/A	C	100	-	G	90	-
NGS_SNP.GL636538.8514	GL636538: 8,514	N/A	N/A	G	100	-	C	100	-
NGS_SNP.GL636538.3960	GL636538: 3,960	N/A	N/A	C	100	-	T	95.7	-
NGS_SNP.GL636537.6464	GL636537: 6,464	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636537.4384	GL636537: 4,384	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636537.1402	GL636537: 1,402	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636536.8746	GL636536: 8,746	N/A	N/A	A	100	-	G	100	-
NGS_SNP.GL636536.6075	GL636536: 6,075	CPSG_10217	15	T	100	E	C	92.3	G
NGS_SNP.GL636536.532	GL636536: 532	N/A	N/A	T	100	-	A	100	-
NGS_SNP.GL636536.4473	GL636536: 4,473	N/A	N/A	T	100	-	C	92.3	-
NGS_SNP.GL636536.1587	GL636536: 1,587	CPSG_10216	738	T	100	T	C	93.3	A
NGS_SNP.GL636536.1541	GL636536: 1,541	CPSG_10216	753	G	100	A	A	95.8	V
NGS_SNP.GL636536.13558	GL636536: 13,558	CPSG_10220	295	A	100	F	G	90	F
NGS_SNP.GL636536.12038	GL636536: 12,038	N/A	N/A	G	100	-	A	91.4	-
NGS_SNP.GL636536.11250	GL636536: 11,250	N/A	N/A	T	100	-	C	91.3	-

Each SNP is also linked to its own record page. Click on the following link to explore the SNP record page in more detail: [NGS\\_SNP.GL636536.6075](#).

Add to basket Add to favorites Download SNP

## SNP: NGS\_SNP.GL636536.6075

**Organism:** Coccidioides posadasii str. Silveira  
**Location:** GL636536: 6,075  
**Type:** coding  
**Number of Strains:** 66  
**Gene ID:** CPSG\_10217  
**Gene Strand:** reverse  
**Major Allele:** C (0.58)  
**Minor Allele:** T (0.42)  
**Distinct Allele Count:** 2  
**Reference Allele:** C  
**Reference Product:** G 15  
**Allele (gene strand):** G  
**SNP context:** TCTGAGACTTATTCTGGTTGCTTCCTTCCTCCAGTTGTGAATGAAT  
**SNP context (gene strand):** ATTTCATTCAACACTGGAAGGACAGGAAGGAAGAAGCAACCAGAATAAGTCTCAGA

SNP record pages provide SNP location, allele summary, associated GeneID, major and minor allele records can be found at the top of the page, followed by DNA polymorphism summary and SNP records table that is searchable by isolate IDs.

A summary of all SNPs detected in this gene across all datasets integrated into FungiDB is displayed in the SNP Genomic Context section:

SNPs are denoted by diamonds that are colored based on the coding potential:

- noncoding (yellow diamonds)
- non-synonymous (dark blue)
- synonymous (light blue)
- nonsense (red)



In the **SNP alignment section**, you can choose to align a group of selected isolates based on the metadata filters:

Select output options:

Multi-FASTA  
 Show Alignment (max 10,000 nucleotides per sequence)  
 Include strain and isolate metadata in the output.

Select strains:

78 Reference Samples Total      53 of 78 Reference Samples selected      Country

expand all | collapse all      Find a variable

Fungal organism		Country		
Host organism		Keep checked values at top		
Data Set		77 (99%) of 78 Reference Samples have data for this variable		
Fungal strain				
Sample		Remaining Reference Samples	Reference Samples	Distribution
Geographic location		77 (100%)	77 (100%)	%
Country				
City, village, or region				
<input checked="" type="checkbox"/> Argentina		1 (1%)	1 (1%)	(100%)
<input type="checkbox"/> Brazil		1 (1%)	1 (1%)	(100%)
<input type="checkbox"/> Guatemala		5 (6%)	5 (6%)	(100%)
<input type="checkbox"/> Mexico		10 (13%)	10 (13%)	(100%)
<input type="checkbox"/> Paraguay		1 (1%)	1 (1%)	(100%)
<input checked="" type="checkbox"/> United States of America		52 (64%)	52 (64%)	(100%)
<input type="checkbox"/> Venezuela		7 (9%)	7 (9%)	(100%)

**View Results**

The **Country Summary** section provides a global overview of the major and minor alleles per country:

▼ Country Summary [Download](#) [Data Sets](#)

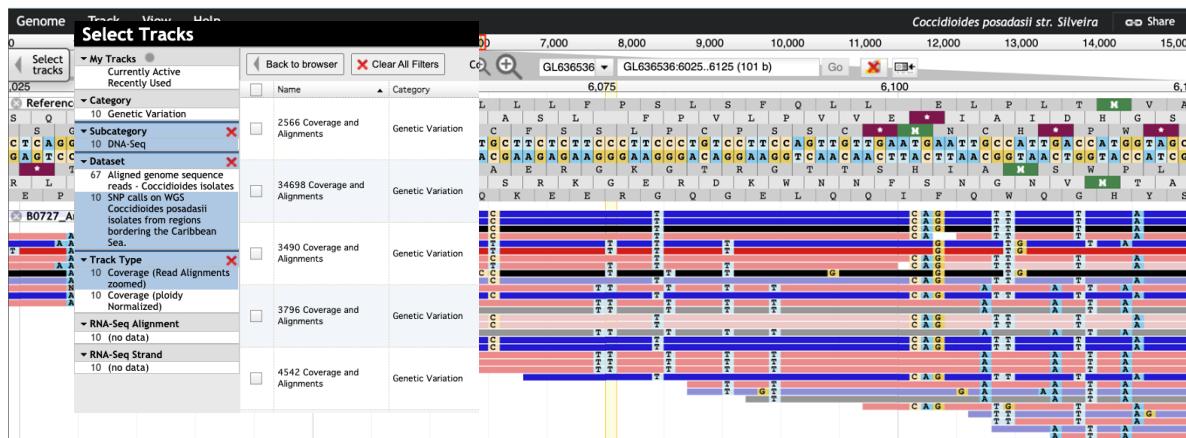
*Search this table...* ?

Geographic Location	#Alleles	Major Allele	Minor Allele	Other Allele
United States of America	65	C (.62)	T (.38)	N/A
Mexico	15	C (.53)	T (.47)	N/A
Venezuela	10	T (.7)	C (.3)	N/A
Guatemala	6	C (.83)	T (.17)	N/A
Argentina	2	C (.5)	T (.5)	N/A
Brazil	2	C (.5)	T (.5)	N/A
Paraguay	2	C (.5)	T (.5)	N/A
unknown	1	C (1)	N/A	N/A

Venezuela    JTORRES    EUAMPL0102-1-7    C    G    C    75    100 [view DNA-seq reads](#)

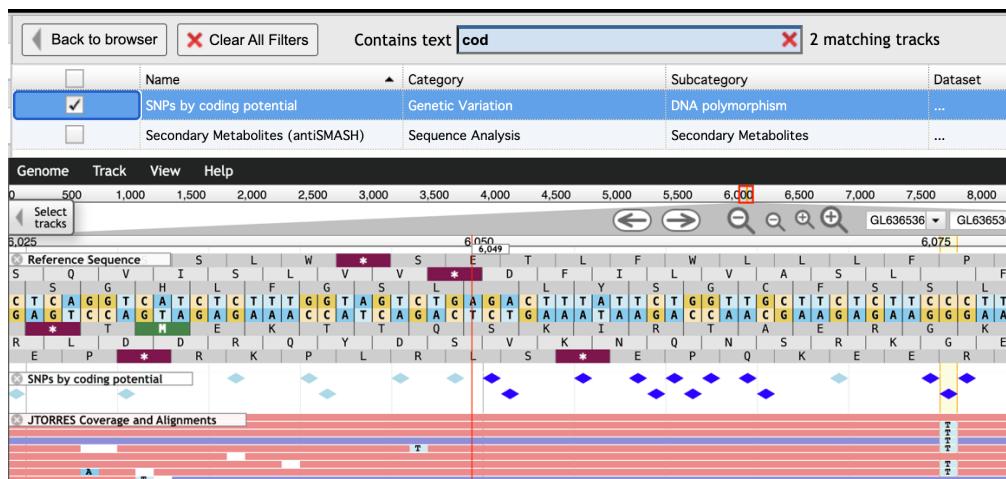
DNA-seq reads can be viewed by clicking on the *view DNA-seq reads* link from within the table.

- Click on the “[view DNA-seq reads](#)” link to navigate to a JBrowse highlighting one of the SNPs detected. (Hint: You can always select more tracks to examine by clicking on the Select Tracks tab on the left).



Note that you can visualize individual mutations reported across all isolates sequenced and integrated in FungiDB.

You can also activate the “SNPs by coding potential” to help you visually differentiate between different kinds of SNP identified.



### C. Learn how to use copy number variation (genes) search.

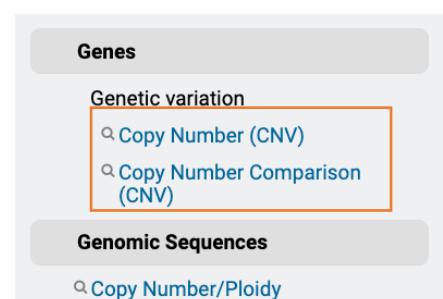
Gene copy number variation can be caused by deletions or duplications. In addition to being useful for variant calling, high-throughput sequencing data can be used to determine regions with copy number variation (CNV). All reads in FungiDB are mapped to the same reference strain as SNP datasets, and, as a result, we can estimate a gene's copy number in each of the aligned strains.

#### Copy Number search (Genes)

##### Using Gene Searches

One way to discover regions of potential segmental aneuploidy is to use the searches for genes by copy number. We have two searches: Gene searches taking advantage of sequence alignment data can be found under the “Genetic Variation” category. Two available searches that define regions of CNV are:

- **Copy number:** This search returns genes that are present at copy numbers (haploid number or gene dose) within a range that you specify.
- **Copy number comparison:** This search compares the estimated copy number of a gene in the re-sequenced strain with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are in the same ortholog group as the gene of interest. We infer that these genes have arisen as a result of tandem duplication of a common ancestor.



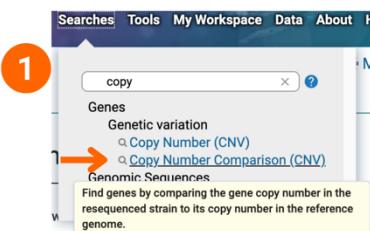
Different metrics for defining copy number:

- **Haploid number** is the number of genes on an individual chromosome.
- **Gene dose** is the total number of genes in an organism, accounting for copy number of the chromosome.

For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2. You can choose to search for genes where at least one of your selected isolates meets your cutoff criteria for the chosen metric (By Strain/Sample), or where the median of the chosen metric across all the selected isolates meets the cutoff (Median of Selected Strains/Samples).

In the next exercise, we will discover aneuploidy in a clinical strain of *Candida albicans*.

- Discover regions of potential segmental aneuploidy in *Candida albicans* isolate 5106.
  1. Deploy the “Copy Number Comparison (CNV)” search.
  2. Select the genome for “*Candida albicans*”.
  3. Navigate to the Fungal strain” metadata field.
  4. Filter isolates for “5106” and check the box to select this isolate.
  5. Leave the “Median or By Strain/Sample” parameter at default.
  - Note: Choosing Median or By Strain/Sample will only make a difference if you have multiple strains.
  6. From the drop-down menu select the “Copy number in resequenced strain is greater than reference” option.



### Identify Genes based on Copy Number Comparison (CNV)

Configure Search    Learn More    View Data Sets Used

Reset values to default

**Organism**

2

**Strain/Sample**

263 Strain/Sample Total    1 of 263 Strain/Sample selected    Fungal strain

expand all | collapse all     Find a variable

Fungal strain

Keep checked values at top    262 (>99%) of 263 Strain/Sample have data for this variable

Rows per page: 100

	Fungal strain	Remaining Strain/Sam...	Strain/Sam...	Distribution	%
<input checked="" type="checkbox"/> Candida albicans 5106	262 (100%)	262 (100%)	1 (< 1%)	1 (< 1%)	(100%)

3  Rows per page: 100

**Median Or By Strain/Sample?**

5

**What comparison do you want to make?**

6

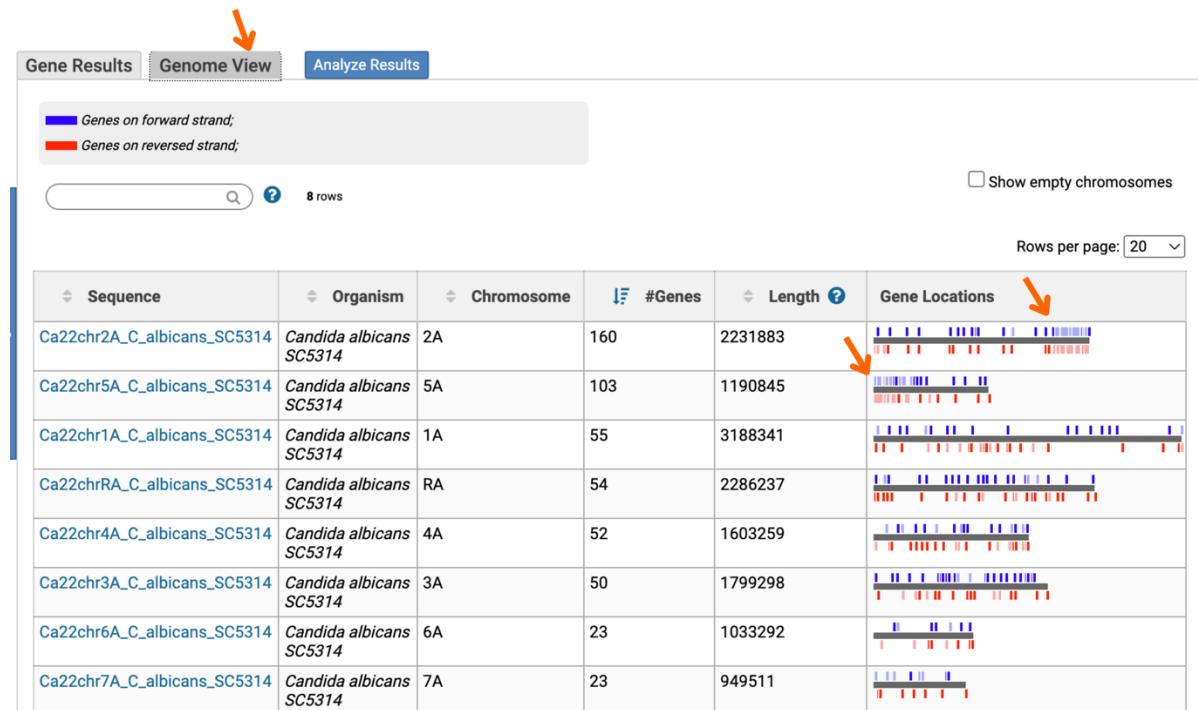
Get Answer

**CopyNumberComparison**  
520 Genes

+ Add a step

Step 1

Examine the results using the Genome View option.



As you can see in the highlighted regions, large numbers of genes predicted to have increased copy numbers are clustered at the right-hand end of chromosome 2 and the left-hand end of chromosome 5, corresponding to the segmental aneuploidies shown in the JBrowse session above.

Strategy URL:

<https://fungidb.org/fungidb/app/workspace/strategies/import/07b439e0de5e9c6a>

## Optional

### Copy Number/Ploidy search (Genomic Sequences)

Find genomic sequences or chromosomes based on their estimated copy number in resequenced strains. Genomic sequences returned by the search will either have a median estimated copy number greater than or equal to the value you entered for the Copy Number across the selected strains/samples **or** will have an estimated copy number greater than or equal to the value you entered for the Copy Number in at least one of the selected strains/samples.

- **Identify trisomic chromosomes in clinical isolates of *Candida albicans*.**

1. Deploy the “Copy Number/Ploidy” search.
2. Set the genome to *Candida albicans* SC5314.
3. Navigate to the Data Set section.
4. Select the dataset called “SNP calls on WGS of *Candida albicans* clinical isolates (oropharyngeal candidiasis)”.
5. Set the Copy Number to “3”.
6. Select to identify ploidy “By strain/sample” and click on the “Get Answer” button.

The screenshot shows the BioNumerics software interface. At the top, there is a navigation bar with links to Searches, Tools, My Workspace, Data, and About. Below the navigation bar, a search bar contains the text "ploid". A dropdown menu labeled "Genomic Sequences" is open, with the option "Copy Number/Ploidy" highlighted. An orange circle labeled "1" is placed over the search bar area. A large orange arrow points downwards from the search bar area to the main search interface.

The main search interface has several sections:

- Organism:** A dropdown menu set to "Candida albicans SC5314". An orange circle labeled "2" is placed over this dropdown.
- Strain/Sample:** A section showing "263 Strain/Sample Total" and "42 of 263 Strain/Sample selected". A "Data Set" button is present. An orange circle labeled "3" is placed over the "Data Set" button.
- Data Set:** A table titled "Data Set" with a header row. The first column contains checkboxes for different datasets. The second column lists the dataset names. The third column shows the count and percentage of strains. The fourth column shows the count and percentage of samples. The fifth column shows the distribution of strains. The last column shows the percentage of samples. An orange circle labeled "4" is placed over the checkbox for the dataset "SNP calls on WGS of Candida albicans clinical isolates (oropharyngeal candidiasis)".
- Copy Number >=:** A text input field containing the value "3". An orange circle labeled "5" is placed over this input field.
- Median Or By Strain/Sample?** A dropdown menu set to "By Strain/Sample (at least one selected strain/sample meets criteria)". An orange circle labeled "6" is placed over this dropdown.

search by strain/sample (i.e., at one or more of the selected strains must match the criteria rather than the median of the selected strains matching) is intended to find chromosomes where the whole chromosome is duplicated. It may find chromosomes where partial aneuploidy involves most of the chromosome but is unlikely to find chromosomes where

partial aneuploidy only covers a small region. Also, because this search currently relies on coverage alone, it will not find instances of global genome duplication (e.g., all chromosomes became triploid).

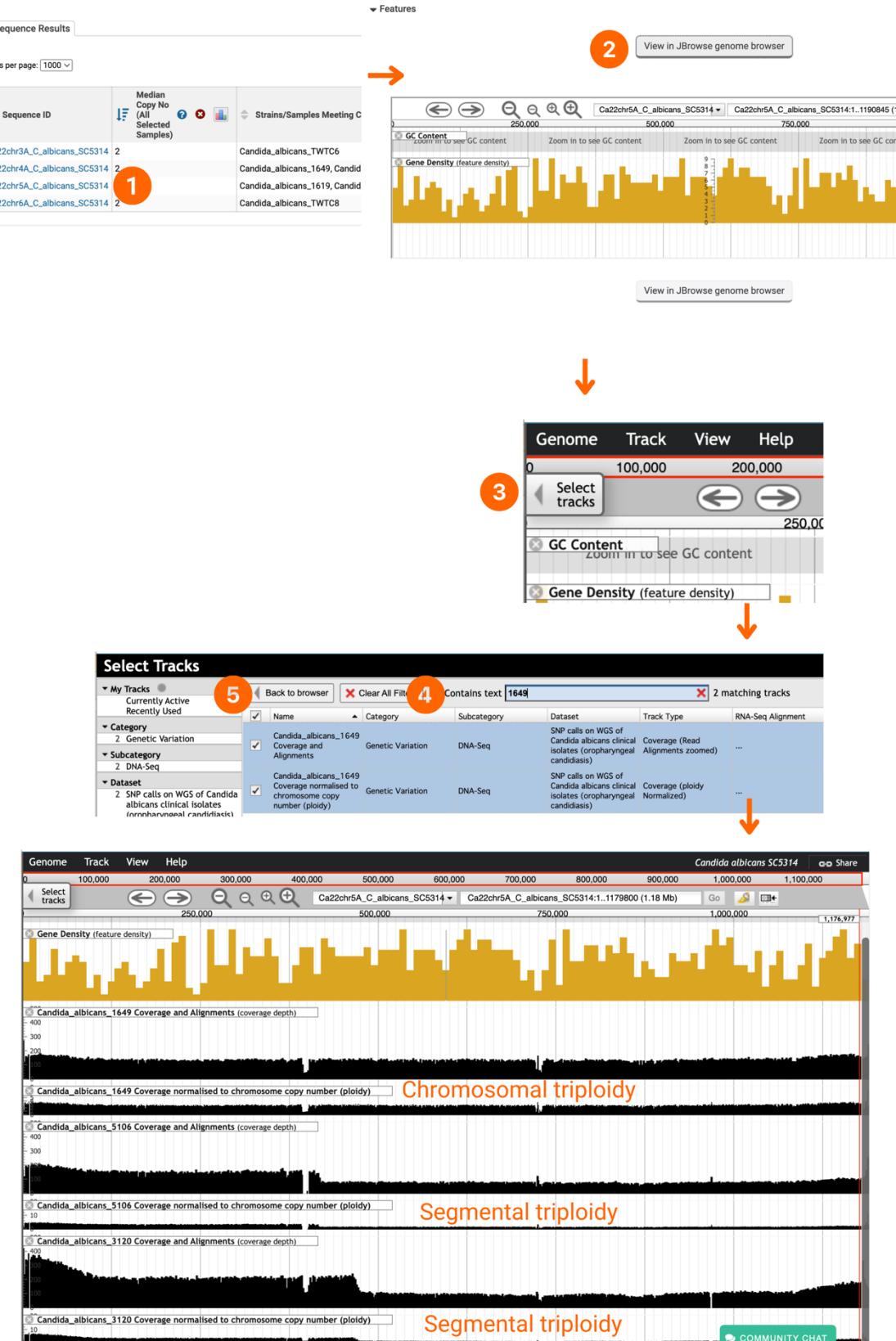
Sequence ID	Median Copy No (All Selected Samples)	Strains/Samples Meeting Criteria	Median Copy No (Samples Meeting Criteria)
Ca22chr3A_C_albicans_SC5314	2	Candida_albicans_TWTC6	3
Ca22chr4A_C_albicans_SC5314	2	Candida_albicans_1649, Candida_albicans_2501, Candida_albicans_3731, Candida_albicans_5106	3
Ca22chr5A_C_albicans_SC5314	2	Candida_albicans_1619, Candida_albicans_1649, Candida_albicans_2823, Candida_albicans_3034, Candi...	3
Ca22chr6A_C_albicans_SC5314	2	Candida_albicans_TWTC8	3

- Explore segmental aneuploidy in JBrowse.**

JBrowse has two coverage tracks:

- Raw coverage from the alignment (available for every isolate where we have whole genome sequencing, whether we ran the copy number pipeline or not)
- Normalized coverage in bins (only available for isolates where we have run the copy number pipeline)
  1. Click on one of the Sequence ID Ca22chr5A\_C\_albicans\_SC5314 (in blue).
  2. Navigate to JBrowse by clicking on the “View in JBrowse genome browser” button.
  3. When in JBrowse, click on the Select tracks tab to customize your view.
  4. Use the “Contains text” filter to identify and select tracks for the following isolates: 1649, 5106, and 3120.
  5. Click on the “Back to browse” tab to return to JBrowse view with selected tracks.

#### ▼ 4.2 Sequence sites, features and motifs



Notice examples of chromosomal (1649) and segmental triploidy (5106 and 3120). The whole chromosome is shown in both screenshots, and both tracks are shown for each sample. Note that VEuPathDB is not currently normalizing for telomere proximity.

URL:

[https://fungidb.org/fungidb/ibrowse/index.html?loc=Ca22chr5A\\_C\\_albicans\\_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fibrowse%2Ftracks%2FcalbSC5314&tracks=gcContent%2CgeneDensity%2CCandida\\_albicans\\_1649%20Coverage%20and%20Alignments%2CCandida\\_albicans\\_1649%20Coverage%20normalised%20to%20chromosome%20copy%20number%20\(ploidy\)%2CCandida\\_albicans\\_5106%20Coverage%20and%20Alignments%2CCandida\\_albicans\\_5106%20Coverage%20normalised%20to%20chromosome%20copy%20number%20\(ploidy\)%2CCandida\\_albicans\\_3120%20Coverage%20and%20Alignments%2CCandida\\_albicans\\_3120%20Coverage%20normalised%20to%20chromosome%20copy%20number%20\(ploidy\)&highlight=](https://fungidb.org/fungidb/ibrowse/index.html?loc=Ca22chr5A_C_albicans_SC5314%3A1..1190845&data=%2Ffungidb%2Fservice%2Fibrowse%2Ftracks%2FcalbSC5314&tracks=gcContent%2CgeneDensity%2CCandida_albicans_1649%20Coverage%20and%20Alignments%2CCandida_albicans_1649%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)%2CCandida_albicans_5106%20Coverage%20and%20Alignments%2CCandida_albicans_5106%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)%2CCandida_albicans_3120%20Coverage%20and%20Alignments%2CCandida_albicans_3120%20Coverage%20normalised%20to%20chromosome%20copy%20number%20(ploidy)&highlight=)

Strategy URL: <https://fungidb.org/fungidb/app/workspace/strategies/import/c899493a551276b3>

References: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383195/>

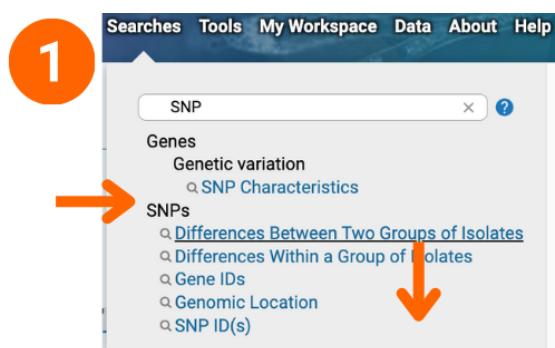
## Supplement

### How to distinguish SNP between two groups of isolates.

Create a strategy using the “Differences Between Two Groups of Isolates” search, which distinguishes SNPs between two groups of *C. posadasii* str Silveira isolates (A and B) collected in the Caribbean (group A) and western hemispheres (group B):

- **Identify SNPs between two groups of *C. posadasii* str. Silveira isolates**

1. Deploy the “Difference Between Two Groups of Isolates” search.
2. Set the genome to *Coccidioides posadasii* strain Silveira.
3. Select Set A isolates from the Data Set menu: Caribbean dataset.
4. Select Set B isolates from the Data Set menu: Western hemisphere dataset.
5. Click on the “Get Answer” button to get the results.



#### Identify SNPs based on Differences Between Two Groups of Isolates

The screenshot shows a search interface with a blue header bar containing 'Configure Search', 'Learn More', and 'View Data Sets Used'. Below the header is a 'Reset values to default' button. The main area is titled 'Organism' with a question mark icon. A search bar contains the text 'sily'. To the right of the search bar are a question mark icon and a 'Reference only' checkbox. Below the search bar, a list of organisms is shown, starting with 'Ascomycota', 'Eurotiomycetes', 'Onygenales', 'Coccidioides', and 'Coccidioides posadasii str. Silveira [Reference]'. An orange circle with the number '2' is positioned at the top left of the screenshot. An orange arrow points from the '2' towards the 'Organism' search bar.

**3**

Set A Isolates

78 Set A Isolates Total  
expand all | collapse all  
Find a variable  ?

- Fungal organism
- absolute proportion mapped reads
- Host organism
- Data Set
- Fungal strain
- Sample
- Sample collection
- Geographic location

10 of 78 Set A Isolates selected Data Set ?

Keep checked values at top

Data Set	Remaining Set A Isolates	Set A Isolates	Distribution	%
<input type="checkbox"/> Coccidioides posadasii str. Silveira Genome Sequence and Annotation	1 (1%)	1 (1%)	<div style="width: 10%;">10%</div>	(100%)
<input checked="" type="checkbox"/> SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea.	10 (13%)	10 (13%)	<div style="width: 100%;">100%</div>	(100%)
<input type="checkbox"/> SNP calls on WGS of Coccidioides isolates from the Western Hemisphere	67 (86%)	67 (86%)	<div style="width: 100%;">100%</div>	(100%)

78 (100%) of 78 Set A Isolates have data for this variable

**4**

Set B Isolates

78 Set B Isolates Total  
expand all | collapse all  
Find a variable  ?

- Fungal organism
- absolute proportion mapped reads
- Host organism
- Data Set
- Fungal strain
- Sample
- Sample collection
- Geographic location
  - Country
  - City, village, or region

67 of 78 Set B Isolates selected Data Set ?

Keep checked values at top

Data Set	Remaining Set B Isolates	Set B Isolates	Distribution	%
<input type="checkbox"/> Coccidioides posadasii str. Silveira Genome Sequence and Annotation	1 (1%)	1 (1%)	<div style="width: 10%;">10%</div>	(100%)
<input type="checkbox"/> SNP calls on WGS Coccidioides posadasii isolates from regions bordering the Caribbean Sea.	10 (13%)	10 (13%)	<div style="width: 100%;">100%</div>	(100%)
<input checked="" type="checkbox"/> SNP calls on WGS of Coccidioides isolates from the Western Hemisphere	67 (86%)	67 (86%)	<div style="width: 100%;">100%</div>	(100%)

78 (100%) of 78 Set B Isolates have data for this variable

**5** Get Answer

Two Groups  
19,147 SNPs

+ Add a step

Step 1

- Change the stringency of your search to major allele frequency  $\geq 90\%$

**1**

Step 1

Two Groups  
19,147 SNPs Edit

+ Add a step

→

View | [Revise](#) | [Insert step before](#) | [Delete](#)

Details for step Two Groups Edit  
4059 SNPs

Modify the configuration of this search

↓

### Set A major allele frequency $\geq$

**2** 90

### Set B major allele frequency $\geq$

**3** 90

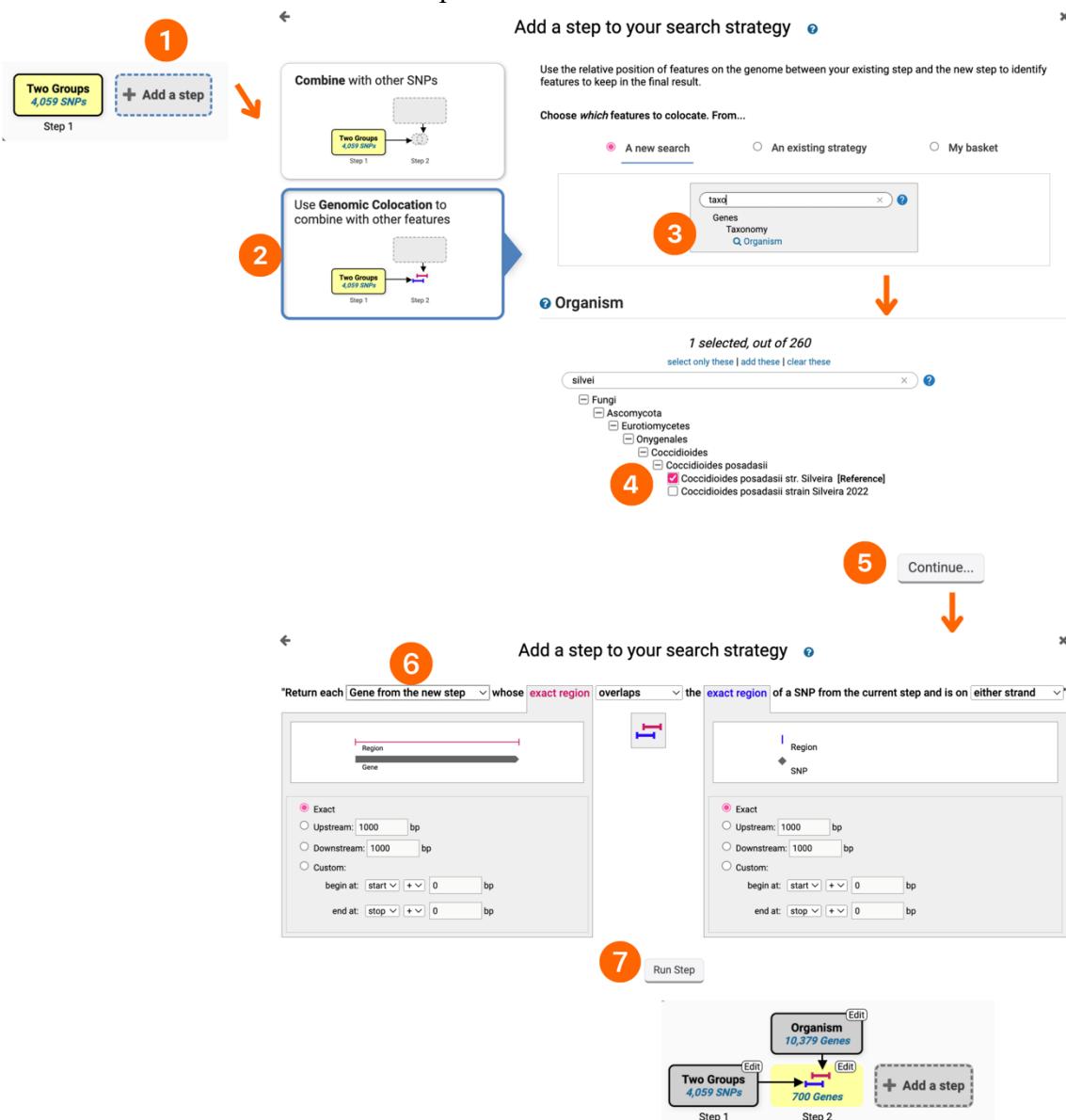
Two Groups  
4,059 SNPs

+ Add a step

Step 1

- Map SNPs from Step 1 to genes in *C. posadasii* str. Silveira.

1. Click on the “Add a step” button.
2. Select the “Use Genomic Colocation to combine with other features” tool.
3. Filter searches on “taxonomy” to identify the “Organism” search.
4. Select *C. posadasii* strain Silveira genome.
5. Click on the “Continue...” button to specify colocation search parameters.
6. Select to return ‘Gene from the new step’ whose exact region overlaps the SNP.
7. Click on the “Run Step” button for results.



In this strategy we compared SNPs in *C. posadasii* collected in different geographical regions and identified 700 genes that overlap with these SNPs. For those genes that are not well characterized (e.g., conserved hypothetical proteins) you can use other searches and tools to understand their function.

Strategy URL:

<https://fungidb.org/fungidb/app/workspace/strategies/import/d9d0fff2dbda229d>

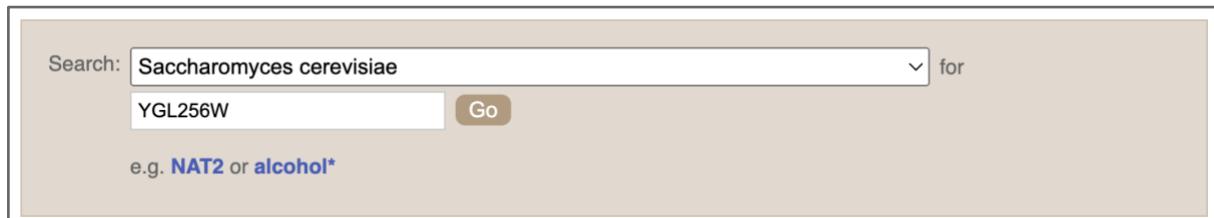
## Exercise: Exploring variants in Ensembl Fungi

Links to be clicked shown in blue, text to be entered shown in red.

In any of the sequence views shown in the ‘Gene’ and ‘Transcript’ tabs, you can view variants on the sequence. You can do this by clicking on [Configure this page](#)

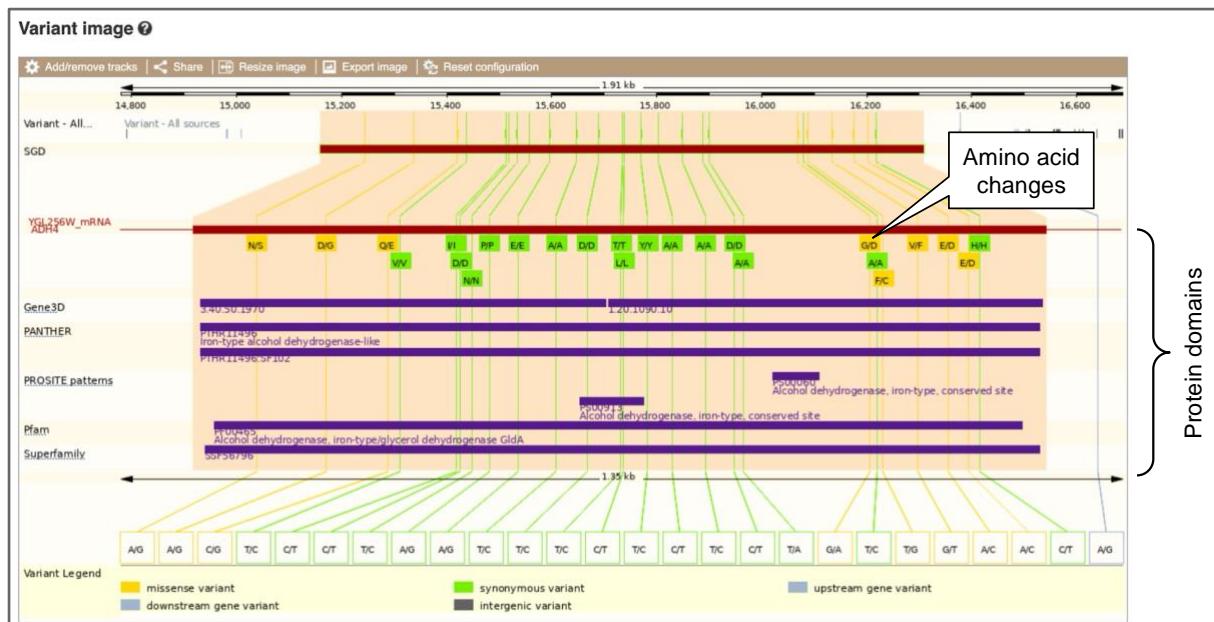
 from any of these views.

Let's take a look at the [Gene sequence](#) view for *ADH4* (gene stable ID: YGL256W). This gene is a ribonuclease protein in *Saccharomyces cerevisiae* (R64-1-1). Select *Saccharomyces cerevisiae* R64-1-1 under [Favourite genomes](#) on the Ensembl Fungi homepage. Search for **YGL256W** and go to the [Variant image](#) view.



Search:  for   e.g. [NAT2](#) or [alcohol\\*](#)

This view shows variants mapped to the gene structure and protein domains.



We can examine all variants and filter to see the ones we are interested in using the ‘Variant’ table. Click on the [Variant table](#) link on the left-hand menu.

This table shows the variants in order of their occurrence throughout the genome, and they are reported on the forward strand. The gene *ADH4* is located on the forward strand, so we are first shown variants upstream of the gene (starting at the 5' upstream region).

(a) How many variants in this gene are predicted to be missense?

You can filter the table to view variants that alter the protein sequence. Click on the **Consequences: All** button above the table. Click the option ‘**PTV and Missense**’ in the pop-up, then **Apply**. You can also filter by other columns such as variant **Class**.

Consequence Type	Count	Status
transcript ablation	(0)	Off
splice acceptor variant	(0)	On
splice donor variant	(0)	On
stop gained	(0)	On
frameshift variant	(0)	On
stop lost	(0)	Off
start lost	(0)	Off
transcript amplification	(0)	Off
inframe insertion	(0)	Off
inframe deletion	(0)	Off
missense variant	(8)	On
protein altering variant	(0)	Off
splice donor 5th base variant	(0)	Off
splice region variant	(0)	Off
splice donor region variant	(0)	Off
splice polypyrimidine tract variant	(0)	Off
incomplete terminal codon variant	(0)	Off
start retained variant	(0)	Off
stop retained variant	(0)	Off
synonymous variant	(17)	Off
coding sequence variant	(0)	Off
mature miRNA variant	(0)	Off
5 prime UTR variant	(0)	Off
3 prime UTR variant	(0)	Off
non coding transcript exon variant	(0)	Off
intron variant	(0)	Off
NMD transcript variant	(0)	Off
non coding transcript variant	(0)	Off
coding transcript variant	(0)	Off
upstream gene variant	(2)	Off
downstream gene variant	(6)	Off

(b) Are there any known variants in this gene predicted to be deleterious?

The SIFT scores (<https://doi.org/10.1093/nar/gkg509>) predict the consequence of the variant on the function of the protein taking into account chemical changes and conservation of amino acids. Scores <0.05 and coloured red are ‘deleterious’ while scores >0.05 and coloured green are ‘tolerated’.

Variant ID	Chr: bp	Alleles	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA co-ord	SIFT	Transcript
s07-15244	VII:15244	A/G	SNP	SGRP	-	-	missense variant	N/S	29	1	YGL256W mRNA
s07-15337	VII:15337	A/G	SNP	SGRP	-	-	missense variant	D/G	60	1	YGL256W mRNA
s07-15420	VII:15420	C/G	SNP	SGRP	-	-	missense variant	Q/E	88	0.72	YGL256W mRNA
s07-16069	VII:16069	G/A	SNP	SGRP	-	-	missense variant	G/D	304	0.1	YGL256W mRNA
s07-16087	VII:16087	T/G	SNP	SGRP	-	-	missense variant	F/C	310	0	YGL256W mRNA
s07-16134	VII:16134	G/T	SNP	SGRP	-	-	missense variant	V/F	326	0.03	YGL256W mRNA
s07-16175	VII:16175	A/C	SNP	SGRP	-	-	missense variant	E/D	339	0.26	YGL256W mRNA
s07-16202							missense variant	E/D	348	0.67	YGL256W mRNA

Let’s have a look at a specific variant. Click on the top result in the filtered table, or

search for [s07-15244](#). This will open the ‘Variant’ tab.

The icons show you what information is available for this variant.

(c) What are the genomic coordinates of this variant?

Location	<a href="#">Chromosome VII:15244</a> (forward strand)   VCF: VII 15244 s07-15244 A G
----------	--

(d) What is the reference allele? (*Hint: Ensembl always reports alleles on the forward strand. The reference allele is given first.*)

You can find some background information on variants, alleles and haplotypes here: <https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/what-are-variants-alleles-and-haplotypes/>. The reference allele for s07-15244 is A.

(e) How many genes are affected by this variant? Does it have the same consequence across different transcripts of different genes?

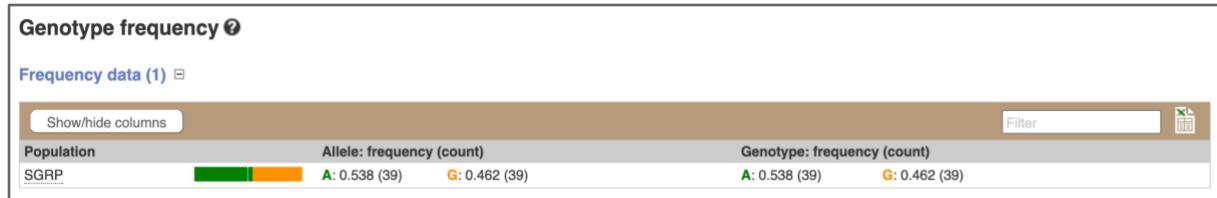
Click on the [Genes and regulation](#) icon, or follow the link in the left-hand panel.

Genes and regulation											
Gene and Transcript consequences											
<a href="#">Show/hide columns</a> <a href="#">Filter</a> <a href="#">Print</a>											
Gene	Transcript (strand)	Allele (Tr. allele)	Consequence Type	Position in transcript	Position in CDS	Position in protein	AA	Codons	SIFT	Detail	
YGL256W	YGL256W_mRNA (+) biotype: protein_coding	G (G)	missense variant	86 (out of 1149)	86 (out of 1149)	29 (out of 382)	N/S	AAC/AGC	1	<a href="#">Show</a>	
No overlap with Ensembl Regulatory features											
No overlap with Ensembl Motif features											

This variant overlaps one gene. It causes a change in the protein sequence (missense variant) in the YGL256W gene we were looking at (note that only missense variants have SIFT scores).

- (f) Which allele is major in the *Saccharomyces* Genome Resequencing Project (SGRP) study?

Click on [Genotype frequency](#) in the left-hand menu. Note that the reference allele A is more frequent than the alternative allele G in this case.



## Additional Exercise – Variation data in *Fusarium oxysporum*

- Select the *Fusarium oxysporum* FO2 genome and search for FOXG\_13574T0 gene. One of its upstream variants is SNP tmp\_10\_6610. What are the possible alleles for this polymorphic position? Which one is on the reference genome?
- What is the most frequent allele at this position? How many heterozygous individuals were observed in the melonis population?
- Which individuals have got genotypes C|T and T|T?

### Answers

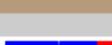
- You can find the alleles in the summary information at the top of the ‘Variant’ tab. The reference allele for tmp\_10\_6610 is C and the alternative allele is T.

**tmp\_10\_6610 SNP**

Most severe consequence	<a href="#">upstream gene variant</a>   <a href="#">See all predicted consequences</a>
Alleles	C/T   Highest population MAF: 0.15
Location	Chromosome 10:6610 (forward strand)   VCF: 10 6610 tmp_10_6610 C T
HGVS name	10:g.6610C>T
External Links	
Original source	
About this variant	This variant overlaps <a href="#">4 transcripts</a> and has <a href="#">10 sample genotypes</a> .

- Click on **Genotype frequency** in the left-hand panel. The most frequent allele is C. There is one heterozygous individual in the melonis population.

**Genotype frequency** ⓘ

Frequency data (1)		
<a href="#">Show/hide columns</a>		
Population	Allele: frequency (count)	Genotype: frequency (count)
melonis	 C: 0.850 (17)    T: 0.150 (3)	C/C: 0.800 (8)    C/T: 0.100 (1)    T/T: 0.100 (1)

- Click on **Sample genotypes** in the left-hand panel. Individual 909454 is heterozygous (C|T genotype) and individual 909455 is homozygous for the minor allele (T|T genotype).

<b>Sample genotypes</b>						
Search for a sample:		Search (e.g. NA18507)				
<a href="#">[back to top]</a>						
<b>Genotypes for melonis</b> ▾						
Sample (Male/Female/Unknown)	Genotype (forward strand)	Population(s)	Father	Mother		
886599 (U)	C/C	melonis	-	-		
889404 (U)	C/C	melonis	-	-		
889405 (U)	C/C	melonis	-	-		
889406 (U)	C/C	melonis	-	-		
889407 (U)	C/C	melonis	-	-		
889408 (U)	C/C	melonis	-	-		
889410 (U)	C/C	melonis	-	-		
909453 (U)	C/C	melonis	-	-		
909454 (U)	C/T	melonis	-	-		
909455 (U)	T/T	melonis	-	-		

## Exercise: The Ensembl Fungi Variant Effect Predictor (VEP)

We have identified four variants in *Verticillium dahliae* JR2: chromosome 5, C->G at 698711, G->T at 698935, G->A at 700313 and C->A at 701484. Use the Ensembl VEP to determine:

- (a) Are your variants novel or have they already been annotated in Ensembl?
- (b) What genes are affected by your variants?
- (c) Do any of your variants affect gene regulation?

Click on **Tools** in the top brown bar from any Ensembl Fungi page, then **Variant Effect Predictor** to open the input form. You will need to change the species to *Verticillium dahliae* JR2 and paste your input data in the provided text box.

The VEP recognises a number of input formats including the Ensembl default format, variant call format (VCF), variant identifiers and HGVS notations. The HGVS nomenclature is a globally recognised standard for describing variants. You can read more about this here: <https://hgvs-nomenclature.org/stable/>.

The Ensembl default format is composed of four compulsory columns and additional ‘strand’ column: Chromosome, Start Position, End Position, Alleles (reference/alternate), Strand (1 for forward; -1 for reverse), with one line per variant. Your variants in this format would look like this:

```
5 698711 698711 C/G
5 698935 698935 G/T
5 700313 700313 G/A
5 701484 701484 C/A
```

The VEP will automatically detect that the data is in Ensembl default format. Clicking on the [Run instant VEP for current line](#) option will generate a pop-up with summarised results for that individual variant.

The screenshot shows the 'Instant results' window for a VEP run. At the top, it says 'Instant results for 5 701484 701484 C/A'. There is a button 'Run instant VEP for current line' in the top right corner. The main area has a yellow header 'Instant VEP' with a warning message: 'The below is a preview of results using the *Verticillium dahliae* Jr2 Ensembl transcript database and does not include all data fields present in the full results set. To obtain these please close this preview window and submit the job using the Run button below.' Below this, it says 'Most severe consequence: upstream\_gene\_variant' and 'Colocated variants: tmp 5 701484 C A'. A table follows, showing variants categorized by gene/feature/type, consequence, and details like distance to transcript:

Gene/Feature/Type	Consequence	Details
<a href="#">VDAG_JR2_Chr5g02160a: VDAG_JR2_Chr5g02160a-00001</a> Type: protein_coding	downstream_gene_variant	Distance to transcript: 2165bp
<a href="#">VDAG_JR2_Chr5g02170a: VDAG_JR2_Chr5g02170a-00001</a> Type: protein_coding	downstream_gene_variant	Distance to transcript: 742bp
<a href="#">VDAG_JR2_Chr5g02170a: VDAG_JR2_Chr5g02170a-00002</a> Type: protein_coding	downstream_gene_variant	Distance to transcript: 778bp
<a href="#">VDAG_JR2_Chr5g02171a: VDAG_JR2_Chr5g02171a-00001</a> Type: protein_coding	upstream_gene_variant	Distance to transcript: 64bp

There are further options that you can choose for your output. These are categorised as [Identifiers](#), [Variants and frequency data](#), [Additional annotations](#), [Predictions](#), [Filtering options](#) and [Advanced options](#). Let's open all the menus and take a look.

The screenshot shows the 'Additional configurations' section of the VEP tool. It includes two boxes with arrows pointing to specific settings:

- A box labeled 'Select which identifiers you want in your output' points to the 'Identifiers' section, which contains checkboxes for Gene symbol, Transcript version, Protein, UniProt, and HGVS.
- A box labeled 'Does this variant already exist?' points to the 'Variants and frequency data' section, which contains a dropdown for 'Find co-located known variants' (set to 'Yes'), and checkboxes for Variant synonyms and Include flagged variants.

Add information about affected transcripts and proteins

**Species:** *Saccharomyces\_cerevisiae* Assembly: R64-1-1 Change species

Name for this job (optional):

**Input data:**

Either paste data:  
[Large text area]

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#)

Or upload file:  No file chosen

Or provide file URL: [Text input]

**Additional configurations:**

**Identifiers** Additional identifiers for genes, transcripts and variants

**Identifiers**

- Gene symbol:
- Transcript version:
- Protein:
- UniProt:
- HGVS:

**Variants and frequency data** Co-located variants and frequency data

**Variants and frequency data**

- Find co-located known variants:
- Variant synonyms:
- Include flagged variants:

**Additional annotations** Additional transcript, protein and regulatory annotations

**Transcript annotation**

- Transcript biotype:
- Exon and intron numbers:
- Identify canonical transcript

Show only coding variants

More filters

Hover over the options to see definitions. When you've selected everything you need, scroll to the bottom of the page and click **Run**.

The screenshot shows the Variant Effect Predictor (VEP) interface. At the top left is a 'New job' button. Below it is a 'Recent jobs' section with a 'Refresh' button. A callout box points to the 'Refresh' button with the text: 'This will count down and refresh the page every 10 seconds'. To the right is a 'Jobs' section showing a single job entry: 'Variant Effect Predictor VEP analysis of Fungal Pathogens VEP Exercise in Verticillium\_dahliaejr2 Done [View results]'. A callout box points to the 'Done' status with the text: 'Options to save, edit, share or delete the job'. Below the job list is a 'Submitted at (GMT)' timestamp: '08/04/2023, 20:16'. A callout box points to the '[View results]' link with the text: 'Click here to view your results'.

A table display will show you the status of your job. It will say **Queued**, then automatically switch to **Done** when the job is done, you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

Click **View results** once your job is done. In your results you will see a graphical summary of your data, as well as a table of your results.

Let's come back to our questions:

- Are your variants novel or have they already been annotated in Ensembl?
- What genes are affected by your variants?
- Do any of your variants affect gene regulation?

The screenshot shows the 'Variant Effect Predictor results' page. At the top left is a 'Job details' section with a 'Summary statistics' button. Below is a table titled 'Category' and 'Count' showing variant counts: Variants processed (4), Variants filtered out (0), Novel / existing variants (3 (75.0) / 1 (25.0)), Overlapped genes (4), Overlapped transcripts (5), and Overlapped regulatory features (-). To the right are two charts: 'Consequences (all)' and 'Coding consequences'. The 'Consequences (all)' chart is a pie chart showing distribution: downstream\_gene\_variant (50%), upstream\_gene\_variant (28%), missense\_variant (11%), 3\_prime\_UTR\_variant (6%), and intron\_variant (6%). The 'Coding consequences' chart is a donut chart showing 100% missense\_variant.

The output table reports one variant consequence per row. If your variants have multiple alternate alleles, hit multiple genes or transcripts, you'll find few lines per variant. If the output table is large, you might want to use the filter option to narrow it down. Once you've added a filter, it will appear in the filter box, allowing you to add other filters.

The image consists of two vertically stacked screenshots of the Ensembl Fungi SNP search interface. Both screenshots show the 'Results preview' section with a navigation bar ('Navigation (per variant)', 'Page: 1 of 1', 'Show: All variants') and a 'Filters' section.

**Screenshot 1 (Top):** This shows the initial state of the filter. A tooltip 'Filter by column (select the column name from the drop-down menu)' points to the 'Consequence' dropdown. A tooltip 'Logical operator' points to the 'is' dropdown. A tooltip 'Start typing to see all matching options' points to the search input field containing 'mis'. A tooltip 'Apply your filter' points to the 'Add' button. The search input field has a dropdown menu open showing 'missense\_variant'.

**Screenshot 2 (Bottom):** This shows the filter applied. The 'Filters' section now displays 'Consequence is missense\_variant'. A tooltip 'Edit filter' points to the edit icon next to the filter. A tooltip 'Remove filter' points to the delete icon. A tooltip 'Applied filter' points to the status message 'Uploaded variant is defined'. A tooltip 'Add another filter' points to the 'Add' button. The status bar at the bottom shows 'All: VCF VEP TXT' and 'Filtered: VCF VEP TXT'.

Filter text box is by default set to ‘defined’, which can be used to filter out empty values, e.g. ‘Existing variant’ ‘is’ ‘defined’ will filter out variants with empty values in the ‘Existing variant’ column, leaving you with known variants only. Note that you should not type ‘defined’ in the search box, just leave it as it is.

**Filter this table**

**Download options**

**Show additional columns**

**Existing variants**

**Variant 1**

**Variant 2**

**Variant 3**

**Variant 4**

Uploaded variant	Location	Allele	Consequence	Gene	Protein ID	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_698711_C/G	5_698711-698711	G	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	intron_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	upstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	upstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5_698711-698711	G	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	-1
5_698935_G/T	5_698935-698935	T	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5_698935-698935	T	3_prime_UTR_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	8/8	1679	-	-	-	-	-	1
5_698935_G/T	5_698935-698935	T	upstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5_698935-698935	T	upstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5_698935-698935	T	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	-1
5_700313_G/A	5_700313-700313	A	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	1
5_700313_G/A	5_700313-700313	A	missense_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	2/2	155	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5_700313-700313	A	missense_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	2/2	161	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5_700313-700313	A	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	-	-1
5_701484_C/A	5_701484-701484	A	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A_1	
5_701484_C/A	5_701484-701484	A	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A_1	
5_701484_C/A	5_701484-701484	A	downstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A_1	
5_701484_C/A	5_701484-701484	A	upstream_gene_variant	VDAG_JR2_Chromosome 5	VDAG_JR2_Chromosome 5	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A_-1	

## Additional Exercise: The Ensembl Fungi Variant Effect Predictor (VEP)

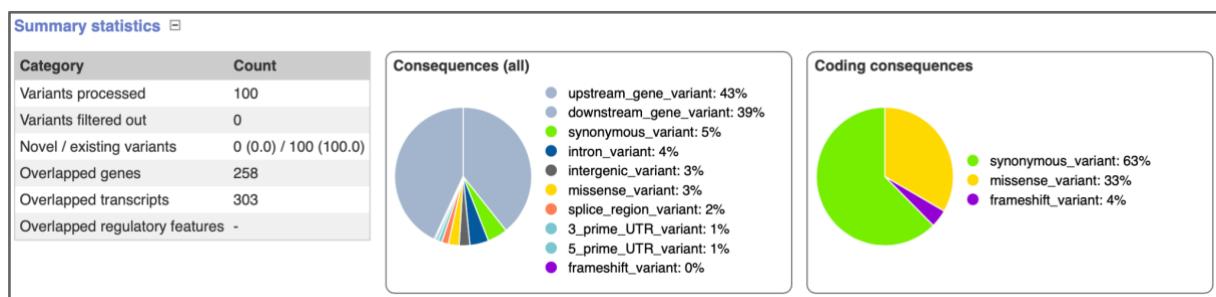
On the course file page, you will find a VCF file labelled VEP\_exercise.vcf. This is a small subset of the outcome of *Puccinia graminis* (Ug99) whole genome sequencing and variant calling experiment. This file can also be found on our FTP site under the following link:  
[http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2021/FungalPathogens/VEP\\_exercise.vcf](http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2021/FungalPathogens/VEP_exercise.vcf)

Run the file through the VEP by downloading and uploading it from your computer, or by attaching it as a remote file hosted online (you will need to provide the FTP file URL).

- (a) How many variants have been processed?
- (b) How many genes and transcripts are overlapped by variants in this file?
- (c) Do any of the variants change the amino acid sequences of any proteins? What genes?  
What is the amino acid change? (*Hint: use the filters above the table to filter by consequences.*)
- (d) What are the HGVS notations of missense variants falling in known protein domains?
- (e) How many variants are frameshift? Which gene(s) do they fall in and which exons? Can you find a UniParc ID of protein(s) affected by this variant?

## Answer

- (a) 100 variants have been processed.
- (b) The variants overlap 258 genes and 303 transcripts.



- (c) Apply the **Consequence is missense\_variant** under ‘Filters’. Under ‘Navigation’ (to the left of the filter options, click on **All**. 8 variants change the amino acid sequence in the encoding protein. The affected genes are:

GMQ\_21813  
 GMQ\_27112  
 GMQ\_04080  
 GMQ\_06767  
 GMQ\_02814  
 GMQ\_20311  
 GMQ\_20457  
 GMQ\_03045

## SNPs & variants: Ensembl Fungi

**Results preview**

**Navigation (per variant)** **Filters** **Download** **New job**

Show: 1 5 10 50 All variants  
Consequence is missense\_variant  
Uploaded variant is defined Add

All: VCF VEP TXT  
Filtered: VCF VEP TXT

Show/hide columns (22 hidden)

Location	Allele	Consequence	Gene	Exon	HGVSc	HGVSp	cdDNA position	cDNA position	Protein position	Amino acids	Codons	Existing variant	Domains
Superconfig_3.1594.801-801	T	missense_variant	GMQ_27112	3/3	GMQ_27112T0:c.265G>A	GMQ_27112T0:p.Glu89Lys	265	265	89	E/K	GAG/AAG	tmp_Superconfig_3.1594.801_C_T	Pfam:PF14892 PANTHER:PTHR45125 PANTHER:PTHR45125-SF3 MobiDB-lite:mobidb-lite
Superconfig_3.156.127654-127654	C	missense_variant	GMQ_21813	2/6	GMQ_21813T0:c.235T>C	GMQ_21813T0:p.Ser79Pro	235	235	79	S/P	TCC/CCC	tmp_Superconfig_3.156.127654_T_C	Gene3D:2.102.10.10 PROSITE_profiles:PS51296 Superfamily:SSF50022
Superconfig_3.58.69935-69935	C	missense_variant	GMQ_20457	4/4	GMQ_20457T0:c.1003G>C	GMQ_20457T0:p.Asp335His	1003	1003	335	D/H	GAT/CAT	tmp_Superconfig_3.58.69935_G_C	Gene3D:3.40.720.10 PANTHER:PTHR23071 Superfamily:SSF53649 CDD:cd16023
Superconfig_3.48.118082-118082	T	missense_variant	GMQ_20311	1/4	GMQ_20311T0:c.73G>A	GMQ_20311T0:p.Glu25Lys	73	73	25	E/K	GAA/AAA	tmp_Superconfig_3.48.118082_C_T	-
Superconfig_3.41.7765-7765	G	missense_variant	GMQ_06767	6/15	GMQ_06767T0:c.1328A>C	GMQ_06767T0:p.Gln443Pro	1328	1328	443	Q/P	CAG/CCG	tmp_Superconfig_3.41.7765_T_G	PANTHER:PTHR46896 PANTHER:PTHR46896-SF3
Superconfig_3.16.171261-171261	T	missense_variant	GMQ_04080	1/2	GMQ_04080T0:c.287G>A	GMQ_04080T0:p.Gly96Glu	287	287	96	G/E	GGA/GAA	tmp_Superconfig_3.16.171261_C_T	-
Superconfig_3.73.160474-160474	G	missense_variant	GMQ_03045	2/3	GMQ_03045T0:c.407G>C	GMQ_03045T0:p.Arg136Thr	407	407	136	R/T	AGA/ACA	tmp_Superconfig_3.73.160474_C_G	Low_complexity_(Seg).seg PANTHER:PTHR31595 PANTHER:PTHR31595-SF1
Superconfig_3.427.55213-55213	A	missense_variant	GMQ_02814	2/2	GMQ_02814T0:c.99G>T	GMQ_02814T0:p.Gln33His	358	358	99	Q/H	CAG/CAT	tmp_Superconfig_3.427.55213_C_A	PANTHER:PTHR31361 PANTHER:PTHR31361-SF15 MobiDB-lite:mobidb-lite

(d) Ensure you selected the following additional configurations in the VEP input form:

Identifiers: **Protein** (to include protein position information), **HGVS** (to include the HGVS notations)

Additional annotations: **Protein matches** (to include any overlapping protein domains)

In the VEP results table, apply the following filters:

**Consequence is missense\_variant**  
**Protein matches is [leave text box empty]**

Under ‘Navigation’ (to the left of the filter options, click on **All**). Ensure the columns **HGVSp** and **Protein matches** are visible under the **Show/hide columns** option above the table. The HGVSp notations of missense variants falling in known protein domains (see ‘Protein matches’ column) are as follows:

GMQ\_21813T0:p.Ser79Pro  
GMQ\_27112T0:p.Glu89Lys  
GMQ\_06767T0:p.Gln443Pro  
GMQ\_02814T0:p.Gln33His  
GMQ\_20457T0:p.Asp335His  
GMQ\_03045T0:p.Arg136Thr

**Results preview**

**Navigation (per variant)** **Filters** **Download** **New job**

Show: 1 5 10 50 All variants  
Protein matches is defined  
Consequence is missense\_variant  
Clear filters Match all of the above rules Update  
Uploaded variant is defined Add

All: VCF VEP TXT  
Filtered: VCF VEP TXT

Show/hide columns (22 hidden)

Location	Allele	Consequence	Gene	Exon	HGVSc	HGVSp	Codons	Existing variant	ENSP	Protein matches
Superconfig_3.156.127654-127654	C	missense_variant	GMQ_21813	2/6	GMQ_21813T0:c.235T>C	GMQ_21813T0:p.Ser79Pro	TCC/CCC	tmp_Superconfig_3.156.127654_T_C	GMQ_21813T0	Superfamily:SSF50022 PROSITE_profiles:PS51296 Gene3D:2.102.10.10
Superconfig_3.1594.801-801	T	missense_variant	GMQ_27112	3/3	GMQ_27112T0:c.265G>A	GMQ_27112T0:p.Glu89Lys	GAG/AAG	tmp_Superconfig_3.1594.801_C_T	GMQ_27112T0	MobiDB-lite:mobidb-lite Pfam:PF14892
Superconfig_3.41.7765-7765	G	missense_variant	GMQ_06767	6/15	GMQ_06767T0:c.1328A>C	GMQ_06767T0:p.Gln443Pro	CAG/CCG	tmp_Superconfig_3.41.7765_T_G	GMQ_06767T0	PANTHER:PTHR46896
Superconfig_3.427.55213-55213	A	missense_variant	GMQ_02814	2/2	GMQ_02814T0:c.99G>T	GMQ_02814T0:p.Gln33His	CAG/CAT	tmp_Superconfig_3.427.55213_C_A	GMQ_02814T0	MobiDB-lite:mobidb-lite
Superconfig_3.58.69935-69935	C	missense_variant	GMQ_20457	4/4	GMQ_20457T0:c.1003G>C	GMQ_20457T0:p.Asp335His	GAT/CAT	tmp_Superconfig_3.58.69935_G_C	GMQ_20457T0	Superfamily:SSF53649 PANTHER:PTHR23071 CDD:cd16023 Gene3D:3.40.720.10
Superconfig_3.73.160474-160474	G	missense_variant	GMQ_03045	2/3	GMQ_03045T0:c.407G>C	GMQ_03045T0:p.Arg136Thr	AGA/ACA	tmp_Superconfig_3.73.160474_C_G	GMQ_03045T0	PANTHER:PTHR31595 Low_complexity_(Seg).seg

(e) Ensure you selected the following additional configuration in the VEP input form:

Identifiers: [UniProt](#) (to display any associated UniProt accession IDs, including UniProtKB/Swiss-Prot and UniParc)

Apply the [Consequence is frameshift\\_variant](#) under 'Filters'. There is 1 frameshift variant which falls in the GMQ\_27001 gene on exon 1 (out of 3). The UniParc ID is UPI0004E9C5AE.

Results preview

Navigation (per variant)    Filters    Download    New job

Show: 1 5 10 50 All variants    Consequence is frameshift\_variant    All: VCF VEP TXT  
Uploaded variant is defined Add    Filtered: VCF VEP TXT

Show/hide columns (27 hidden)

Location	Allele	Consequence	Gene	Exon	Codons	Existing variant	ENSP	SWISSPROT	UNIPARC
Supercontig_3.1482:1095-1097	-	frameshift_variant	GMQ_27001	1/3	CCG/CG	Imp_Supercontig_3.1482_1095_CGG(CG	GMQ_27001T0	-	UPI0004E9C5AE#P

Show: 1 5 10 50 All variants

## Exercise: Ensembl Fungi whole-genome alignments

Links to be clicked shown in blue, text to be entered shown in red.

Ensembl Fungi contains whole genome alignments for pairs of key species, generated using LastZ. Let's look at some of these comparative genomics views in the Location tab.

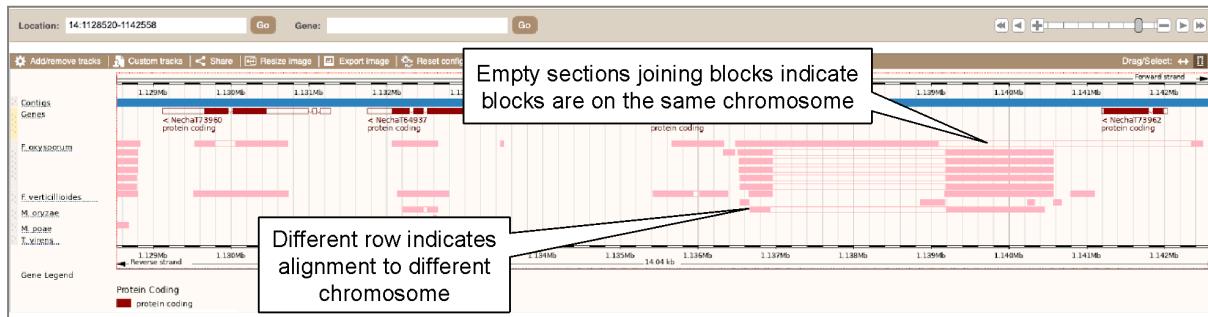
- (a) Find the region **14:1128520-1142558** in *Fusarium solani* and go to the [Region in detail](#) page. This region includes four genes we identified from our first BioMart query: *PEP5*, *PDA1*, *ESP3* and *PEP5*.

The screenshot shows the Ensembl Fungi search interface. In the search bar, 'Fusarium solani' is entered, followed by a dropdown menu set to 'for'. Below the search bar is a text input field containing '14:1128520-1142558' and a brown 'Go' button. Below the input field is the text 'e.g. NAT2 or alcohol\*'. The background is light beige.

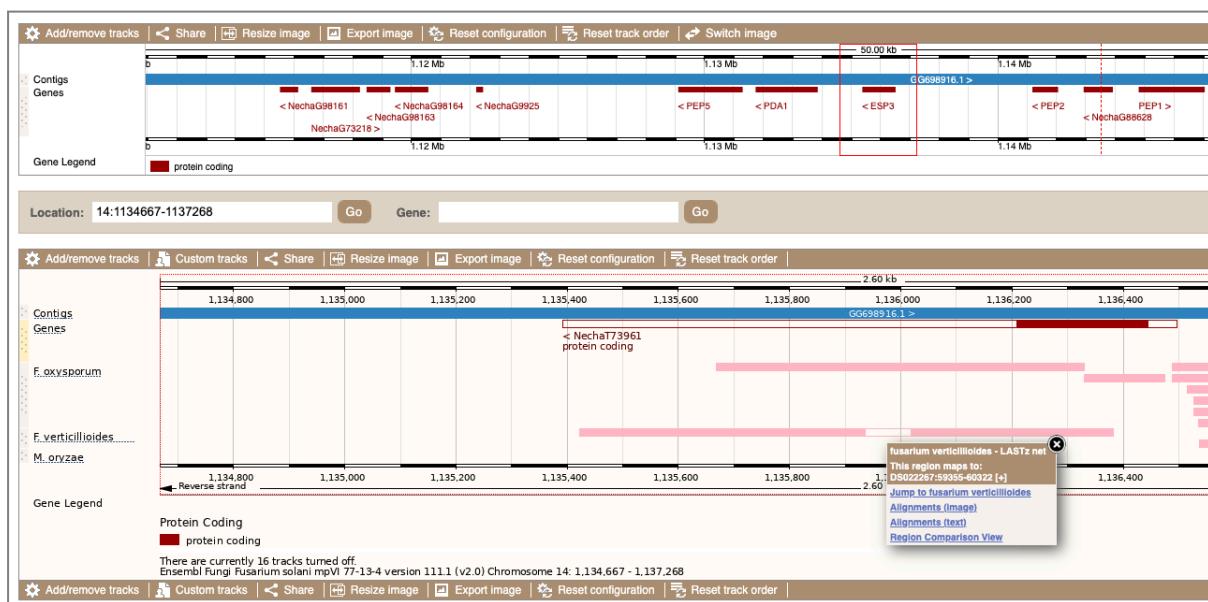
We can look at individual species' comparative genomics tracks in this view by clicking on [Configure this page](#). In the 'Comparative genomics' section, turn on all of the available species' alignments in the normal style.

The screenshot shows the 'Region in detail' page for *Fusarium solani*. On the left, there is a sidebar with various links like 'Location-based displays', 'Comparative Genomics', and 'Genetic Variation'. The main content area has a header 'Comparative genomics' with a sub-section 'Enable/disable all tracks'. A list of species is shown with checkboxes next to their names. To the right, there is a large area for viewing comparative genomic tracks, which is currently empty. At the bottom, there is a 'Key' section with icons for 'Track style', 'Forward strand', 'Reverse strand', 'Favourite track', and 'Track information'.

We can now see some pink alignments shown on the display. Alignments to the same chromosome are presented in a single row, and gaps in the alignment are shown by linking blocks. If there are alignments to multiple chromosomes in the aligned species these are represented on different rows.



- (b) Looking at the pink alignment blocks, does this region in *F. solani* s align to multiple different chromosomes in the other species?
  - (c) Which chromosome(s) does the *F. solani* *ESP3* gene align to in *F. verticillioides*?



We can see that alignments in this region are quite poor for these species, with alignments spanning different chromosomes. This supports the lack of orthologues between these species.

We can view more detailed alignments in the alignment's text/image and region comparison views. Let's first view a text alignment in this region. Click on [Alignments \(text\)](#) on the left and choose *Fusarium verticillioides* from the drop-down menu.

Because this single chromosome region in *F. solani* aligns to regions that are far spread in other genomes, you need to select a specific block for the alignment, as we cannot display a single sequence alignment from more than one region.

The screenshot shows the 'Alignments (text)' section of the Ensembl Fungi WGA interface. At the top, there is a search bar for 'Alignment' set to 'Fusarium verticillioides - lastz' with a 'Go' button. Below it, 'Location' is set to '14:1128520-1142558' and 'Gene' is empty, with 'Go' buttons for both. A chromosome browser track is visible. A download button is present. A message indicates 'A total of 11 alignment blocks have been found.' Below this, a table lists the 11 blocks ordered by size. The table has columns for 'Alignment (click to view)', 'Length (bp)', 'Location on *Fusarium solani*', and 'Location on *Fusarium verticillioides*'. Each row contains a link to 'Block 1' through 'Block 11'. A note says 'Entire alignment can be viewed by selecting a Block from the Alignment column.' A callout box points to the table with the text 'Blocks ordered by size'. Another callout box points to the right side of the table with the text 'F. verticillioides alignment regions across the region'. A final callout box points to the bottom of the table with the text 'All *F. solani* alignment regions on chromosome 14'.

Alignment (click to view)	Length (bp)	Location on <i>Fusarium solani</i>	Location on <i>Fusarium verticillioides</i>
<a href="#">Block 1</a>	1395	<a href="#">14:1139178-1140572</a>	<a href="#">9:1319698-1321143</a>
<a href="#">Block 2</a>	1218	<a href="#">14:1129517-1130734</a>	<a href="#">11:1354930-1356096</a>
<a href="#">Block 3</a>	961	<a href="#">14:1135422-1136382</a>	<a href="#">DS022267:59355-60322</a>
<a href="#">Block 4</a>	662	<a href="#">14:1132135-1132796</a>	<a href="#">10:1292005-1292692</a>
<a href="#">Block 5</a>	326	<a href="#">14:1138852-1139177</a>	<a href="#">5:2632133-2632458</a>
<a href="#">Block 6</a>	305	<a href="#">14:1140792-1141096</a>	<a href="#">1:1367865-1368184</a>
<a href="#">Block 7</a>	299	<a href="#">14:1136656-1136954</a>	<a href="#">3:163711-164009</a>
<a href="#">Block 8</a>	275	<a href="#">14:1128520-1128794</a>	<a href="#">2:124185-124450</a>
<a href="#">Block 9</a>	119	<a href="#">14:1136537-1136655</a>	<a href="#">DS022270:2615-2734</a>
<a href="#">Block 10</a>	101	<a href="#">14:1140573-1140673</a>	<a href="#">DS022267:3013-3103</a>
<a href="#">Block 11</a>	88	<a href="#">14:1140238-1140325</a>	<a href="#">3:4306258-4306345</a>

Let's click on [Block 3](#). This takes you to a new page with a sample of the aligned sequence. Then click the button to [Display full alignment](#). You will see a list of the regions aligned, followed by the sequence alignment. Exons are shown in red. Click on [Configure this page](#), you can turn on the options to view [Show conservation regions](#) and [Mark alignment start/end](#). Remember to click the tick at the top right when closing this window to save your choices. This will add highlights where the sequence matches.

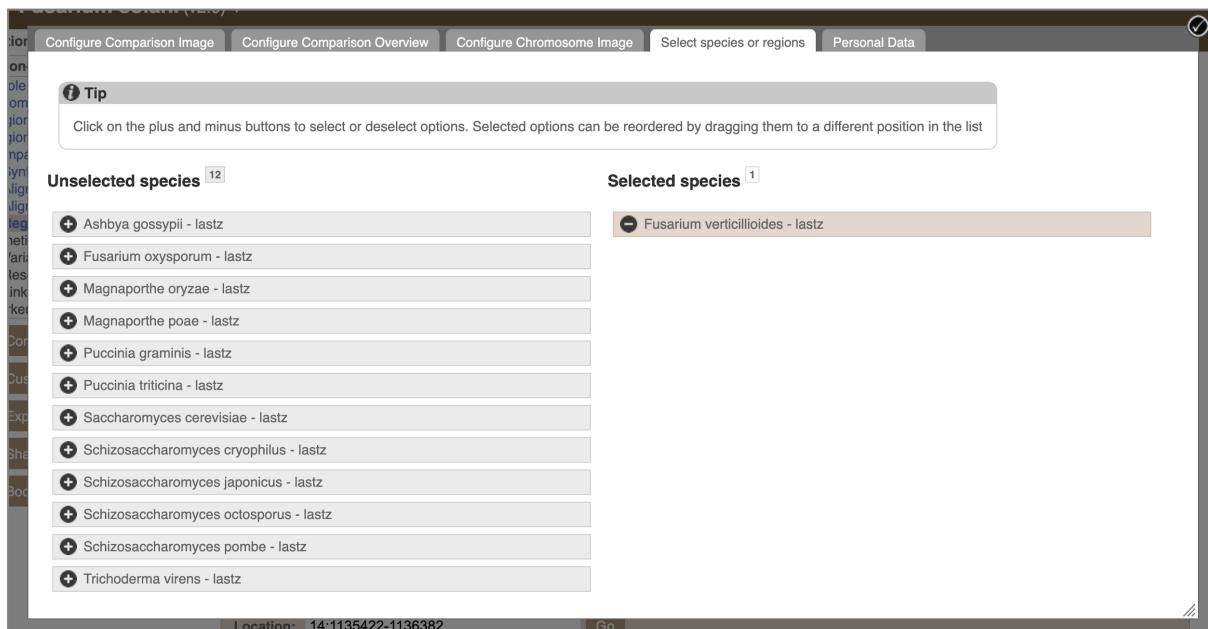
The screenshot shows the 'Display options' configuration menu. It includes buttons for 'Display options', 'Manage configurations', and 'Reset configuration'. A dropdown menu 'Select from available configurations:' is set to 'Default'. A 'Save configurations and close pop-up menu' button is at the top right. The main area contains several configuration options with dropdown menus:

- Strand: Forward
- Number of base pairs per row: 120 bps
- Additional exons to display: Core exons
- Orientation of additional exons: Display exons in both orientation
- Line numbering: None
- Codons: Do not show codons
- Show conservation regions:
- Mark alignment start/end:

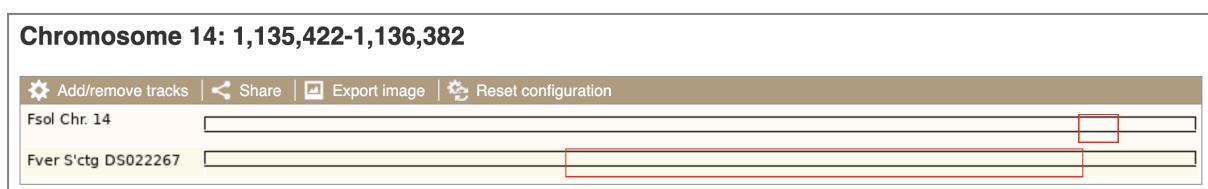
Comparative Genomics & Orthology (WGA): Ensembl Fungi

To view an image of the alignments, click on [Region comparison](#) in the left-hand navigation panel. This view is like the ‘Region in detail’ page as it shows three images of the genome at different scales. You can add multiple species to this view.

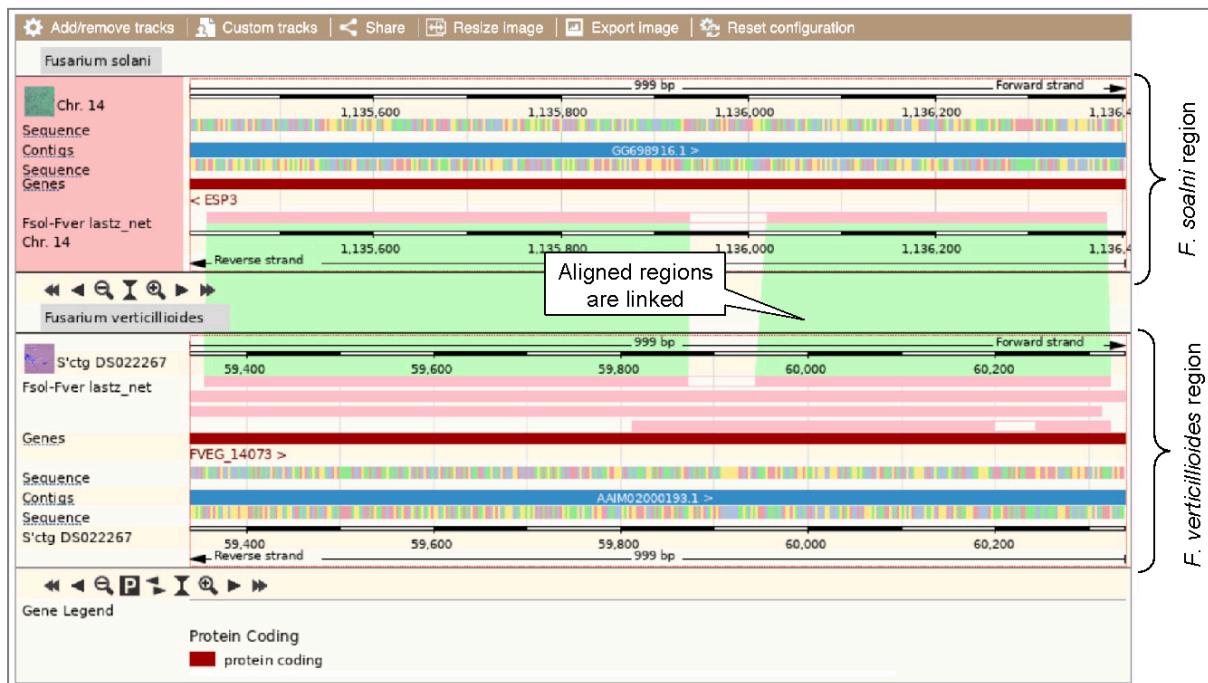
Click on the brown **Select species or regions** button. In the pop-up menu, select *Fusarium verticillioides* from the list. Close the window.



This page, similar to the region in detail page, shows the chromosome positions first. We can see the location of this alignment on the scaffold in *F. verticillioides*.



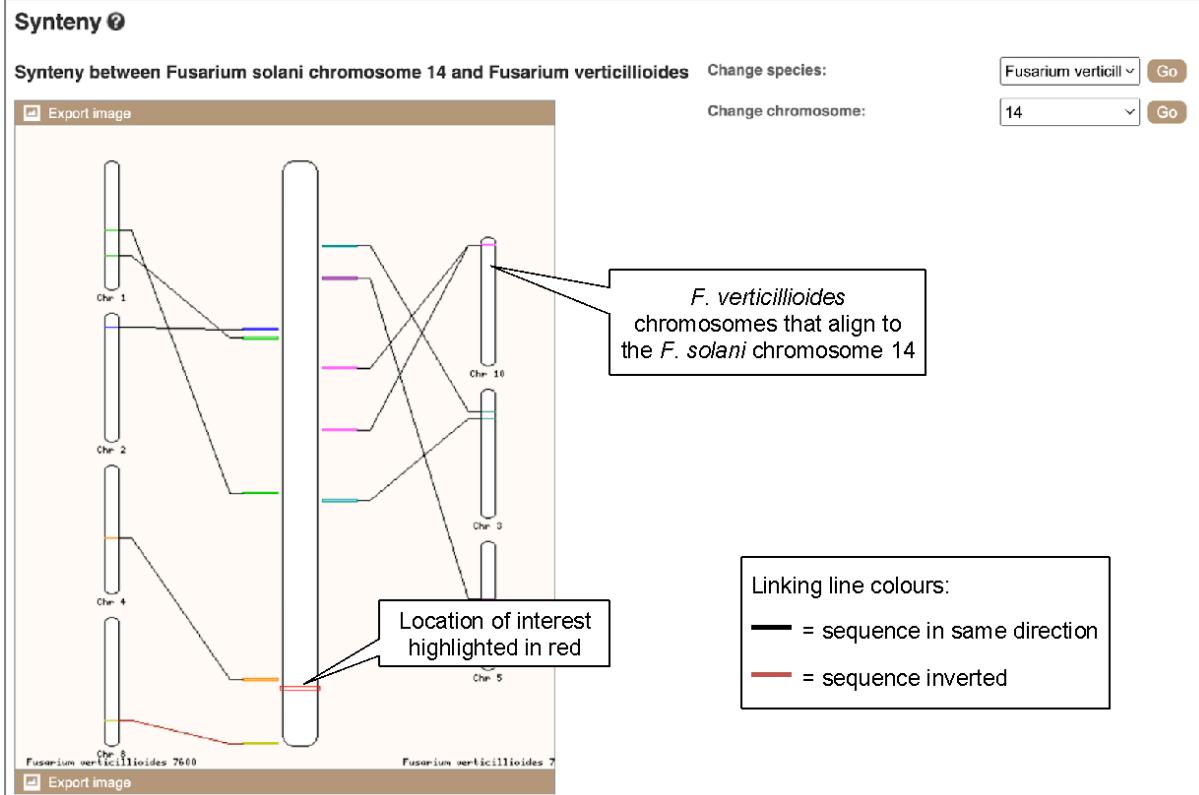
Scroll down to the most detailed image. An example image (of another alignment block) is below, and you should see something similar on your browser.



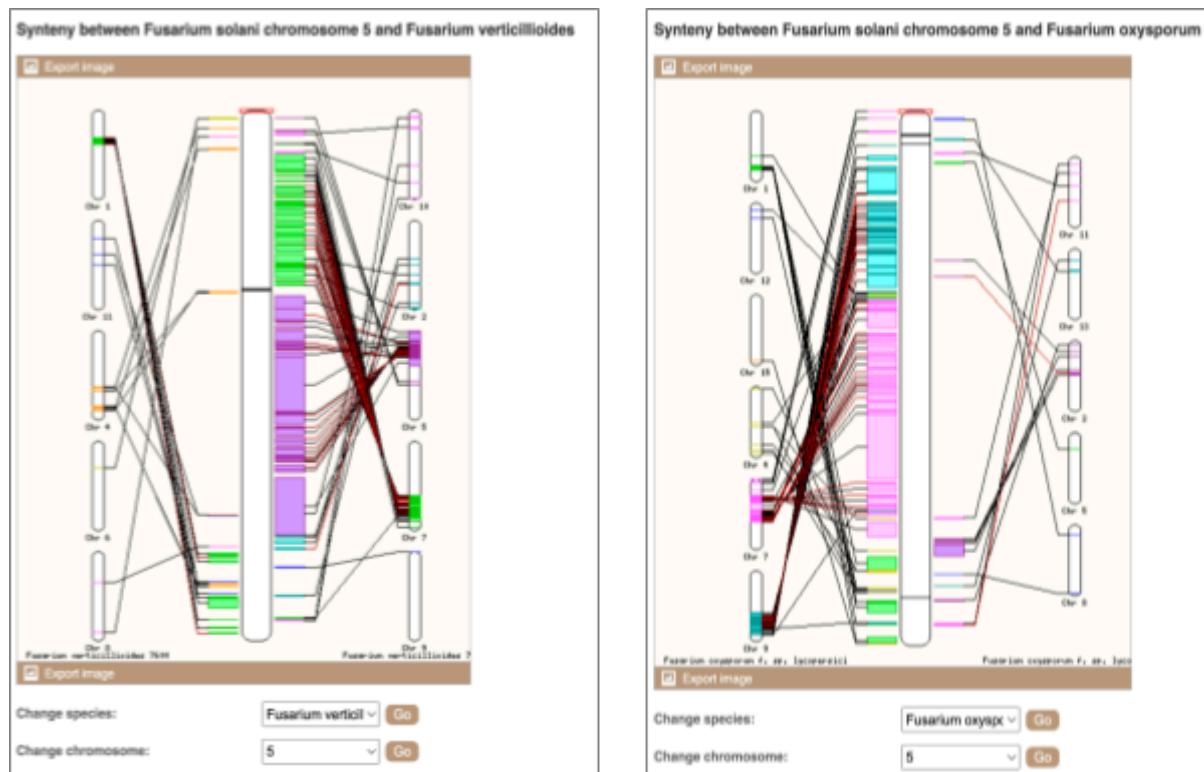
You can add data to both of these views with the same options you had in the ‘Region in detail’ page. Click on [Configure this page](#) and look at the top of the menu.

The configuration menu allows selecting species and configurations. The 'Species to configure:' dropdown includes 'Fusarium solani' and 'Fusarium verticillioides'. The 'Select from available configurations:' dropdown is set to 'Default'.

We can view chromosomal rearrangements in the ‘Synteny’ view. Click on [Synteny](#) in the left-hand navigation panel. (note - there is a bug causing this view to fail to load in later Ensembl versions, you can access the view in this archive - [https://feb2023-fungi.ensembl.org/Fusarium\\_solani/Location/Synteny?db=core;otherspecies=Fusarium\\_verticillioides;r=14:1128520-1142558](https://feb2023-fungi.ensembl.org/Fusarium_solani/Location/Synteny?db=core;otherspecies=Fusarium_verticillioides;r=14:1128520-1142558))



- (d) Which chromosome in *F. verticillioides* is most similar to *F. solani* chromosome 5?  
Change the display to show *F. oxysporum*. Does this give you the same answer as for *F. verticillioides*?



## Additional Exercise - Rearrangements in *Magnaporthe* species

In the publication '[PacBio sequencing reveals transposable elements as a key contributor to genomic plasticity and virulence variation in \*Magnaporthe oryzae\*](#)', Bao et al (2017) identified a region on chromosome 1 that is shown to be a region of inter-chromosomal rearrangement and inversion. We're going to take a look at this region and see how it looks in *Magnaporthe oryzae* and *Magnaporthe poae*.

- Search for the region **1:5603535-5611402** in *Magnaporthe oryzae*.
- Click on [Region comparison](#) and choose *Magnaporthe poae* from the [Select species or regions](#) pop-up to display an alignment.
- Scroll down to the most detailed image. To what region (chromosome/scaffold/contig) does this region align to on the *M. poae* assembly?
- Which genes are present in the aligned region for *M. oryzae* and *M. poae*? What are their biotypes?

## Answer - Rearrangements in *Magnaporthe* species

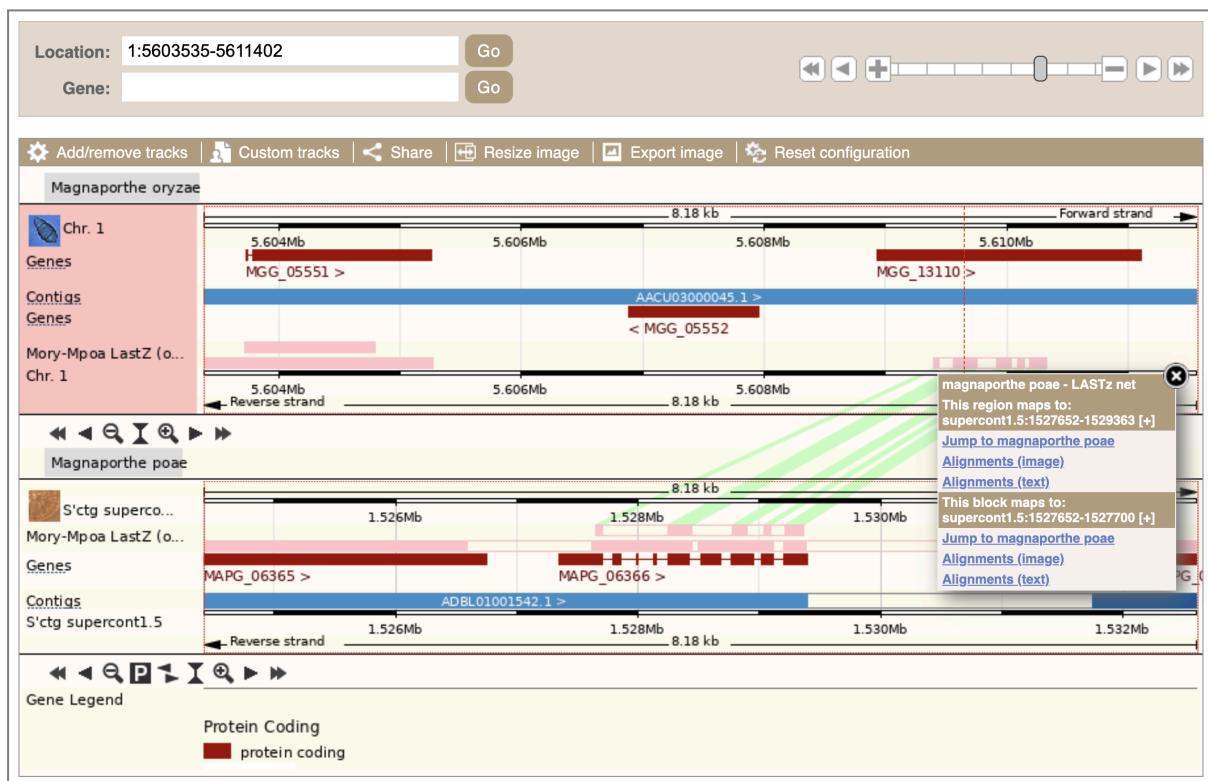
- Go to [fungi.ensembl.org](http://fungi.ensembl.org) in your browser and search for the region **1:5603535-5611402** in *Magnaporthe oryzae*.

Search: Magnaporthe oryzae  
1:5603535-5611402 Go  
e.g. NAT2 or alcohol\*

- Click on [Region Comparison](#) in the left-hand panel. Click on [Select species or regions](#) and select [Magnaporthe poae - lastz](#) in the pop-up menu to display the alignment.

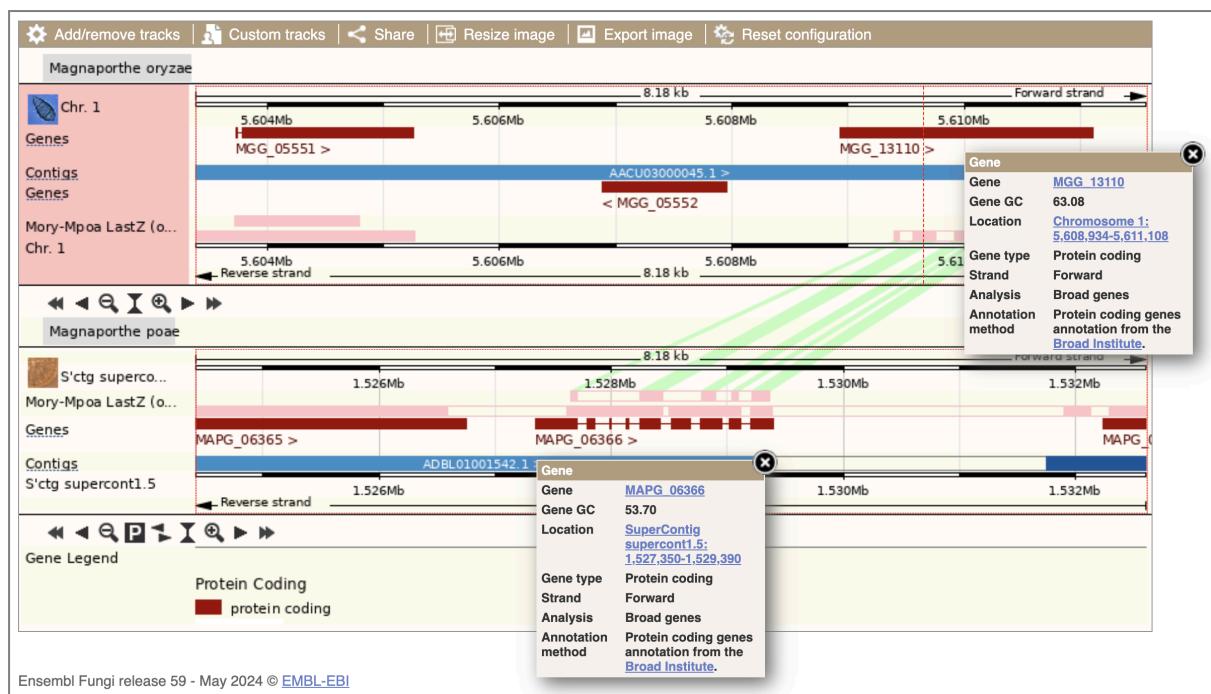
## Comparative Genomics & Orthology (WGA): Ensembl Fungi

- (c) Scroll down to the ‘Region in detail’ view. The region aligns to SuperContig (S’ctg) supercont1.5 in the *M. poae* assembly.



- (d) In the ‘Genes’ track, find out which features overlap the alignment regions. Click on the feature to find out more information. In *M. oryzae*, the gene MGG\_13110 is present. In *M. poae*, the gene MAPG\_06366 is present. Both genes are protein coding.

## Comparative Genomics & Orthology (WGA): Ensembl Fungi



# MycoCosm: Comparative Analysis of Gene Families

**Objective:** Compare genomes of wood decay fungi to identify gene families which can be used to distinguish white rot and brown rot fungi

Many fungi of the phylum Basidiomycota are capable of degrading wood, including the recalcitrant polymer lignin, which gives wood its structural strength and resistance to microbial attack (Floudas et al. 2012; Riley et al. 2014). These wood decaying fungi are often classified as either **white rot**, in which lignin is completely degraded and cellulose is left somewhat intact; or **brown rot**, in which cellulose is degraded and lignin is left somewhat intact. While the precise enzymatic mechanisms vary from one fungus to another, in general white rot genomes encode class II peroxidase enzymes to break down lignin, carbohydrate-binding motif enzymes to bind cellulose, and glycoside hydrolases to break down cellulose. By contrast, brown rot genomes tend to have relatively reduced numbers of these enzymes, or even lack them entirely.

Suppose we are comparing the genomes of four wood decaying fungi: *Auricularia subglabra*, *Calocera cornea*, *Gloeophyllum trabeum*, *Phanerochaete chrysosporium* RP-78. Suppose, also, that we don't know which of them are white-rot or brown-rot fungi. How can we use MycoCosm to make predictions about their mode of decay?

Start by going to the genome group page created for this example (in real life we would use a similar genome group page, but with a larger, ecologically- or phylogenetically-relevant selection of organisms):

[https://mycocosm.jgi.doe.gov/WR\\_BR\\_example\\_2017/](https://mycocosm.jgi.doe.gov/WR_BR_example_2017/)

Info • White rot/brown rot example 2017						
SEARCH	BLAST	ANNOTATIONS	MCL CLUSTERS	GEO MAPPING	DOWNLOAD	INFO
<a href="#">HELP!</a>						
##	Name		Assembly Length	# Genes	Published	
1	<a href="#">Auricularia subglabra v2.0</a>		76,853,599	25,459	<a href="#">Floudas D et al., 2012</a>	
2	<a href="#">Calocera cornea v1.0</a>		33,244,933	13,177	<a href="#">Nagy LG et al., 2016</a>	
3	<a href="#">Gloeophyllum trabeum v1.0</a>		37,181,821	11,846	<a href="#">Floudas D et al., 2012</a>	
4	<a href="#">Phanerochaete chrysosporium RP-78 v2.2</a>		35,149,519	13,602	<a href="#">Ohm RA et al., 2014</a>	

## CAZy browser

CAZymes (Carbohydrate-Active Enzymes) are enzymes that degrade, modify, and/or create glycosidic bonds (Levasseur et al. 2013). They can be classified into families of structurally-

related catalytic and carbohydrate-binding modules (or functional domains). The classifications used by the CAZy database are incorporated into MycoCosm for comparative analyses.

Click on the CAZYMES item under ANNOTATIONS in the Main menu.

The screenshot shows the CAZymes search interface. At the top, there's a navigation bar with links for SEARCH, BLAST, ANNOTATIONS (which is currently selected), MCL CLUSTERS, GEO MAPPING, DOWNLOAD, INFO, and HELP!. Below the navigation bar is a search bar with fields for 'Search for:' and 'Keywords', and buttons for 'Filter' and 'Exact'. There are also checkboxes for 'Show only filtered results' and 'Show only filtered totals', and a 'Display Type' dropdown set to 'Table'. A sidebar on the left lists categories: PFAM DOMAINS, SECONDARY METABOLISM CLUSTERS, CAZYMES, PEPTIDASES, TRANSPORTERS, and TRANSCRIPTION FACTORS. The CAZYMES category is expanded, showing its sub-categories. The main content area displays a table of predicted CAZymes across four species: *Aureobasidium*, *Calco*, *Glotr*, and *Pchcr*. The table includes columns for 'Annotations/Genomes', 'Aureobasidium', 'Calco', 'Glotr', 'Pchcr', 'Total', and 'Annotation Description'. The data shows counts for various CAZy families, such as CAZy (2,008 total), AA (292 total), CBM (231 total), and CE (109 total).

Annotations/Genomes	Aureobasidium	Calco	Glotr	Pchcr	Total	Annotation Description
CAZy	827	350	368	463	2,008	CAZy
AA	130	27	43	92	292	Auxiliary Activities family
CBM	123	18	19	71	231	Carbohydrate-Binding Module family
CE	61	14	14	20	109	Carbohydrate Esterase family

Here you will see a table representation of the predicted CAZymes in each species. The organisms are labeled along the top by genome portal identifier (“portal ID”). The CAZymes are organized hierarchically by family and labeled along the sides: CAZy family identifier on the left, and family description on the right. The numbers in the table represent how many proteins from each organism’s gene catalog were annotated with a given CAZyme, with a total provided for each row. Notice that the CAZymes are hierarchically organized: you can see the total number of genes assigned to the top level enzyme category (e.g. “AA”). To view family (e.g. “AA1”, “AA2”) and subfamily (e.g. “AA1\_1”, “AA1\_2”) designations, click on the small arrow to the left of each category, or use the “Expand All” button at the top of the page.

This screenshot shows a more detailed view of the CAZymes search interface. The layout is similar to the previous one, with a navigation bar, search bar, and sidebar. The main content area displays a table with the same structure as the first screenshot, but it includes more detailed information for the CAZy family. The table shows data for the AA, AA1, AA2, and AA3 families across the four species. For example, under the AA family, it shows subfamilies AA1, AA1\_1, AA1\_2, AA1\_3, AA1\_dist, AA2, AA2\_dist, AA3, AA3\_1, and AA3\_2, with their respective counts and descriptions. The 'Display Type' is set to 'Table'.

Annotations/Genomes	Aureobasidium	Calco	Glotr	Pchcr	Total	Annotation Description
CAZy	848	352	372	466	2,038	CAZy
AA	131	29	44	93	297	Auxiliary Activities family
AA1	10	5	5	5	25	Auxiliary Activity Family 1
AA1_1					4	Auxiliary Activity Family 1 / Subf 1
AA1_2		2	1	1	4	Auxiliary Activity Family 1 / Subf 2
AA1_3			7		7	Auxiliary Activity Family 1 / Subf 3
AA1_dist		1			1	Multicopper oxidase
AA2	20	1	1	17	39	Auxiliary Activity Family 2
AA2_dist	1	1	1	1	4	Class II peroxidase
AA3	50	15	24	39	128	Auxiliary Activity Family 3
AA3_1	1	1	1		3	Auxiliary Activity Family 3 / Subf 1
AA3_2	38	13	20	34	105	Auxiliary Activity Family 3 / Subf 2

If we read Levasseur et al. 2013, we know that the AA2 family consists of peroxidases that may degrade lignin. Browsing the table, we see that *P. chrysosporium* and *A. subglabra* possess 20 and 17 copies of AA2, whereas *G. trabeum* and *C. cornea* each possess only one copy of AA2. This might suggest that the former two are white rot fungi and the latter two brown rot fungi!

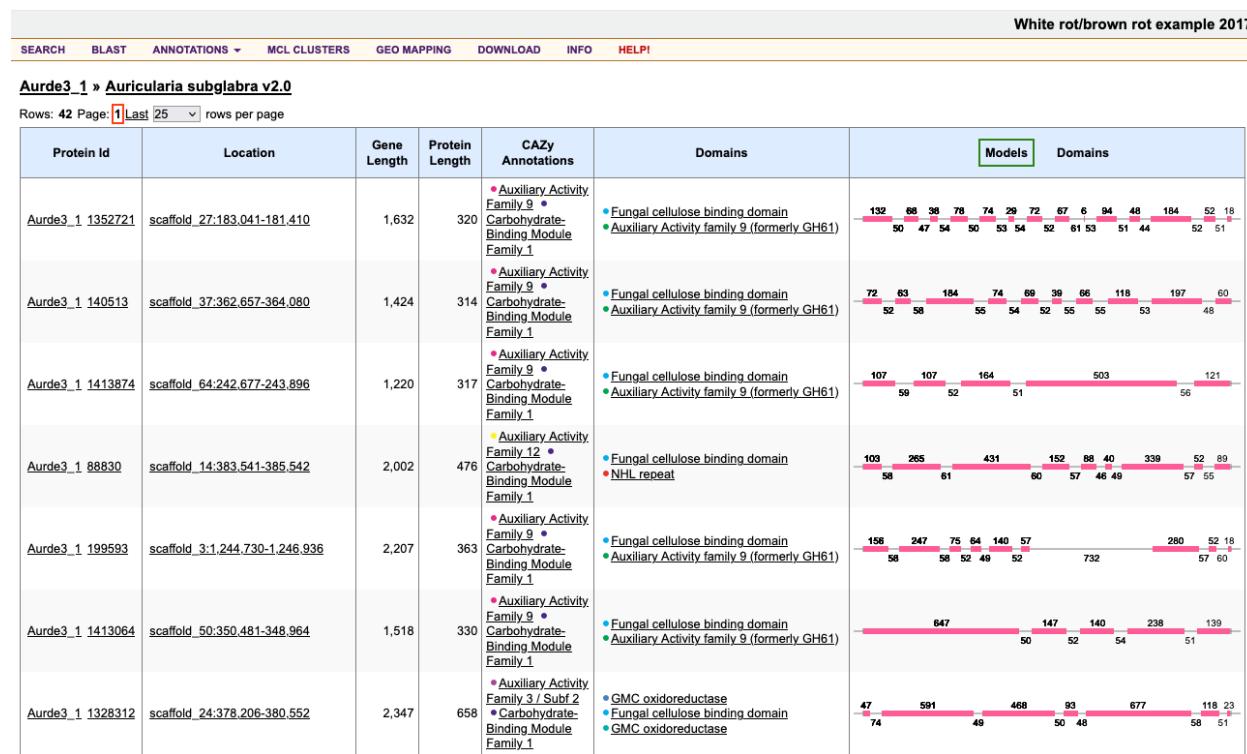
What about the carbohydrate binding motifs, CBM1? Let's say we don't want to scroll through the entire list of CAZymes. Type "CBM1" into the "CAZY terms" search box and click "Filter". This will limit the view to only those CAZymes that have a CBM1. Why do so many CAZymes besides CBM1 show up? Because CBM1 co-occurs on the same protein chain with many other CAZymes of diverse function. The numbers in the table will now show, for each CAZyme's row, the number of proteins that also have a CBM1.

The screenshot shows the CAZy database search interface. At the top, there are tabs for SEARCH, BLAST, ANNOTATIONS (with a dropdown), MCL CLUSTERS, GEO MAPPING, DOWNLOAD, INFO, and HELP!. Below the tabs, a search bar contains the text "CBM1" under the heading "Search for:". To the right of the search bar are buttons for "Any", "Keywords", "Filter", "Exact", and "Clear". The main content is a table with the following columns: Annotations/Genomes, Aurde3\_1, Calco1, Glotr1\_1, Phchr2, Total, and Annotation Description. The table lists various CAZy families and their counts, with rows for CAZy, AA, AA3, AA3\_2, AA8, AA9, AA12, CBM, CBM1, CE, CE1, CE5, CE15, CE16, GH, GH3, GH5, GH5\_5, GH5\_7, GH6, GH7, GH10, GH11, and GH12. The CBM1 row shows a count of 86 in the Total column.

Annotations/Genomes	Aurde3_1	Calco1	Glotr1_1	Phchr2	Total	Annotation Description
CAZy	83	2	2	68	155	CAZy
AA	8		7		15	Auxiliary Activities family
AA3	2				2	Auxiliary Activity Family 3
AA3_2	2				2	Auxiliary Activity Family 3 / Subf 2
AA8			1		1	Auxiliary Activity Family 8
AA9	5		6		11	Auxiliary Activity Family 9
AA12	1				1	Auxiliary Activity Family 12
CBM	48	1	1	36	86	Carbohydrate-Binding Module family
CBM1	48	1	1	36	86	Carbohydrate-Binding Module Family 1
CE	7		4		11	Carbohydrate Esterase family
CE1	1		2		3	Carbohydrate Esterase Family 1
CE5	2				2	Carbohydrate Esterase Family 5
CE15	3		1		4	Carbohydrate Esterase Family 15
CE16	1		1		2	Carbohydrate Esterase Family 16
GH	20	1	1	21	43	Glycoside Hydrolase family
GH3			1		1	Glycoside Hydrolase Family 3
GH5	4	1	4		9	Glycoside Hydrolase Family 5
GH5_5	3	1	2		6	Glycoside Hydrolase Family 5 / Subf 5
GH5_7	1		2		3	Glycoside Hydrolase Family 5 / Subf 7
GH6	2		1		3	Glycoside Hydrolase Family 6
GH7	4		6		10	Glycoside Hydrolase Family 7
GH10	2	1	4		7	Glycoside Hydrolase Family 10
GH11	2		1		3	Glycoside Hydrolase Family 11
GH12	1				1	Glycoside Hydrolase Family 12

Notice the abundance of CBM1-encoding genes in *P. chrysosporium* and *A. subglabra*, while *G. trabeum* and *C. cornea* have only a single CBM1-encoding gene each (co-occurring with GH5\_5 and GH10 proteins). All of this indicates that we might be looking at two white-rot and two brown-rot fungi.

Click on the number (e.g., 48 for Aurde3\_1) to see the CBM1-containing proteins of *A. subglabra* in more detail. Notice a variety of CAZymes co-occur with CBM1, including GH5 (various subfamilies), GH6, and many others.



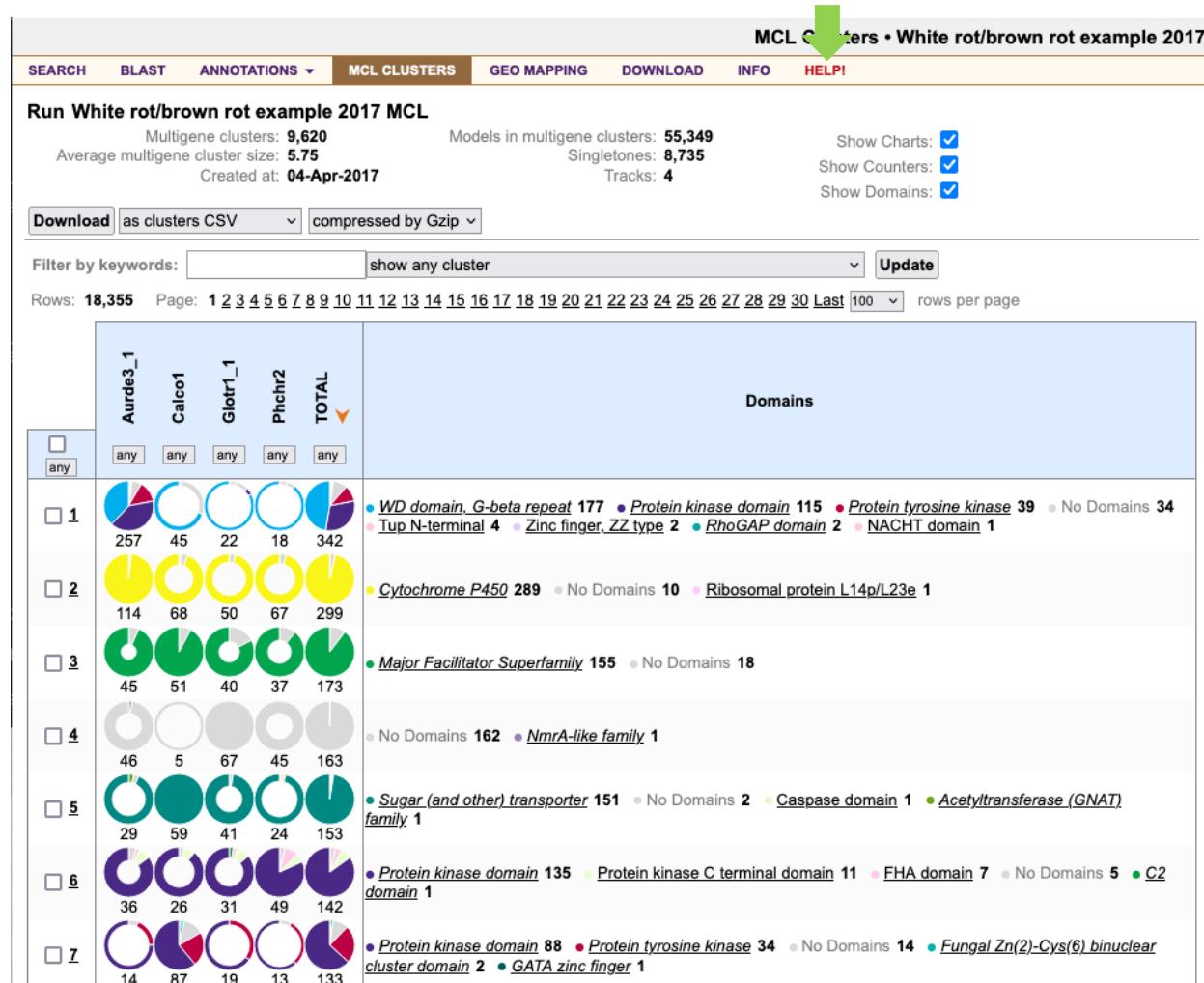
As an exercise, repeat the same search with GH6, GH7, and also the AA9 family of lytic polysaccharide monooxygenases, which may oxidatively act on lignin (Levasseur et al. 2013). Do the presence/absence patterns of these genes indicate the same conclusions about these fungi's mode of decay as we found with AA2 and CBM1? Is it a strict dichotomy, or are there some grey areas in the distribution of these genes?

(Answer: *P. chrysosporium* and *A. subglabra* induce white rot wood decay; *G. trabeum* and *C. cornea* brown rot. Notice that brown rot *G. trabeum* has a few AA9 genes, however, indicating that these genes may play a role in brown rot, not just white rot, where AA9s are expanded.)

## Cluster page

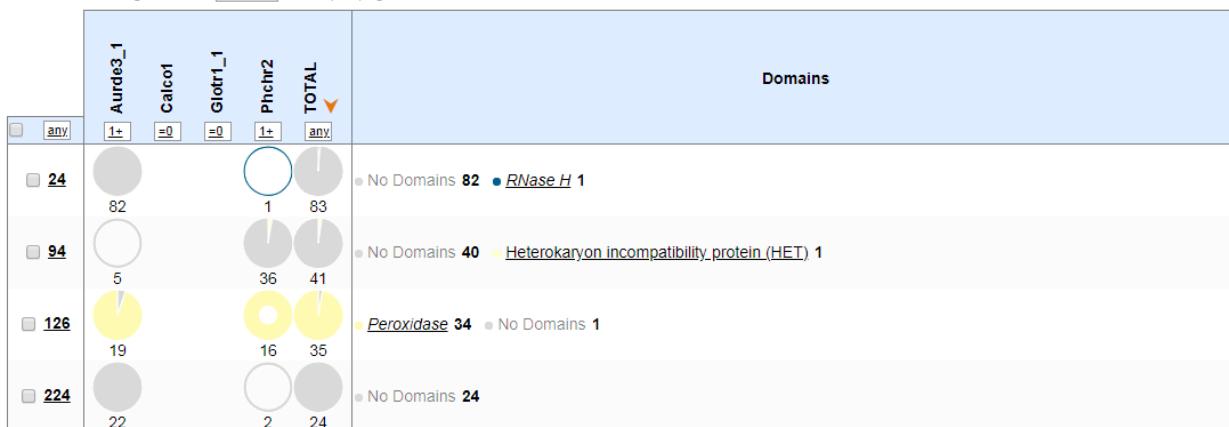
Now that we have an idea which fungus uses which decay mode, let's ask the reverse question: what are the genes present in one lifestyle, and absent in the other? To do this, click the 'MCL CLUSTERS' item of the Main menu. Here you will see the results of protein sequence clustering by the TRIBE-MCL method (Enright et al. 2002). As with the CAZy browser, the columns indicate organisms. Each row indicates a single protein cluster (analogous to a protein family),

where the number corresponds to the proteins from each organism in the cluster. The donut charts provide visualizations for the relative number of proteins and functional content provided by each organism in the cluster. See the HELP Menu for a full explanation of the cluster page.



Notice that under each organism label is a button “any” that can be used to filter clusters by the number of proteins that organism contributes to a cluster, and thus limit which clusters are shown. As an experiment, set the white rot fungi (Aurde3\_1 and Phchr2) to “1+” and the brown rot fungi (Calco1 and Glotr1\_1) to “=0”. Doing so will return only those clusters which are present in Aurde3\_1/Phchr2 and absent in Calco1/Glotr1\_1.

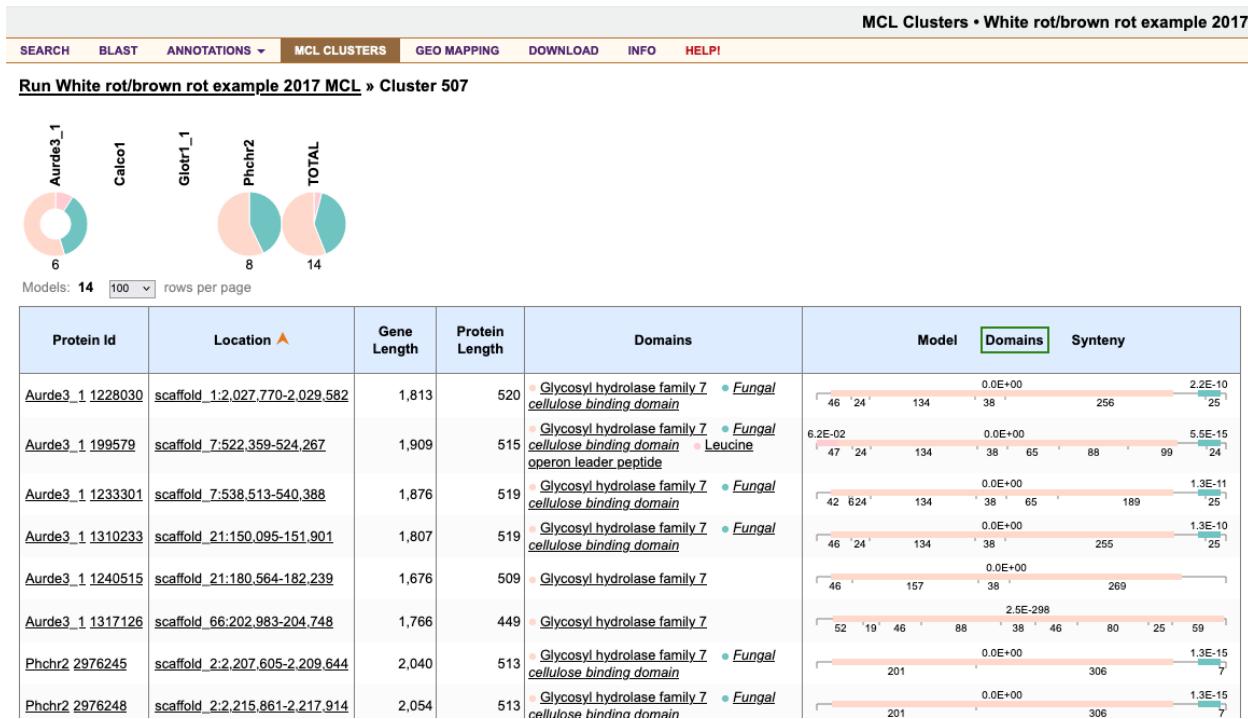
Rows: 150 Page: 1 Last 100 rows per page



150 clusters fit these criteria. These clusters might include genes important to the white rot decay mode, because they are present in white rot fungi and absent in brown rot fungi. However, some of these clusters might have no functional connection to wood decay mode - they are present/absent from the respective kinds of wood decay fungi merely by chance. These clusters nevertheless represent candidates for further analysis of possible connections to decay mode.

How does one begin interpreting the results? To help with this, each cluster row shows the Pfam domains (<https://www.ebi.ac.uk/interpro/entry/pfam>) that are found in that cluster. Notice that the third row has a “Peroxidase” (PF00141) domain. Notice that the numbers are very close to what we found for the AA2 class II peroxidases in the CAZy browser. It turns out that PF00141 is a superfamily that includes the AA2 enzymes, but it is important to note that not all members of PF00141 can degrade lignin - some have other functions.

Scroll through the rest of the 150 clusters and you will see domains such as “Glycosyl hydrolase family 7” and “Fungal cellulose binding domain” in cluster 507, which roughly overlap with the GH7 and CBM1 families from the CAZy exercise. Click the “507” to explore that cluster in more detail. On the cluster detail page, a table is presented with one protein per row. Click the “Domains” view on the rightmost column to see the domain structure of each protein. Notice that all of the proteins have the GH7 domain, and that most (but not all) have a single CBM1 motif at the C-terminus.

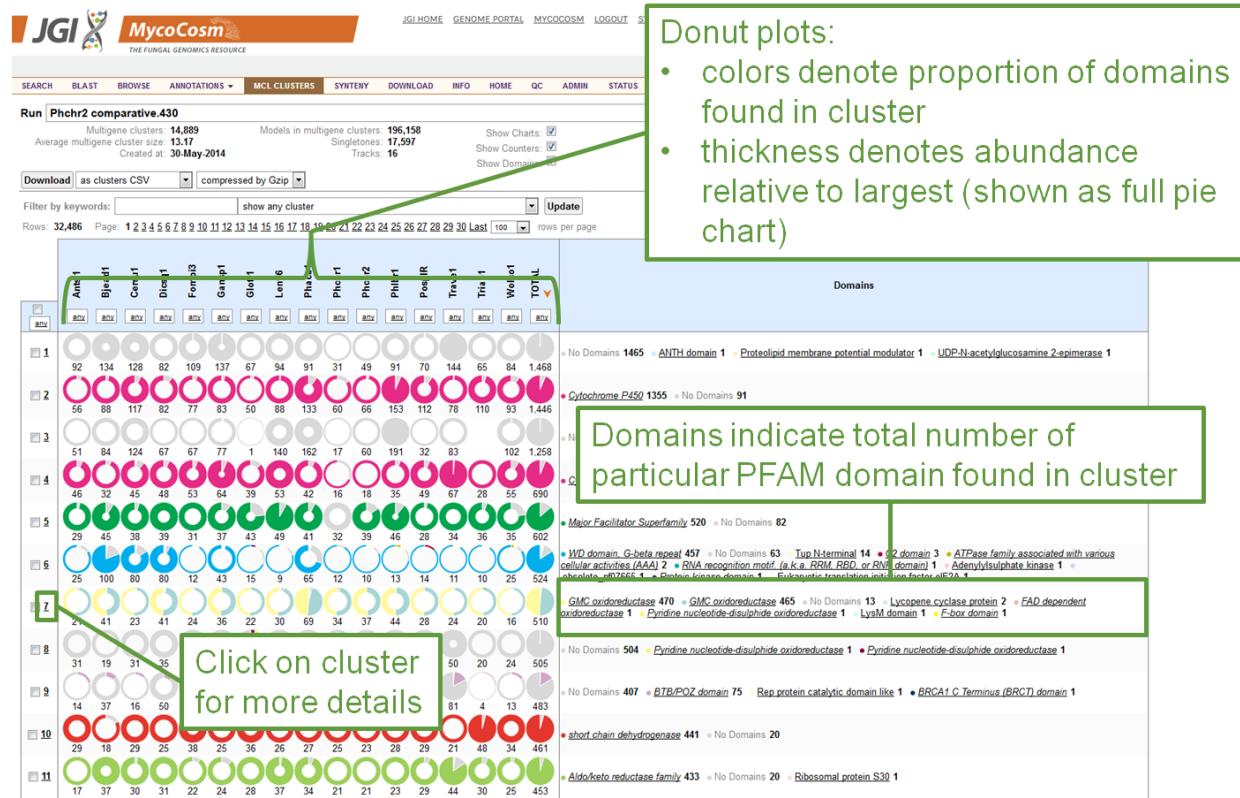


Let's look at what other proteins have the CBM1 carbohydrate-binding motifs in them.

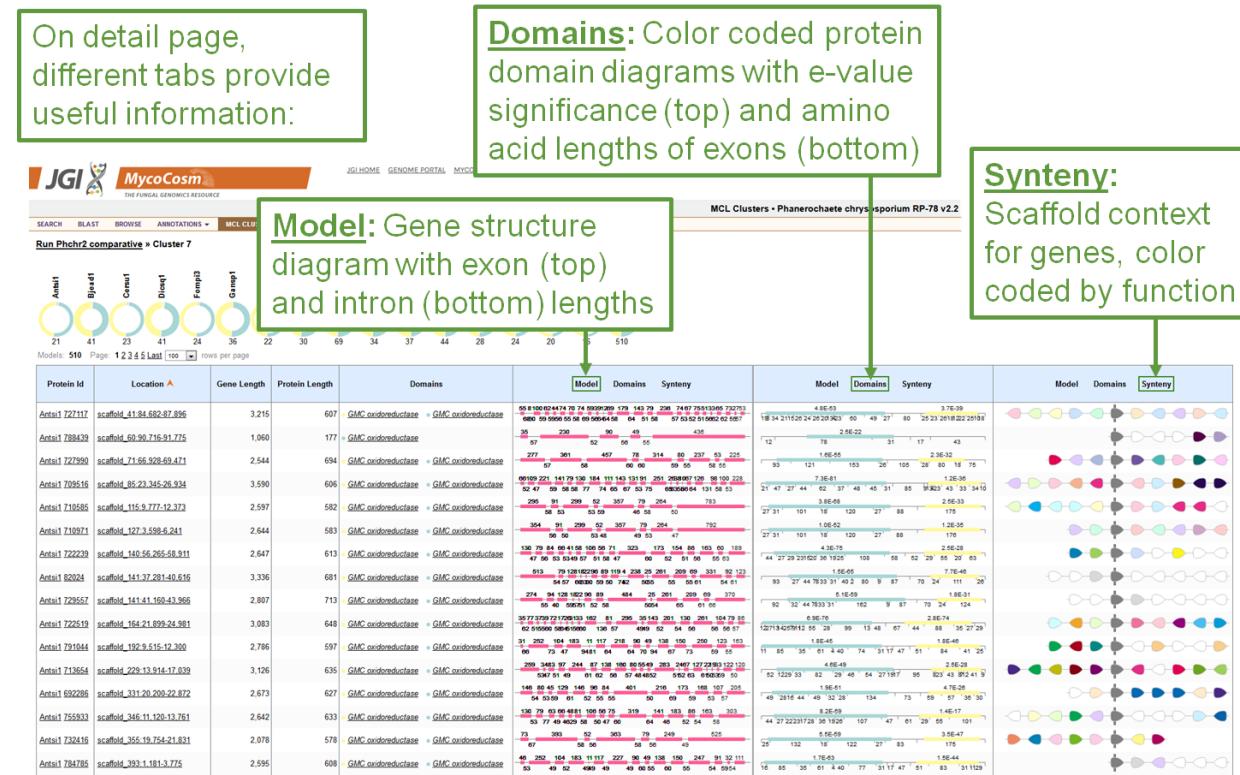
Returning to the cluster run page (click the “MCL CLUSTERS” tab). Enter the phrase “fungal cellulose binding domain” (be sure to include the quotes) into the “filter by keywords” field and select “Update”. This returns some 26 clusters, all of which have the Pfam domain CBM\_1 (PF00734). We see that CBM1 motifs occur in a wide array of domain combinations: often with GMC oxidoreductases, AA9 lytic polysaccharide monooxygenases (formerly Glycosyl hydrolase family 61), and many hydrolytic enzymes such as GH5, GH6, and GH7. Notice that while these proteins typically are found in expanded copy number in the white rot fungi (Aurde3\_1 and Phchr2) they are sometimes found, albeit in lower copy number, in the brown rot fungi (Calco1 and Glotr1\_1).

As additional exercises you can (a) search for gene families absent in both white rot fungi; (b) find gene families absent in white rot but present in both brown rot fungi and look at functional domains associated with these families; (c) check if any of these domains are present only in brown rot fungi by resetting filters back to “any” and searching for names of these domains.

A summary of tools available in MCL clustering are shown below:



Clicking in Cluster number provides additional tools as shown below:



## References:

- Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R. A., Henrissat, B., Martinez, A. T., Otillar, R., Spatafora, J. W., Yadav, J. S., Aerts, A., Benoit, I., Boyd, A., Carlson, A., Copeland, A., Coutinho, P. M., de Vries, R. P., Ferreira, P., Findley, K., Foster, B., Gaskell, J., Glotzer, D., Gorecki, P., Heitman, J., Hesse, C., Hori, C., Igarashi, K., Jurgens, J. A., Kallen, N., Kersten, P., Kohler, A., Kues, U., Kumar, T. K., Kuo, A., LaButti, K., Larrondo, L. F., Lindquist, E., Ling, A., Lombard, V., Lucas, S., Lundell, T., Martin, R., McLaughlin, D. J., Morgenstern, I., Morin, E., Murat, C., Nagy, L. G., Nolan, M., Ohm, R. A., Patyshakuliyeva, A., Rokas, A., Ruiz-Duenas, F. J., Sabat, G., Salamov, A., Samejima, M., Schmutz, J., Slot, J. C., St John, F., Stenlid, J., Sun, H., Sun, S., Syed, K., Tsang, A., Wiebenga, A., Young, D., Pisabarro, A., Eastwood, D. C., Martin, F., Cullen, D., Grigoriev, I. V., & Hibbett, D. S. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science*, 336(6089): 1715-1719.
- Riley, R., Salamov, A. A., Brown, D. W., Nagy, L. G., Floudas, D., Held, B. W., Levasseur, A., Lombard, V., Morin, E., Otillar, R., Lindquist, E. A., Sun, H., LaButti, K. M., Schmutz, J., Jabbour, D., Luo, H., Baker, S. E., Pisabarro, A. G., Walton, J. D., Blanchette, R. A., Henrissat, B., Martin, F., Cullen, D., Hibbett, D. S., & Grigoriev, I. V. 2014. Extensive sampling of basidiomycete genomes demonstrates inadequacy of the white-rot/brown-rot paradigm for wood decay fungi. *Proc Natl Acad Sci U S A*, 111(27): 9923-9928.
- Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., & Henrissat, B. 2013. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol Biofuels*, 6(1): 41.
- Enright, A. J., Van Dongen, S., & Ouzounis, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7): 1575-1584.

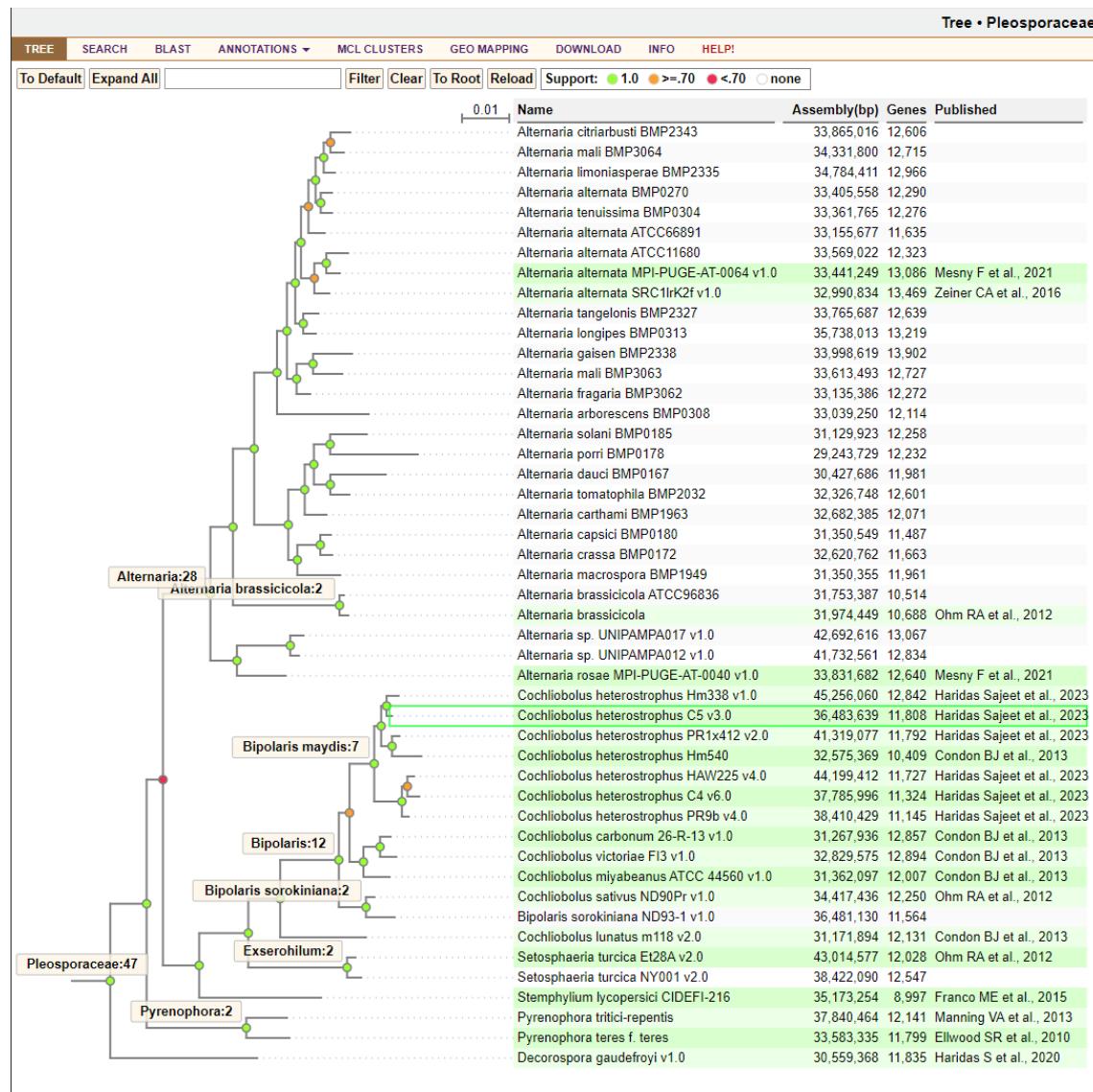
# MycoCosm: Synteny Tutorial

**Objective:** Explore genome synteny of *Cochliobolus heterostrophus* C5 with related genomes using the Pleosporaceae group page and the *Cochliobolus heterostrophus* C5 genome portal.

The SYNTENY tab is used for pairwise whole genome comparisons, enabling visual comparative analysis of complete genome assemblies at different levels of resolution. Since this uses one genome as the comparator, the SYNTENY tab is only available on single genome portals (i.e., absent from groups).

First, go to the Pleosporaceae group page at <https://mycocosm.jgi.doe.gov/Pleosporaceae>

Click on the TREE tab and locate *Cochliobolus heterostrophus* C5 in the tree.



Note the green selection box while mousing over the tree. Left-clicking will collapse and expand the selection box. Shift+clicking will isolate the selection in a new view. To restore the default view, click the TREE tab (the browser back button does not work on the tree page). Click on “*Cochliobolus heterostrophus* C5” to go to the organism genome portal. Ideally, you should do this in another tab or window so that you can follow the exercises below keeping the phylogenetic placement of this organism in mind.

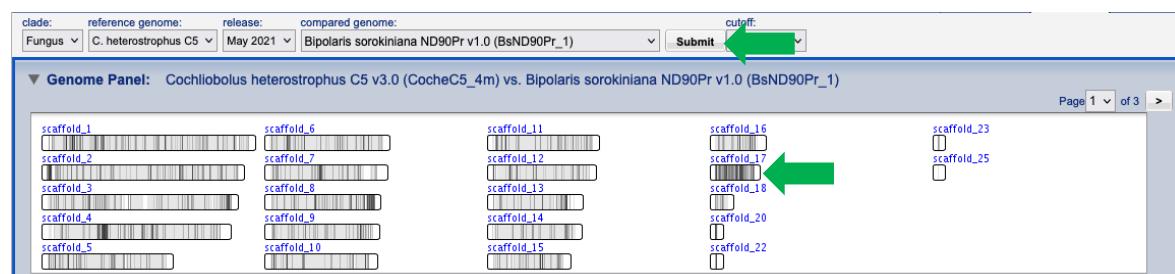
Click on the SYNTENY tab in the organism portal (*Cochliobolus heterostrophus* C5).

Genomic synteny is displayed in three collapsible panels in the Synteny Browser:

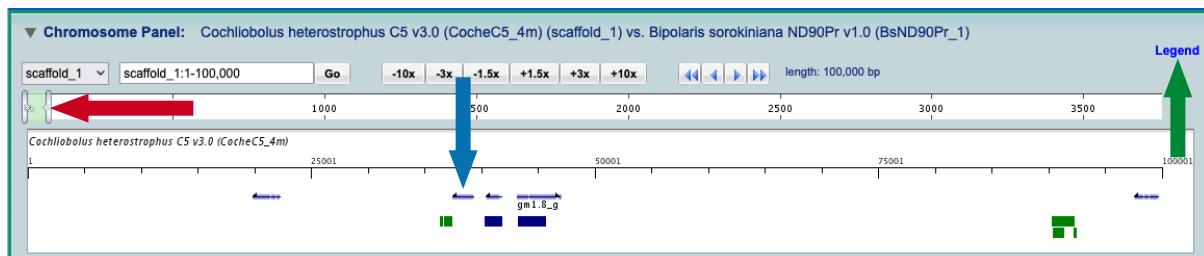
- A. the Genome Panel
- B. the Chromosome Panel
- C. the Comparison Panel.

The compared genome can be changed from the dropdown menu and clicking “Submit”.

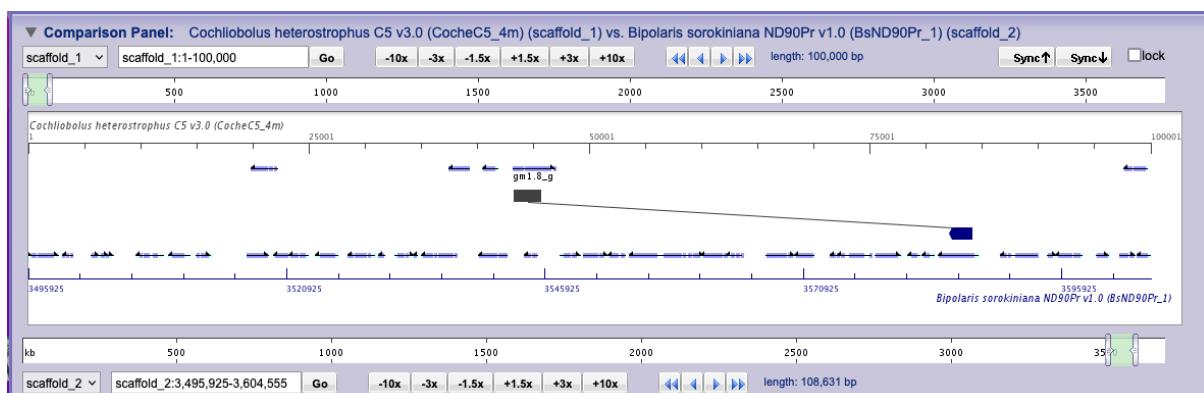
- A. The Genome Panel depicts alignment density for all scaffolds in the reference genome against all chromosomes in the compared genome. Here, alignment density is defined for a region in the reference genome as the number of syntenic regions in the compared genome. Darker regions in the image have higher density of coverage. Clicking on a particular scaffold selects that for the Chromosome and Comparison panels below.



- B. The Chromosome Panel shows all of the alignments in the compared genome to a particular interval on a single chromosome in the reference genome. Synteny is depicted as "blocks" along the reference-genome interval. Each block represents an alignment of two sequences, where the position of the block indicates the alignment's location on the reference genome and the color of the block indicates the chromosome where the match is found on the compared genome. Click on Legend (green arrow) to reveal the color-coding schema. The blocks appear stacked on top of each other when a fragment of the reference genome has synteny with multiple locations in the compared genome. The navigation buttons along with the chromosome slider (red arrow) allow for zooming and panning along the interval of the reference chromosome. A protein model (blue arrow) leads to the protein page, which shows annotations and a link to the genome browser.



C. The Comparison Panel zooms further to depict synteny between a specific interval on the reference genome and a specific interval on the compared genome. In this view, each aligned region is depicted as a pair of blocks, one along the reference chromosome (grey) and one along the compared chromosomes (colored), connected by a line. Also displayed in the Comparison Panel are gene model tracks (if available) for the reference and compared chromosomes.



Syntenic blocks and gene models are both interactive, as described above for the Chromosome Panel. Navigation controls allow the user to switch chromosomes, zoom and pan independently over the reference and compared genomes.

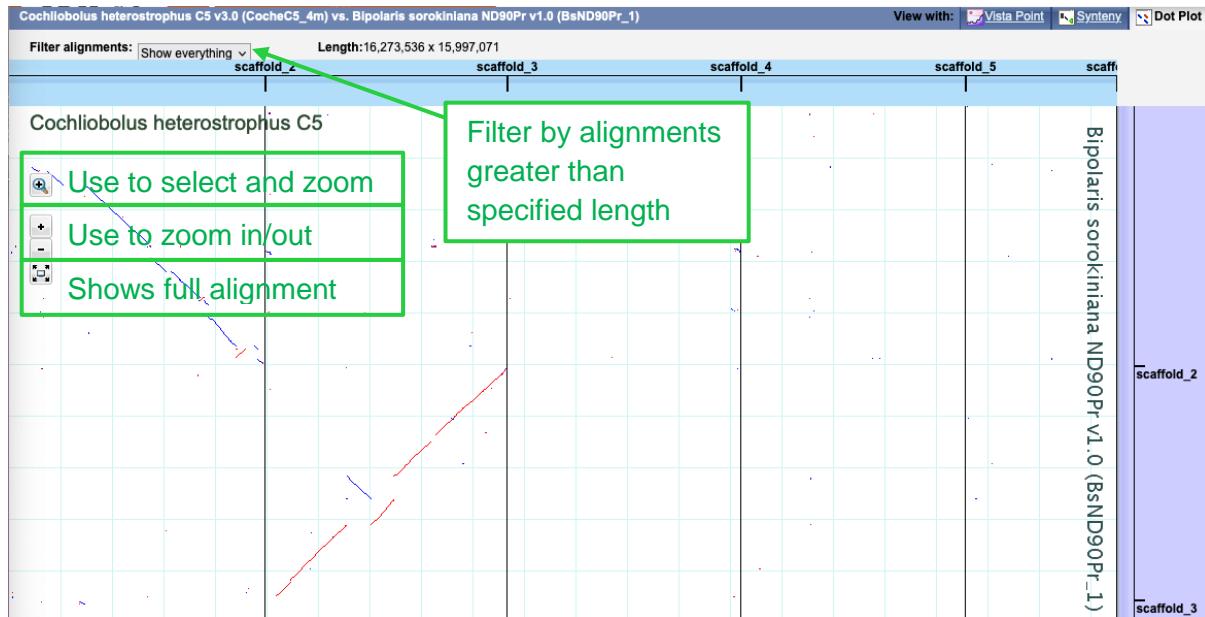
The SYNTENY page also allows whole genome pairwise comparison and comparison of one-to-many using the ‘Dot Plot’ and ‘Vista Point’ views respectively.

‘Dot Plot’ is an interactive tool that enables users to look at the DNA conservation between two genome assemblies at different levels of resolution and across multiple chromosomes/scaffolds.



In the main view window, DNA coordinates of the reference genome are presented on the X axis, and DNA coordinates of the compared genome are presented on the Y axis. All chromosomes or scaffolds are concatenated together, usually in a descending order by size. The diagonal lines in the image display the homologous regions between the two genomes. If the line is blue, the regions are on the same strand. If the line is red, the regions are on opposite strands. The grid in black lines indicates scaffold/chromosome boundaries. Use the

toolbar on the left to zoom or select specific regions on the plot. The map can also be navigated using click+drag similar to google maps. A cutoff control above the main window allows you to filter alignments to show only syntenyic regions greater than a specified length.



‘Dot Plot’ hides the genome portal navigation bar. You can click the “Synteny” view to restore it.

‘Vista Point’ shows multiple genome alignment using “peaks and valleys” graph as seen on the genome browser. Regions of high conservation are colored according to the annotation as exons (dark blue), UTRs (light blue) or non-coding (pink). The thresholds that determine what gets colored, as well as minimum and maximum percentage bounds can be adjusted by the user. The order of the curves and the zoom can be adjusted using drag-and-drop and click-and-drag respectively.



## Exercises:

1. Study the phylogenetic tree of the Pleosporaceae.
2. Use the SYNTENY tab in the *Cochliobolus heterostrophus* C5 genome portal and compare it to the genome of *Cochliobolus heterostrophus* C4. Compare the *Cochliobolus heterostrophus* C5 genome with other genomes of increasing phylogenetic distance like *Cochliobolus sativus*, *Setosphaeria turcica* and *Alternaria brassicicola*. Increase the viewed area by dragging the slider to cover a greater percentage of the scaffold. Note how increasing the cutoff from the default (50bp) can remove spurious alignments often caused by repeats.
3. Use the 'Dot Plot' view to study the high congruence between the two *Cochliobolus heterostrophus* assemblies. Compare the *Cochliobolus heterostrophus* C5 genome with other genomes of increasing phylogenetic distance as above. Note the breakdown of large scale synteny with increasing phylogenetic distance into mesosynteny as described by Ohm et al. (2012). In mesosynteny, genes are conserved within homologous chromosomes (scaffolds), but with randomized orders and orientations. Mesosynteny becomes more pronounced moving further phylogenetically to *Stagonospora nodorum* (Phaeosphaeriaceae). Ohm et al. showed that this type of genome evolution can be explained by repeated intra-chromosomal inversions.

## Reference:

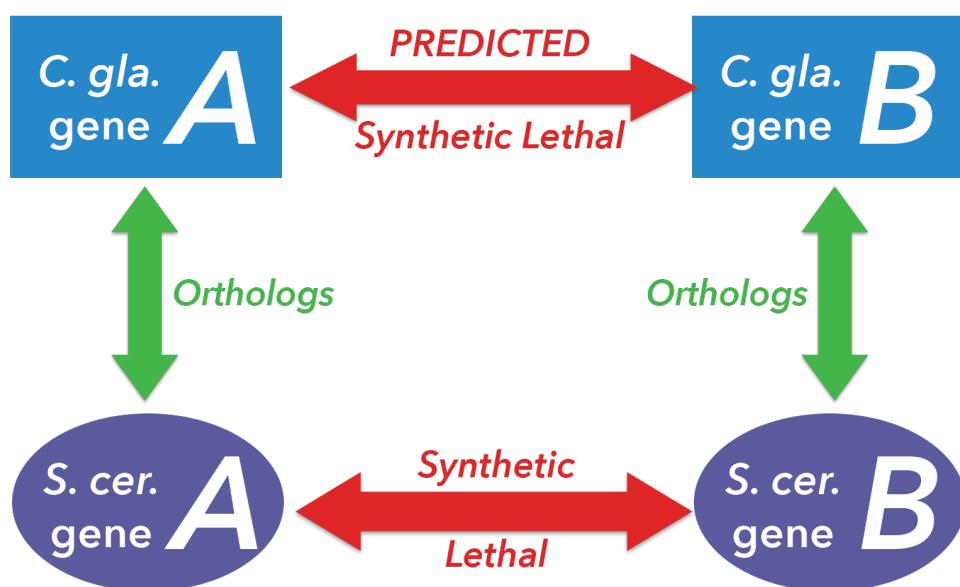
- Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, et al. (2012) Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi. PLOS Pathogens 8(12): e1003037.

## Using *S. cerevisiae* Orthologs to Predict Fungal Pathogen Biology

Antifungal agents such as azoles are used to treat infections with *Candida* species. Unfortunately, the opportunistic fungal pathogen *C. glabrata* possesses a relatively high intrinsic resistance to azoles, and also becomes resistant to azole treatment quickly.

Mitochondrial dysfunction and loss of the mitochondrial genome have been proposed as mechanisms by which *C. glabrata* acquires azole resistance. To exploit the loss of mitochondrial function in resistant *C. glabrata* isolates, researchers may be able to target proteins or pathways that become essential only when the mitochondrial genome is absent. This is based on the idea of synthetic lethality—a type of genetic interaction where the loss of two or more nonessential genes in combination results in cell inviability.

Genetic interactions such as synthetic lethality are richly documented for the budding yeast *S. cerevisiae*, but not as much for many other fungal species. By examining known genetic interactions in *S. cerevisiae*, we can predict synthetic lethal relationships in *C. glabrata* and other fungal pathogens.



If conserved, these synthetic lethal interactions may reveal future antifungal targets for use against azole-resistant strains in the clinic. Using known synthetic lethal interactions in the *S. cerevisiae* genome, predict potentially conserved synthetic lethal interactions for mitochondrial genes in *C. glabrata*.

## 1. Obtain a list of all genes encoded in the mitochondrial genome of *C. glabrata*:

- On the CGD homepage (<http://www.candidagenome.org>), open the Search tab in the yellow toolbar and select Advanced Search.

The screenshot shows the CGD homepage with a yellow header bar containing links for Home, Search, GBrowse, JBrowse, Sequence, GO, Tools, Literature, Download, and Community. The 'Search' tab is highlighted. In the sidebar under 'Search', there is a list of tools: BLAST, GO Term Finder, GO Slim Mapper, Text Search, Primers, PatMatch, and Advanced Search. A purple arrow points to the 'Advanced Search' link. Below the sidebar, there is a photograph of GFP-labeled Dam1 Complex proteins in DAPI-stained nuclei, with the caption: 'Courtesy of Laura Burack and Judy Berman, University of Minnesota'.

- In Step 1 of the Advanced Search, select ***Candida glabrata CBS138*** as your strain.
- In Step 2, check the “Select all chromosomal features” checkbox.
- In Step 3, specify that that you are looking for mitochondrial genes by selecting “**mito\_C\_glabrata\_CBS138**” as the chromosome.
- Click on “Search” (bottom left). A results page will follow, listing out 37 features in the *C. glabrata* mitochondrial genome.

The screenshot shows the 'Advanced Search' interface. It has three main sections: Step 1 (Select strain), Step 2 (Select chromosomal feature), and Step 3 (Narrow results). In Step 1, 'Candida glabrata CBS138' is selected. In Step 2, the 'Select all chromosomal features' checkbox is checked. In Step 3, the 'Chromosome' dropdown menu is open, showing options like 'mito\_C\_glabrata\_CBS138', 'ChrJ\_C\_glabrata\_CBS138', etc. A purple arrow points to each of these three specific areas.

- Scroll to the bottom of the page and click on the “Download All Search Results” link. The results will download in an Excel sheet.

CagIMt30	tRNA: Uncharacterized	tL(UAA)4mt	Mitochondrial leucine tRNA, has UAA anticodon	mito_C_glabrata_CBS138:17616 to 17697   GBrowse	Relative Coordinates	Chromosomal Coordinates
				Noncoding_exon 1 to 82	17,616 to 17,697	
Sort by : Systematic Name <input style="float: right;" type="button" value="Go!"/>						
<b>Analyze gene list: further analyze the gene list displayed above or download information for this list</b>						
Further Analysis:	GO Term Finder Find common features of genes in list	GO Slim Mapper Sort genes in list into broad categories	View GO Annotation Summary View all GO terms used to describe genes in list			
Download:	<a href="#">Download All Search Results</a>	<a href="#">Download</a>	Download selected information for entire gene list. Available information types include Sequence, Coordinates, GO annotations, Phenotype.			
Result Page : 1 2 Next						

## 2. Use FungiDB to find *S. cerevisiae* orthologs of *C. glabrata* mitochondrial genes:

- Open the FungiDB homepage (<http://fungidb.org/>). In the “Search for...” section on the left, select "Genes," then “Annotation, curation and identifiers” section and click on “List of IDs”.

**Search for...**

[expand all](#) | [collapse all](#)

?

▼ Genes

- ▼ Annotation, curation and identifiers
  - [List of IDs](#)
  - [User Comments](#)
- ▶ Epigenomics
- ▶ Function prediction
- ▶ Gene models
- ▶ Genetic variation
- ▶ Genomic Location
- ▶ Immunology
- ▶ Orthology and synteny

- Using your exported file from CGD, copy and paste the ORF names of the *C. glabrata* mitochondrial genes into the box. Click on “Get Answer”.
- In the Search Strategy panel, click on the “Add Step” button. In the resulting pop-up window, click on “Transform into related records” and then “Orthologs.”
- In the “Organism” list, search for “cerevisiae”. Select “Saccharomyces cerevisiae S288C”, click “select only these,” and then hit “Run Step”.
- 12 orthologs in *S. cerevisiae* will be returned. Download this list by clicking on the “Download” link on the top right side of the table.

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
Q0130	Q0130-l26_1	S. cerevisiae S288c	KP263414:46,723..46,953(+)	F0 ATP synthase subunit c	CagIMp10	OG5_126818	0	78
Q0045	Q0045-l26_1	S. cerevisiae S288c	KP263414:13,818..26,701(+)	cytochrome c oxidase subunit 1	CagIMp04, CagIMp07	OG5_128358	1	43
Q0070	Q0070-l26_1	S. cerevisiae S288c	KP263414:13,818..23,167(+)	intron-encoded DNA endonuclease al5 alpha	CagIMp04, CagIMp07	OG5_128358	1	43
Q0105	Q0105-l26_1	S. cerevisiae S288c	KP263414:36,540..43,647(+)	cytochrome b	CagIMp03	OG5_128504	1	31
Q0120	Q0120-l26_1	S. cerevisiae S288c	KP263414:36,540..42,251(+)	intron-encoded RNA maturase bl4	CagIMp03	OG5_128504	1	31

- In the download options menu, select “**Tab delimited (openable in Excel) – choose a pre-configured table**”. Set the Download Type as **Comma-delimited (.csv) file\***, then hit **Get**.

### 3. Import the *S. cerevisiae* orthologs into AllianceMine:

- Access AllianceMine from Alliance home page (<https://www.alliancegenome.org>); click on AllianceMine from the Data and Tools section at the top of the page.

The screenshot shows the Alliance home page with a navigation bar at the top. The 'Data and Tools' menu is open, displaying several options: Downloads, API, AllianceMine (which has an orange arrow pointing to it), JBrowse 2, Submit Data, Textpresso, and Tools and Prototypes. To the right of the menu, there is a search bar with the text 'All' and 'Search'.

- Open the file of *S. cerevisiae* orthologs that you downloaded from FungiDB. Sort this file by Column C to get just the exons, then copy the QXXXX exon names. To import these orthologs into AllianceMine, go to the **Upload** tab and then create a new list by choosing the organism as *S. cerevisiae* and then pasting in the list of orthologs.

## Create a new list

Select the type of list to create and then enter your identifiers or upload them from a file.

**i** • Separate identifiers by a comma, space, tab or new line  
• Qualify any identifiers that contain whitespace with double quotes like so: "even skipped"

List type  
Gene

Organism  
S. cerevisiae

Identifiers are case sensitive

Free Text  File Upload

Q0045  
Q0045  
Q0045  
Q0045  
Q0045

**SHOW EXAMPLE** **RESET** **CONTINUE**

- You can save this list of genes as “**List 1: S. cerevisiae orthologs**” (type it in). Click on the blue “**Save List**” button.

Upload /  Save

36 of your 12 identifiers matched a Gene

12 Matches 24 Synonyms

List Name **List 1: S. cerevisiae orthologs**   Save List

Matches (12)  Synonyms (24)

① An exact match was found for the following identifiers

PREVIOUS  NEXT  Show 10 results on page  Page 1 of 2

Your Identifier	Matches	Primary DBID	Systematic Name	Organism > Short Name	Standard Name	Name
Q0060	S000007263	Q0060	S. cerevisiae		AI3	
Q0070	S000007265	Q0070	S. cerevisiae		AI5_ALPHA	
Q0250	S000007281	Q0250	S. cerevisiae		COX2	Cytochrome c OXidase
Q0080	S000007267	Q0080	S. cerevisiae		ATP8	ATP synthase
Q0140	S000007275	Q0140	S. cerevisiae		VAR1	
Q0130	S000007274	Q0130	S. cerevisiae		OL11	OLigomycin resistance
Q0085	S000007268	Q0085	S. cerevisiae		ATP6	ATP synthase
Q0045	S000007260	Q0045	S. cerevisiae		COX1	Cytochrome c OXidase
Q0065	S000007264	Q0065	S. cerevisiae		AI4	
Q0120	S000007273	Q0120	S. cerevisiae		BI4	

4. After you save the list, you'll get query results with options for running searches. In the **Widgets** section below the table, click "view all" in the **Interactions** section. You'll get roughly 1000 results that you can see by clicking the **forward button** (opposite the back button at the top left)

## Widgets

### Interactions

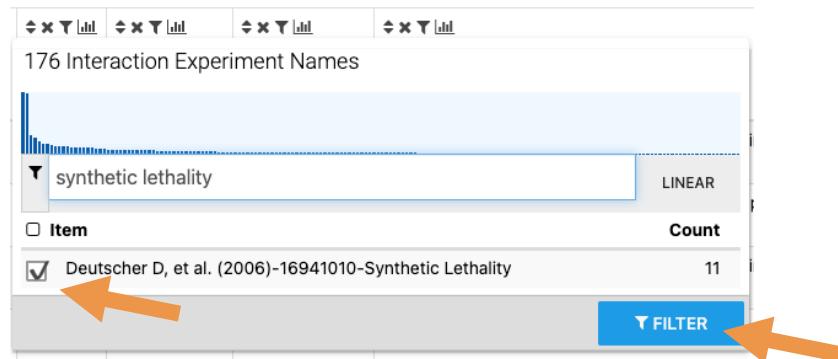
Genes (from the list or not) that interact with genes in this list. Counts may include the same interaction more than once if observed in multiple experiments.

All Genes in the table have been analysed in this widget.

[VIEW ALL](#)

<input type="checkbox"/> BioEntity.secondaryIdentifier	BioEntity.name
<input type="checkbox"/>	YGL187C Cytochrome c OXidase
<input type="checkbox"/>	YER154W cytochrome OXidase Activity
<input type="checkbox"/>	YIR024C INner membrane Assembly 22 kDa
<input type="checkbox"/>	YKR016W Mitochondrial contact site and Cristae organizing system
<input type="checkbox"/>	YLR203C Mitochondrial Splicing Suppressor
<input type="checkbox"/>	YOL027C Mitochondrial Distribution and

- In the column for "Experiment Name," click the **bar graph icon** and type in "synthetic lethality." Click the box next to any results, and then click **Filter**.



- You'll get a list of eleven genes that are synthetically lethal with a member of your original ortholog list. (COX1)
- Click "Save List" at the top right and then choose "**Pick items from the table**." Radio buttons will appear in the table and you want to check all the items in the "**Participant 2 Standard Name**" column (this will automatically select the DBID as well). These are the genes known to be synthetically lethal with the list you used as input.

SAVE LIST </> PYTHON EXPORT

- Genes (1)**
- Gene > Organisms (1)**
- Gene > Interactions > Interaction Details (11)**
- Gene > Interactions > Bio-Entities (9)**
- Gene > Interactions > Interaction Details > Int**

Pick items from the table

Participant 2 Primary DBID	Participant 2 Standard Name	Experiment Name
<input checked="" type="checkbox"/> S000000773	<input checked="" type="checkbox"/> FRD1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000001624	<input checked="" type="checkbox"/> SDH3	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000001699	<input checked="" type="checkbox"/> URA1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000001699	<input checked="" type="checkbox"/> URA1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000002708	<input checked="" type="checkbox"/> PRO1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000004132	<input checked="" type="checkbox"/> PUT1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000004132	<input checked="" type="checkbox"/> PUT1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000004340	<input checked="" type="checkbox"/> DIC1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000005850	<input checked="" type="checkbox"/> PRO2	Deutscher D, et al. (2006)-16941010-Synthetic Lethality
<input checked="" type="checkbox"/> S000006183	<input checked="" type="checkbox"/> FUM1	Deutscher D, et al. (2006)-16941010-Synthetic Lethality

- Save this list as "List 2 synthetic lethals with orthologs"

Save a list of 11 InteractionDetails

Name

Optional attributes

Description

CANCEL

- Access your new gene list by clicking on the **Lists** link in the top toolbar and selecting your new list name.
- Export the list of synthetic lethal interactors by clicking on the **Export** button, and then on the **Download file** button.

Export this table as...

File name and type  TSV

Preview (first 3 rows)

ORF > Primary DBID	ORF > Systematic Name	ORF > Organism . Sh
S000000773	YEL047C S. cerevisiae	FRD1 Fumarate Reductase
S000001624	YKL141W S. cerevisiae	SDH3 Succinate DeHydroge
S000001699	YKL216W S. cerevisiae	URA1 URAcil requiring

Column headers

No column headers  
 Use human readable headers (e.g. Gene > Organism Name)  
 Use raw path headers (e.g. Gene.organism.name)

Select rows

Size: 8 (all rows)

Offset: 0

Select columns

ORF > Primary DBID  
 ORF > Systematic Name

## 5. Import the *S. cerevisiae* synthetic lethal interaction genes into FungiDB for further analysis:

- Open the FungiDB homepage (<http://fungidb.org/>). In the “Search for...” section on the left, select “Genes,” then “Annotation, curation and identifiers” section and click on “List of IDs”.
- Using your exported file from AllianceMine, copy and paste the ORF names of the *S. cerevisiae* interactors (e.g. YEL047C, YKL141W, etc.) into the ID box. Click on “Get Answer”.
- In the Search Strategy panel, click on the “Add Step” button. In the resulting pop-up window, click on “Transform into related records” and then “Orthologs.”
- In the “Organism” list, search for “glabratus”. Select “**Nakaseomyces glabratus CBS 138 [Reference]**,” (make sure to unclick *S. cerevisiae*), then hit “Run Step”.
- 9 orthologs of the *S. cerevisiae* interactors will be returned. These are *C. glabrata* genes are predicted to have synthetic lethal interactions with *C. glabrata* mitochondrial genes. You can download this list.
- Then, above the Gene Results table, click on the **Analyze Results** button. Select **Gene Ontology Enrichment** and run an enrichment for **Biological Process**. Are the results surprising? Remember that these *S. cerevisiae* genes have synthetic lethal interactions with mitochondrial genes. Do the results suggest any biological processes that, if disrupted, might possibly inhibit mitochondria-defective *C. glabrata* clinical isolates?

## Exercise: Ensembl Fungi Gene Trees and Homologues

Links to be clicked shown in blue, text to be entered shown in red.

Let's look at the homologues of *Saccharomyces cerevisiae* (R64-1-1) **TAZ1** (gene stable ID: YPR140W). This gene is involved in stress response and conserved across different taxonomic domains. Click on the gene ID **YPR140W** to open the 'Gene' tab.

The screenshot shows the Ensembl Fungi website interface for the *Saccharomyces cerevisiae* (R64-1-1) genome. The URL is [https://www.ensembl.org/Fungi/Saccharomyces\\_cerevisiae/R64-1-1/Gene?g=TAZ1](https://www.ensembl.org/Fungi/Saccharomyces_cerevisiae/R64-1-1/Gene?g=TAZ1). The main content area displays the gene details for TAZ1, including its description as a Lysophosphatidylcholine acyltransferase required for normal phospholipid content of mitochondrial membranes, and its chromosomal location on Chromosome XVI: 814,391-815,536 forward strand. A callout box highlights the 'Gene tree' link under the 'Fungal Compara' section of the left-hand menu. The 'Gene' tab is currently selected.

Click on **Fungal Compara: Gene tree** on the left-hand menu, which will display the current gene (in the context of a phylogenetic tree) used to determine orthologues and paralogues

# Evolutionary Analysis (Gene Trees): Ensembl Fungi

**EnsemblFungi** • HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

Login/Register

**Saccharomyces cerevisiae (R64-1-1) ▾**

Location: XVI:814,391-815,536 Gene: TAZ1 Transcript: TAZ1

**Gene-based displays**

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence**
  - Secondary Structure
  - Gene families
  - Literature
- Fungal Comparisons
  - Genomic alignments
  - Gene tree**
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Pan-taxonomic Comparisons
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Biological process
  - GO: Cellular component
  - GO: Molecular function
  - PHI: Phibase identifier
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Molecular interactions
- Regulation
- External references
- Supporting evidence
- ID History
- Gene history

**Configure this page**

**Custom tracks**

**Export data**

**Share this page**

**Bookmark this page**

**Gene: TAZ1 YPR140W**

**Description**

Lyso-phosphatidylcholine acyltransferase; required for normal phospholipid content of mitochondrial membranes; major determinant of the final acyl chain composition of the mitochondrial-specific phospholipid cardiolipin; mutations in human ortholog tafazzin (TAZ) cause Barth syndrome, a rare X-linked disease characterized by skeletal and cardiomyopathy and bouts of cyclic neutropenia; a specific splice variant of human TAZ can complement yeast null mutant [Source:SGD;Acc:S000006344]

**Location**

Chromosome XVI: 814,391-815,536 forward strand. R64-1-1:BK006949.2

**About this gene**

This gene has 1 transcript (splice variant) and 326 orthologues.

**Transcripts**

Show transcript table

**Gene tree ?**

**Unique gene tree stable ID**

**GeneTree** EFGT01050000064920

**Summary statistics**

Number of genes	327
Number of speciation nodes	295
Number of duplication nodes	13
Number of ambiguous nodes	18
Number of gene split events	0

Highlight annotations

Show  GO  InterPro

Show 10 entries

**Filter tree by Gene Ontology (GO) terms or InterPro protein domains**

highlight	Accession	Description
<input type="radio"/> 327 members	GO:0003674	molecular_function
<input type="radio"/> 327 members	GO:0003824	catalytic activity
<input type="radio"/> 327 members	GO:0006629	lipid metabolic process
<input type="radio"/> 327 members	GO:0006644	phospholipid metabolic process
<input type="radio"/> 327 members	GO:0006793	phosphorus metabolic process
<input type="radio"/> 327 members	GO:0006796	phosphate-containing compound metabolic process
<input type="radio"/> 327 members	GO:0008150	biological_process
<input type="radio"/> 327 members	GO:0008152	metabolic process
<input type="radio"/> 327 members	GO:0009987	cellular process
<input type="radio"/> 327 members	GO:0016740	transferase activity

Showing 1 to 10 of 119 entries

**Protein alignments**

**Collapsed nodes**

**Gene and species of interest**

**Basidiomycete fungi:** 94 homologs

**Budding yeasts:** 9 homologs

**CTG clade:** 14 homologs

**Saccharomycetaceae:** 8 homologs

**TAZ1, *Saccharomyces cerevisiae*:** 1 homolog

**CAGL0D04972g, *Candida glabrata*:** 1 homolog

**KAFR\_0B06580, *Kazachstania africana CBS 235*:** 1 homolog

**Naumovozyma:** 2 homologs

**Saccharomycetaceae:** 5 homologs

**Wickerhamomyces:** 2 homologs

**Budding yeasts:** 2 homologs

**HGUI\_02154, *Hanseniaspora guilliermondii*:** 2 homologs

**Cyberlindnera:** 2 homologs

**NADFDRAFT\_4515, *Nadsonia fulvescens* var. *elongata* DSM 6958:** 2 homologs

**Trichomycetaceae:** 2 homologs

**LIPSTDRAFT\_68925, *Lipomyces starkeyi* NRRL Y-11557:** 1 homolog

**CANCADRAFT\_19190, *Tortospora caseinolytica* NRRL Y-17796:** 1 homolog

**YALID\_C14036g, *Yarrowia lipolytica*:** 1 homolog

**Filamentous ascomycetes:** 154 homologs

**Taphrinomycotina:** 4 homologs

**Fungi incertae sedis:** 21 homologs

**LEGEND**

**Branch Length**

- x1 branch length
- x10 branch length
- x100 branch length

**Genes**

- Gene ID** gene of interest
- Gene ID** within-sp. paralog

**Nodes**

- gene node
- speciation node
- duplication node
- ambiguous node
- gene split event

**Collapsed Nodes**

- collapsed sub-tree
- collapsed (paralog)
- collapsed (gene of interest)

**Collapsed Alignments**

- 0 - 33% aligned AA
- 33 - 66% aligned AA
- 66 - 100% aligned AA

**Expanded Alignments**

- gap
- aligned AA

**Legend**

(a) How many duplication events are there in this tree?

Scroll to ‘View options’ at the bottom of the page. Here, you can find some quick filtering options. You can view paralogues and quickly expand or collapse nodes based on class, phylum etc.

**View options:**

- View current gene only (Default)
- View paralogues of current gene
- View all duplication nodes
- View fully expanded tree
- Collapse all the nodes at the taxonomic rank

Use the 'configure page' link in the left panel to see more options available from menus on individual tree nodes.

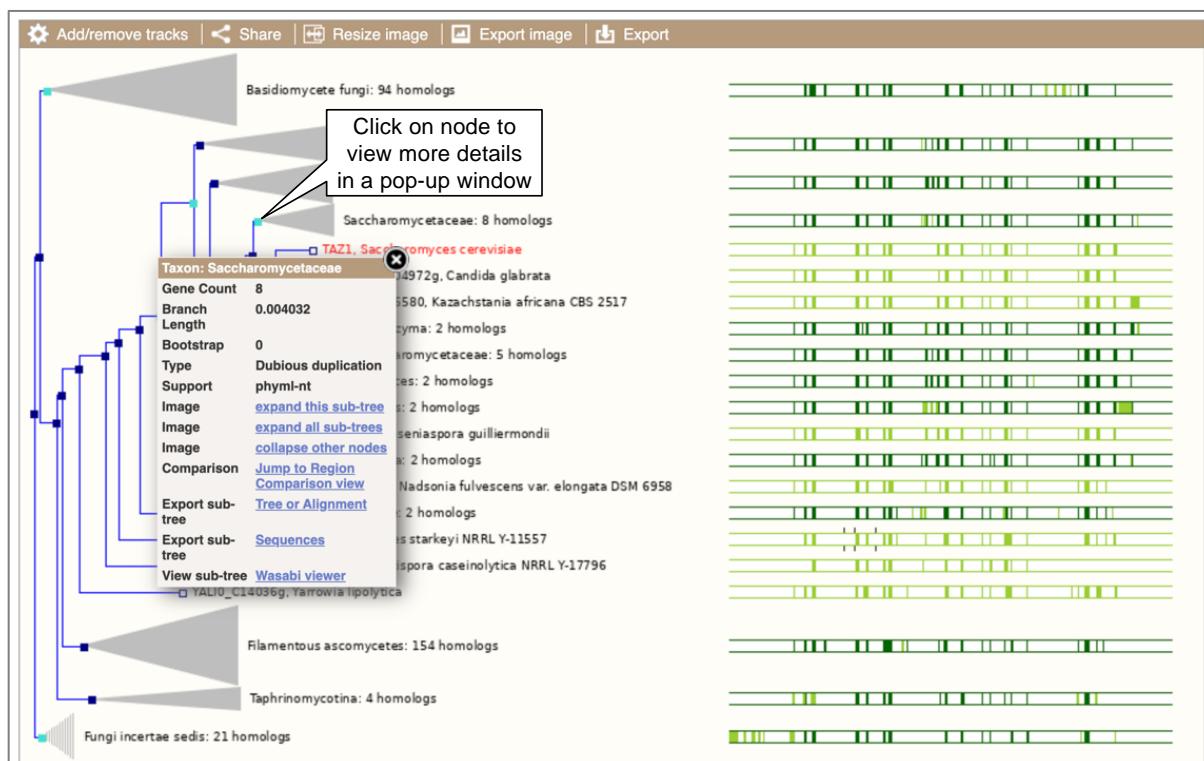
Ensembl Fungi release 58 - January 2024 © EMBL-EBI

About Us      Get help      Our sister sites      Follow us

About us      Using this website      Ensembl      Blog

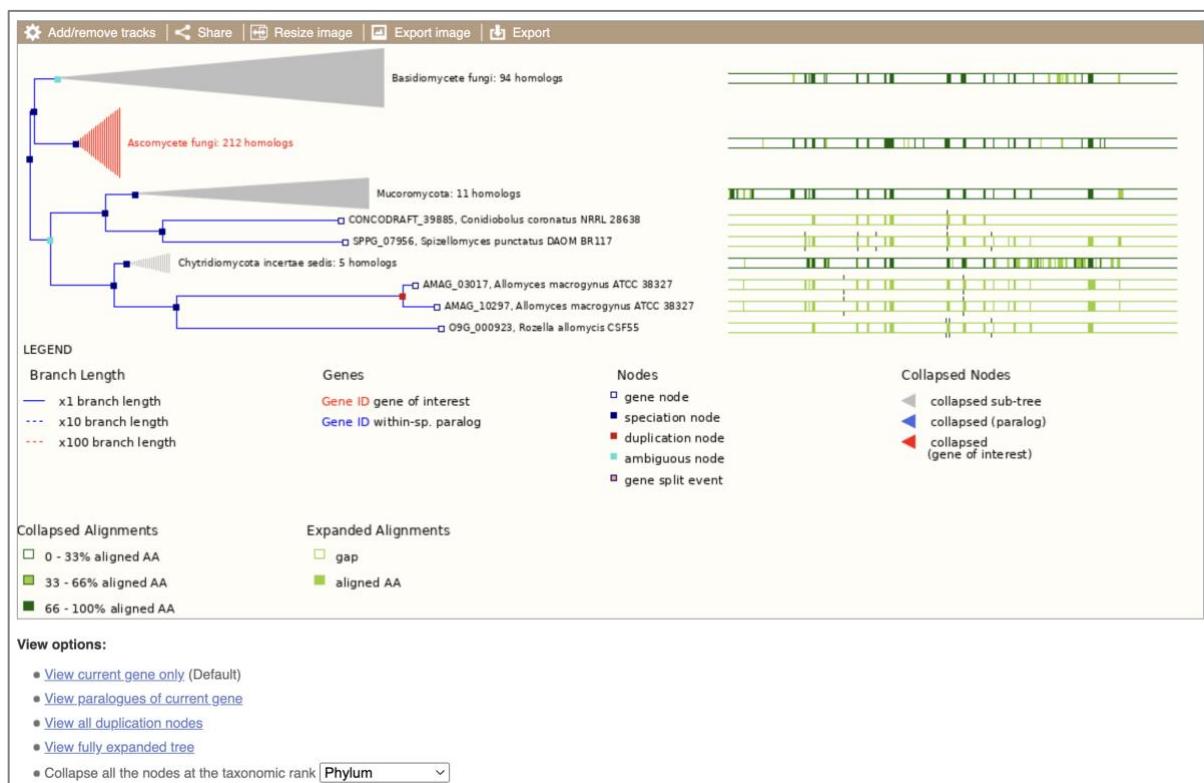
Click on **View all duplication nodes**. This will expand the tree so that all duplication nodes are visible. Count the number of red nodes. There are 13 duplication events in the tree.

Funnels indicate collapsed nodes. Click on a node (coloured square) to open a pop-up window, which tells you what type of node it is, some statistics and options to expand or export the sub-tree:



- (b) What is the Phylum with the highest number of *TAZ1* homologues?

Under ‘View options’, collapse all nodes at the taxonomic rank **Phylum**.



The phylum with the highest number of *TAZ1* homologues is Ascomycete fungi.

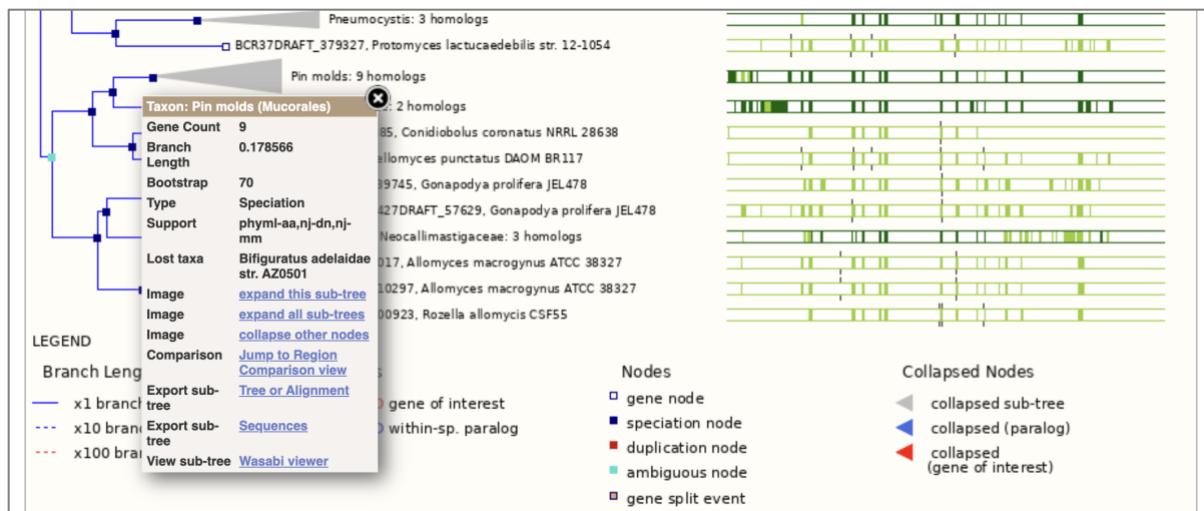
- (c) What is the bootstrap support of the pin moulds (*Mucorales*) class in this tree?

Bootstrap values in a phylogenetic tree indicate that out of 100, how many times the same branch is observed when repeating the generation of a phylogenetic tree on a resampled set of data. Bootstrap values in Ensembl gene trees are calculated using a tool called TreeBeST, and the final consensus trees consist of clades chosen to minimise the number of duplications, losses inferred and have the highest bootstrap support. More on this process is available at

[https://www.ensembl.org/info/genome/compara/homology\\_method.html](https://www.ensembl.org/info/genome/compara/homology_method.html).

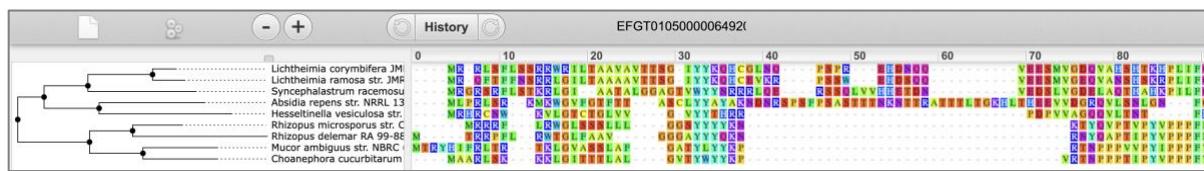
Click on the **Pin molds** node to view more details. In the pop-up window, you will find the bootstrap value to be 70.

## Evolutionary Analysis (Gene Trees): Ensembl Fungi

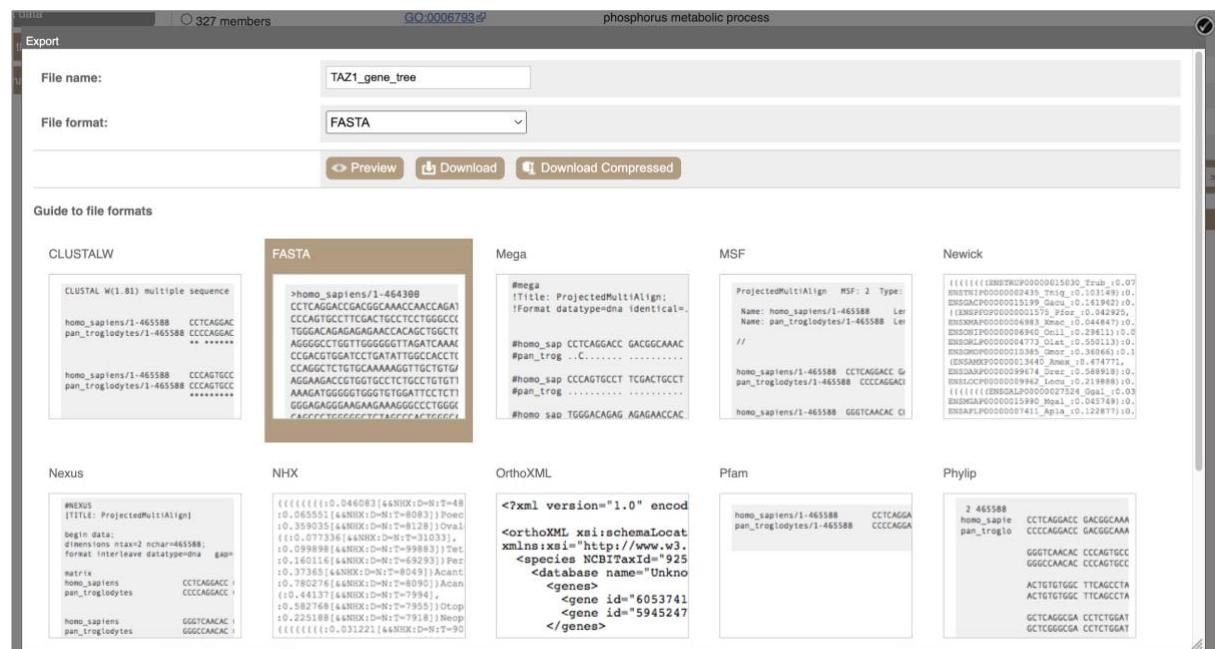


- (d) Can you display the sequence alignment of all the homologues in this Class (*Hint: Use the Wasabi viewer*)?

[Wasabi](#) is an open-source, web-based environment for visualising sequence data alongside phylogenetic trees. You can read more about the platform in this publication: <https://europemc.org/article/MED/26635364>.



You can download the tree in a variety of formats. Click on the [Export](#) icon  in the bar at the top of the image. This opens a pop-up window where you can choose your format. You can preview this file before you download it.



## Evolutionary Analysis (Gene Trees): Ensembl Fungi

We can look at homologues in the [Orthologues](#) and [Paralogues](#) pages, which can be accessed from the left-hand menu. If there are no orthologues or paralogues, then the link(s) will be greyed out. Click on [Orthologues](#) to see the orthologues available.

**Orthologues ?**

[Download orthologues](#)

**Summary of orthologues of this gene Hide ⊖**

Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'C' to show details for all species.

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (1504 species)	<input checked="" type="checkbox"/>	298	14	0	<a href="#">1192</a>
Acidomyces (2 species)	<input type="checkbox"/>	1	0	0	<a href="#">1</a>
Agaricales (36 species)	<input type="checkbox"/>	18	1	0	<a href="#">17</a>
Atheliales (2 species)	<input type="checkbox"/>	1	1	0	0
Blastocladiales (1 species)	<input type="checkbox"/>	0	1	0	0
Boletales (12 species)	<input type="checkbox"/>	6	0	0	<a href="#">6</a>
Botryosphaerales (7 species)	<input type="checkbox"/>	2	0	0	<a href="#">5</a>
Cantharellales (10 species)	<input type="checkbox"/>	1	1	0	<a href="#">8</a>
Capnodiales (35 species)	<input type="checkbox"/>	3	0	0	<a href="#">24</a>
Chaetothyriomycetidae (31 species)	<input type="checkbox"/>	0	0	0	<a href="#">24</a>
Chytridiomycota (14 species)	<input type="checkbox"/>	4	1	0	<a href="#">9</a>
Corticales (1 species)	<input type="checkbox"/>	1	0	0	0

Hover over the column names with your mouse to view a description

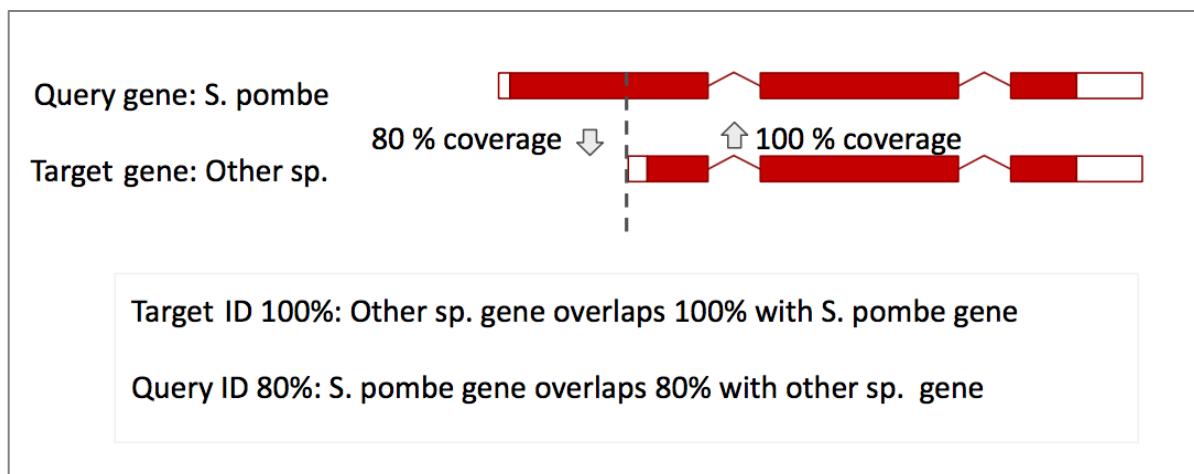
**Selected orthologues Hide ⊖**

Show All entries [Show/hide columns](#)

Species	Type	Orthologue	Link to orthologue gene tab	Filter table	Download table
Absidia repens str. NRRL 1336	1-to-1	<a href="#">BCR42DRAFT_405738</a>	<a href="#">View Gene</a> <a href="#">Compare Regions</a> (BCR42scaffold_281454)	Target %id: 23.74 %   Query %id: 17.32 %   GOC Score: n/a   WGA Coverage: n/a   High Confidence: No	<a href="#">Filter</a>
Acaromyces str. MCA 4198	1-to-1	<a href="#">AFT_281454</a>	<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (KZ819639:712,383-713,796:1) <a href="#">View Sequence Alignments</a>	Target %id: 26.17 %   Query %id: 24.93 %   GOC Score: n/a   WGA Coverage: n/a   High Confidence: No	
Acidomyces richmondensis BFW	1-to-1	<a href="#">M433DRAFT_132335</a>	<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (scaffold_55:22,999-24,384:1) <a href="#">View Sequence Alignments</a>	Target %id: 27.41 %   Query %id: 28.35 %   GOC Score: n/a   WGA Coverage: n/a   High Confidence: Yes	
Acremonium chrysogenum ATCC 11550	1-to-1	<a href="#">ACRE_050350</a>	<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (scaffold53:63,537-64,720:-1) <a href="#">View Sequence Alignments</a>	Target %id: 32.27 %   Query %id: 29.13 %   GOC Score: n/a   WGA Coverage: n/a   High Confidence: Yes	
Agaricus bisporus var. burnettii JB137-S8	1-to-1	<a href="#">AGABI1DRAFT_91626</a>	<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (JH971389:1,858,411-1,859,589:-1) <a href="#">View Sequence Alignments</a>	Target %id: 62 %   Query %id: n/a   GOC Score: n/a   WGA Coverage: n/a   High Confidence: No	

- (e) What is the difference between Target %id and Query %id? (*Hint: Mouse over*)

The sequence identity is reported in two ways. Target %id is how much of the orthologue (target gene) overlaps with the query gene (our *S. cerevisiae* gene). The Query %id is the inverse of this. For example:



Click on [Hide](#)  above the table or scroll to the bottom of the page to see a list of the species that do not have any orthologues with *TAZ1* in *S. cerevisiae*... there are a lot!

Species without orthologues
1190 species are not shown in the table above because they don't have any orthologue with YPR140W.
<ul style="list-style-type: none"> <li>● Ancestral sequence</li> <li>● [Candida] arabinofermentans NRRL YB-2248</li> <li>● [Candida] auris str. 6684</li> <li>● [Candida] auris</li> <li>● [Candida] glabrata</li> <li>● [Candida] glabrata</li> <li>● [Candida] glabrata</li> <li>● [Candida] glabrata</li> </ul>

*S. cerevisiae* is part of Ensembl's pan-taxonomic-compara (often shortened to pan-compara), which compares a subset of fungal species with representative species from other taxa, such as plants, protists, bacteria and vertebrates. This offers a broad view of homologous relationships from across the taxonomy. Go to [Pan-taxonomic Compara: Gene Tree](#). Let's look at the pan-taxonomic tree with nodes collapsed at the Kingdom rank.

## Evolutionary Analysis (Gene Trees): Ensembl Fungi



Click on Pan-taxonomic Compara: Orthologues.

Orthologues					
<a href="#"> Download orthologues</a>					
<b>Summary of orthologues of this gene</b> <a href="#">Hide</a>					
<small>Click on 'Show details' to display the orthologues for one or more groups of species. Alternatively, click on 'Configure this page' to choose a custom list of species.</small>					
Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (1504 species)	<input checked="" type="checkbox"/>	4	0	0	<a href="#">1500</a>
Acidomyces (2 species)	<input type="checkbox"/>	0	0	0	<a href="#">2</a>
Agaricales (36 species)	<input type="checkbox"/>	0	0	0	<a href="#">36</a>
Atheliales (2 species)	<input type="checkbox"/>	0	0	0	<a href="#">2</a>
Blastocladiales (1 species)	<input type="checkbox"/>	0	0	0	<a href="#">1</a>
Boletales (12 species)	<input type="checkbox"/>	0	0	0	<a href="#">12</a>
Botryosphaerales (7 species)	<input type="checkbox"/>	0	0	0	<a href="#">7</a>
Cantharellales (10 species)	<input type="checkbox"/>	0	0	0	<a href="#">10</a>

## Evolutionary Analysis (Gene Trees): Ensembl Fungi

Show All entries		Show/hide columns		Filter				
Species	Type	Orthologue		Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Aedes aegypti (Yellow fever mosquito, LVP_AGWG)	1-to-1 <a href="#">View Gene Tree</a>	<a href="#">AAEL001564</a> 2:21,496,991-21,541,309:-1 <a href="#">View Sequence Alignments</a>		23.81 %	18.37 %	n/a	n/a	No
Amborella trichopoda	1-to-1 <a href="#">View Gene Tree</a>	<a href="#">AMTR_s00022p00068080</a> AmTr_v1.0_scaffold00022:710,032-717,504:-1 <a href="#">View Sequence Alignments</a>		23.43 %	17.59 %	n/a	n/a	No
Amphimedon queenslandica (Demosponge)	1-to-1 <a href="#">View Gene Tree</a>	<a href="#">LOC100632622</a> GL345242.1:108,662-110,163:-1 <a href="#">View Sequence Alignments</a>		24.73 %	18.11 %	n/a	n/a	No
Anopheles gambiae (African malaria mosquito, PEST)	1-to-1 <a href="#">View Gene Tree</a>	<a href="#">AGAP007599</a> 2L:48,133,715-48,137,634:-1 <a href="#">View Sequence Alignments</a>		24.57 %	18.64 %	n/a	n/a	No

- (f) How many species with predicted orthologues for this gene are there in Fungal Compara? What about in Pan-compara?

Fungal Compara:

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (370 species)	<input checked="" type="checkbox"/>	298	14	0	58

Pan-taxonomic Compara:

Species set	Show details	With 1:1 orthologues	With 1:many orthologues	With many:many orthologues	Without orthologues
All (370 species)	<input checked="" type="checkbox"/>	4	0	0	366

- (g) How many animal orthologues are there? Does this number agree with the Pan-taxonomic tree above? Hint: Click the Show details box for Vertebrates and Metazoa, and count the number of orthologues in the table below).
- (h) Filter the second table to view the human orthologue. How much sequence identity does the human protein have to the *S. cerevisiae* one? Is it a high-confidence homologue? Click on the View Sequence Alignment link in the ‘Orthologue’ column to View Protein Alignment in ClustalW format. Does it support your conclusions?

## Evolutionary Analysis (Gene Trees): Ensembl Fungi

## Orthologue Alignment

 Download homology

### Type: 1-to-1 orthologues

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Saccharomyces cerevisiae	YPR140W	YPR140W	381 aa	18 %	65 %	XVI:814391-815536
Human	ENSG00000102125 ↗	ENSP0000469981 ↗	292 aa	24 %	85 %	X:154411524-154421726 ↗

CLUSTAL W (1.81) multiple sequence alignment

## Additional Exercise 1: *Zymoseptoria* Orthologues

We will now explore an orthologue that we identified using BioMart (additional exercise 1 in the BioMart module). We identified 18 genes associated with the GO term detoxification in *Zymoseptoria tritici*. We then found a single high-confidence orthologue in *Cryptococcus neoformans* (CNM01690) which we will now explore further.

Search for CNM01690 in *Cryptococcus neoformans* var. *neoformans* JEC21 and go to the gene page.

- Does this gene in *C. neoformans* have a UniProtKB-Gene Ontology annotation?
- Find the *Z. tritici* orthologue in the [Orthologues](#) page and view a protein alignment.
- At which end of the protein (N- or C-terminus) does the alignment between these two genes become worse?

## Additional Exercise 1 Answer: *Zymoseptoria* Orthologues

Go to [fungi.ensembl.org](http://fungi.ensembl.org) in your browser. In the species-specific search box, select *Cryptococcus neoformans* var. *neoformans* JEC21 from the drop-down list and enter **CNM01690**. In the results page, click on the gene stable ID **CNM01690** to navigate to the ‘Gene’ tab.

- In the left-hand panel under ‘Gene-based displays’, click on [External references](#). Yes, this gene has a UniProtKB-Gene Ontology annotation. The database ID for the UniProtKB-Gene Ontology annotation is Q5K7P9.

External database	Database identifier
NCBI gene (formerly Entrezgene)	<a href="#">3255046</a> [view all locations]
UniGene	<a href="#">Fne.7413</a> [view all locations]
UniProtKB-Gene Ontology Annotation	<a href="#">Q5K7P9</a> [view all locations]

- Go to [Fungal Compara: Orthologues](#) in the left-hand panel. Click on the [Hide ⊖](#) button for the ‘Summary of orthologues of this gene’ table. In the ‘Selected orthologues’ table, use the search bar in the top right-hand corner to search for *Zymoseptoria tritici*. Click on [View Sequence Alignments](#) and in the pop-up menu select [View Protein Alignment](#).

**Orthologues ?**

[Download orthologues](#)

**Summary of orthologues of this gene Show +**

**Selected orthologues Hide ⊗**

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Zymoseptoria tritici	1-to-1	<a href="#">Mycgr3G54449</a>	71.17 %	71.54 %	n/a	n/a	Yes
		<a href="#">View Gene Tree</a>	1:5,638,024-5,639,654:1		<a href="#">Orthologue Alignment</a>	<a href="#">View Protein Alignment</a>	
					<a href="#">View Sequence Alignments</a>	<a href="#">View cDNA Alignments</a>	

- (c) You can find a description of the different symbols by clicking the question mark icon  next to ‘Orthologue alignment’. This opens the corresponding help page in a new tab.

**Orthologue alignment ?**

[Download homology](#)

Type: 1-to-1 orthologues

Click on  to open the corresponding help page in a new browser tab

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Cryptococcus neoformans var. neoformans JEC21	CN01690	AAW46801	383 aa	71 %	98 %	13:510531-512507
Zymoseptoria tritici	<a href="#">Mycgr3G54449</a>	<a href="#">Mycgr3P54449</a>	385 aa	71 %	98 %	<a href="#">1:5638024-5639654</a>

CLUSTAL W (1.81) multiple sequence alignment

```

AAW46801/1-383      MSTEQVITCKAAIAWEAGKPLSIEVTVEVAPPKDGEVRIKILYTGLCHTDAYTLSGNPDE
Mycgr3P54449/1-385  MSTEQQTITCKAAVAWEAGKDLVIEDVEVLPPRAHEVRIKVAYTGVCHTDAYMLSGKDPE
*****.*****.***** * * * * *: *****: * *:*****: * *;**

```

In the help page, we can find a description of the conservation codes:

For protein alignments, the conservation codes are:

- \* when amino acids are identical
- : when amino acids are different but the function is conserved
- . when amino acids are different but the function is semi-conserved.
- space when amino acids are different and there is no conservation of function.

Dashes in the sequence (for both nucleotides and amino acids) indicate gaps in the alignment.

Looking at the ClustalW alignment and referring to the conservation codes, we can see that the N-terminus is more highly conserved than the C-terminus as there is a gap in the alignment in the C-terminus:

## Evolutionary Analysis (Gene Trees): Ensembl Fungi

CLUSTAL W (1.81) multiple sequence alignment

AAW46801/1-383 Mycgr3P54449/1-385	MSTEGQVITCKAAIAWEAGKPLSIEΤVEVAPPKDGEVRIKILYTLGLCHTDAYTLSGNPDE *****,*;*****:***** * * * * *; * * *; * * *; * * *; ***
AAW46801/1-383 Mycgr3P54449/1-385	GAFPVILGHEGGGIVESVGEGVDNVKVGDHVVPLYTAECRECKFCKSGKTNLGRVRTTQ GAFPVIAGHEGAGIIVESIGEVTVNVKGDTVALYTPECKECKFCKSGKTNLGKIRATQ ***** * * *;*****:***** * * *;*****:*****:*****:*****:*****:*****
AAW46801/1-383 Mycgr3P54449/1-385	GKGVMPDGTTRFKCKGQDILHFMGCSTFAQYTVVSKFSVVAINPKAPLKTSCLLGCGITT GKGVMPDGSSRFRCKGKDLLHFMGCSTFSQYTVVADISVVAVTDKAPMDRTCLLGCGITT *****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****
AAW46801/1-383 Mycgr3P54449/1-385	GYGAATKSP---GI-EGSNVAIFGVGCVGLSVLQGAKAKGCKRIFAIDTNPKKEWAVKF GYGAATITAGKNGVEKGDNVAVFGAGCVGLSVIQGAASRNAGKIIIVDVNDSKKEWASKF ***** : . *: :*.***:*,*****:**** : . : *: :*. * .***** **
AAW46801/1-383 Mycgr3P54449/1-385	GATDFINP-KDLPEGKTIIVDYLIEETDGGLDFTFDATGNVGVMRNALEACHKGWGVCTII GATDFVNPTKDLKEGEKIQDRLVEMTDGGCDYTFDCTGNVHVMRSALEACHKGWGESIII *****:*** *** ***:.* * *;* **** *;****.**** *** ,*****:*** . **
AAW46801/1-383 Mycgr3P54449/1-385	GVAPAGAEISTRPFQLVTGRVWKGSAGGGVKGRTELPGIVEDYLAGKLWVNEFVTHNETL GVAAAGQEIASTRPFQLVTGRVWKGCAGGGVKGRSQMGGKLKVDEFITHRQNL ***.** *:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****
AAW46801/1-383 Mycgr3P54449/1-385	EGINKGFDDMHAGDCIRCVVDMGF-NEAP GGINDAFHDMHAGDCIRCVVDMQKL---- ***..* .*****:*****

You can read more about Clustal alignments in the '[The Clustal Omega Multiple Alignment Package](#)' publication by Sievers and Higgins (2020).

## Additional Exercise 2: Mushroom Genes

We're going to take a look at the gene CC1G\_05700 in *Coprinopsis cinerea* okayama7#130.

From the ‘Gene’ tab, click to view the [Gene tree](#). At the bottom of the image click to collapse all the nodes at the taxonomic rank of [Class](#).

- What do you notice about the types of fungi shown in the gene tree?
- Does this match with what you would expect from the gene description? (*Hint: Agaricomycetes class belongs to the Basidiomycota phylum*)
- Based on the protein alignment shown at the right, can you predict which end of the gene/protein is most conserved?
- Click to view the [Orthologues](#) page. In the Selected orthologues table, find the entry for the species *Amanita thiersii* and click to view a protein alignment. Does this support your conclusion about the conserved region of the gene/protein?

## Additional Exercise 2 Answer: Mushroom Genes

Go to [fungi.ensembl.org](http://fungi.ensembl.org) in your browser. In the species-specific search box, select *Coprinopsis cinerea* okayama7#130 from the drop-down list and enter **CC1G\_05700**.

The screenshot shows a search interface for the Ensembl Fungi database. A search bar at the top contains the text "Search: Coprinopsis cinerea okayama7#130" followed by a dropdown arrow and the word "for". Below the search bar is a text input field containing "CC1G\_05700" and a brown "Go" button to its right. At the bottom of the interface, there is a note "e.g. NAT2 or alcohol\*".

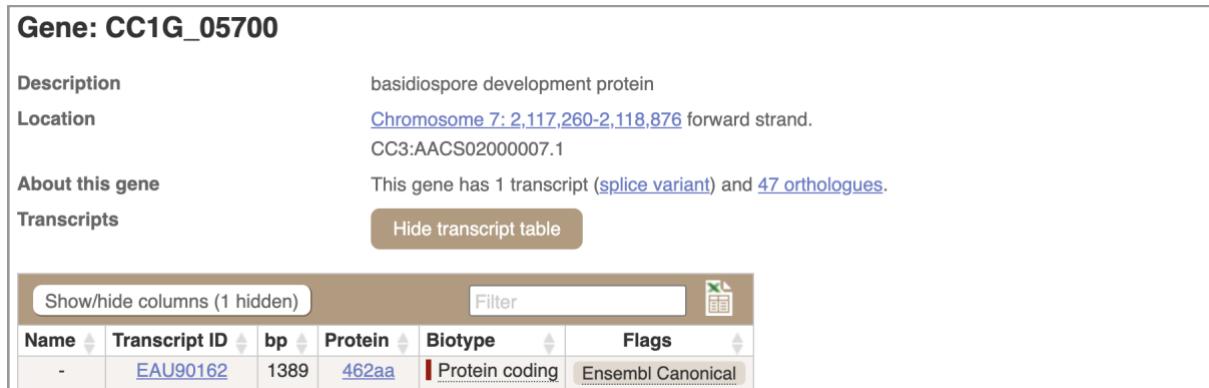
Click on the gene stable ID [CC1G\\_05700](#) to open the ‘Gene’ tab. In the left-hand panel, click on [Fungal Compara: Gene tree](#). Scroll to the bottom of the gene tree and collapse all the nodes at the taxonomic rank [Class](#) under ‘View options’.

- All fungi shown in the gene tree are Agaricomycetes:

## Evolutionary Analysis (Gene Trees): Ensembl Fungi



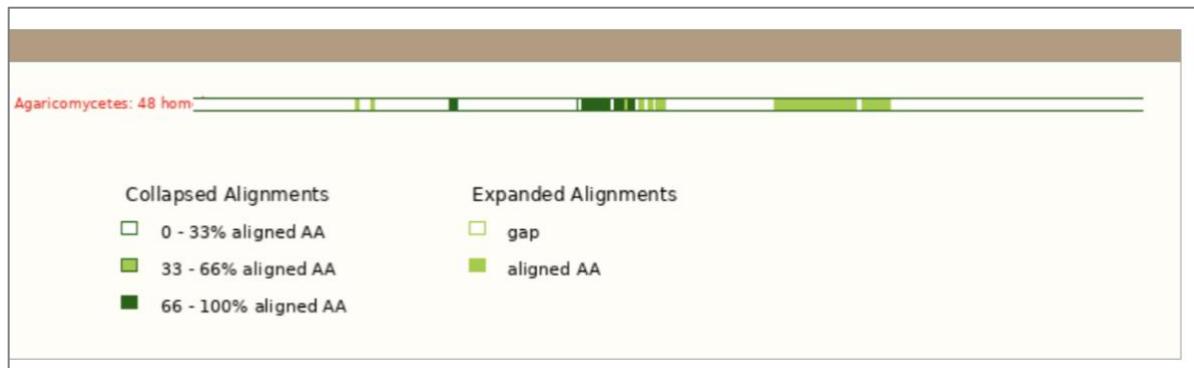
(b) The gene description is as follows:



The class Agaricomycetes belongs to the phylum Basidiomycota, therefore we would expect the gene encoding the basidiospore development protein to be conserved across Agaricomycetes species.

(c) Dark green regions in the alignment indicate highly conserved sequences (see ‘Collapsed Alignments’ legend):

## Evolutionary Analysis (Gene Trees): Ensembl Fungi



- (d) Go to **Fungal Compara: Orthologues** in the left-hand panel. Click on the **Hide ⊖** button for the ‘Summary of orthologues of this gene’ table. In the ‘Selected orthologues’ table, use the search bar in the top right-hand corner to search for *Amanita thiersii*. Click on **View Sequence Alignments** and in the pop-up menu select **View Protein Alignment**.

Selected orthologues <a href="#">Hide ⊖</a>							
Show All ▾ entries		Show/hide columns					
Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Amanita thiersii Skay4041	1-to-1	<a href="#">AMATHDRAFT_122148</a>	42.73 %	10.17 %	n/a	n/a	No
		KZ301993:102,546-102,928:-1					
		<a href="#">View Gene Tree</a>					
		<a href="#">View Sequence Alignments</a>	<a href="#">Orthologue Alignment</a> <a href="#">View Protein Alignment</a> <a href="#">View cDNA Alignment</a>				

**Orthologue alignment** ⓘ

[Download homology](#)

Type: 1-to-1 orthologues

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Coprinopsis cinerea okayama#7#130	CC1G_05700	EAU90162	462 aa	10 %	21 %	7:2117260-2118876
Amanita thiersii Skay4041	<a href="#">AMATHDRAFT_122148</a>	<a href="#">PFH51030</a>	110 aa	42 %	90 %	<a href="#">KZ301993:102546-102928</a>

CLUSTAL W (1.81) multiple sequence alignment

```

EAU90162/1-462 -----MRVLLHDTCQMNLEKFGSHGEALISNVKETSQELRKTSSTFEEHQHDKLLG
PFH51030/1-110 PLTPLDKNATSMRVLHHDTQANFEKFTSTRVDNFNLNGLAETKSEINLVKSLFERQETLTLN
*****:****:***: :...: **.*: ..* **: ::.* .
```

```

EAU90162/1-462 DIIIDLVNRSQKQLQSSIGSPAQSAAALDAFMNKVELRLESLDQRLRDLAMQAFNQTHSQALQT
PFH51030/1-110 DIIIDLVNRCQSQIQTGLGSFAQASGMEQQLSKSD1INQRDLCDLDRDAIQTV-----
*****:*,*:;*****:;::: : ::: **:****:*,.
```

```

EAU90162/1-462 QIQAIQNLIQAQQNLILNAVTPLLPLQLSPFRLAPSTSLANPSQTQRTDASSQTIEKRQ
PFH51030/1-110
```

```

EAU90162/1-462 PSYHQETLRKRQRVDSDIQEISPCKPLPGSAQKKRIESPRSVQKPSLELTQRLFPSSP
PFH51030/1-110
```

```

EAU90162/1-462 DLIKYSTDSEGPKPQVNERSAPLVTPLRPLQDLFPFFPGSNQRSVSKRPMPPSSTRLV
PFH51030/1-110
```

```

EAU90162/1-462 GPGKSATPGPSRVGAESRAALARPLRIKPLAIAPLAFSSTSCKTPVHISNFTPVPVPSL
PFH51030/1-110
```

```

EAU90162/1-462 RNAVAGEGRALKIAQTPQVLKNERMITSQAAKNTTMPPPGMVSLSRSSTTTATATKPTS
PFH51030/1-110
```

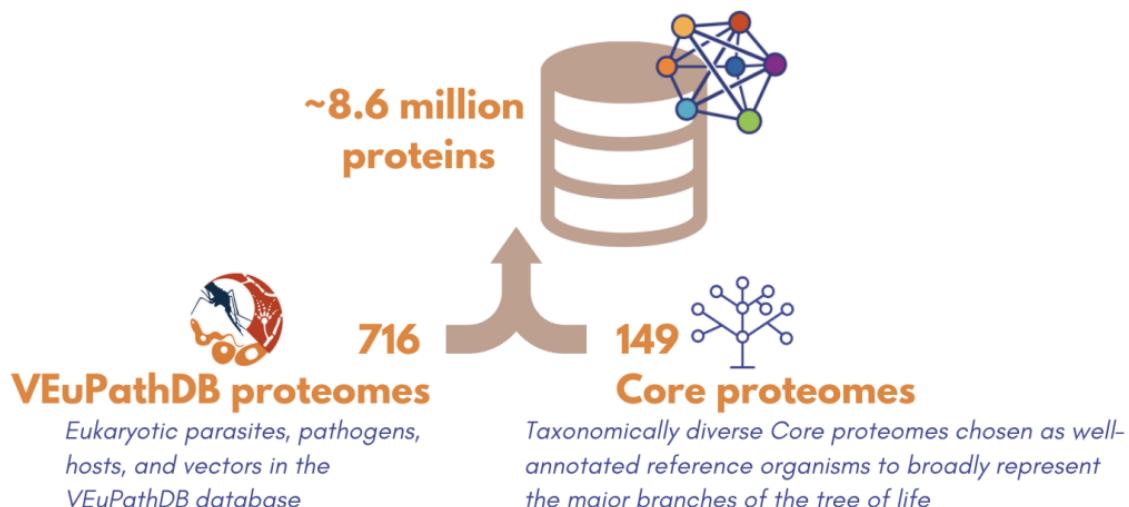
```

EAU90162/1-462 NTPRGPEANKPLLLRAPTNNGPRPLQERMKEPVREGRRFIPLVDTDDDEDS
PFH51030/1-110
```

## Phyletic Pattern search in OrthoMCL

### Learning objectives:

- Run searches in OrthoMCL.
- Run phyletic pattern searches using checkboxes or an expression.
- Combine searches using the strategy system.
- Explore individual ortholog group pages.
- Explore the group cluster graphs.

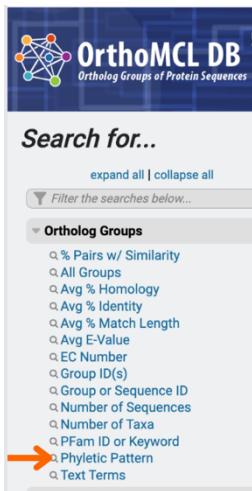


**OrthoMCL 7** is a genome-scale database and website ([orthomcl.org](http://orthomcl.org)) that uses protein sequence similarity and phylogenetic relationships among proteins to create groups of orthologous protein sequences called orthogroups. Orthogroups are clustered with OrthoFinder, which uses DIAMOND for much faster BLAST analysis, improves normalization of BLAST scores, and includes phylogenetic information to improve clustering.

OrthoMCL offers a number of tools for exploratory data analysis. Its records can be mined using search strategies that take advantage of the ability of OrthoMCL to group both known and unknown proteins into multi-species orthogroups that share protein function. Proteins with known function can be identified in a model species, and then orthologs with analogous functions can be found in less studied species. Text searches can bring in useful annotation from any organism and find related proteins in an organism of interest. For example, the Phyletic pattern search looks at the taxonomic distribution of the proteins in an orthogroup. This search can find ortholog groups that **include or exclude taxa or species that you specify**.

Below are a few exercises about using two approaches to setting up the phyletic pattern search criteria (using check boxes and regular expressions). One is not necessarily better than the other, but rather allows for the selection of different levels of granularity.

Navigate to the **Phyletic Pattern**. It is located under the Ortholog Groups search (in the left “Search for...” panel or from the Searches menu in the header) on the [OrthoMCL.org](#) home page.



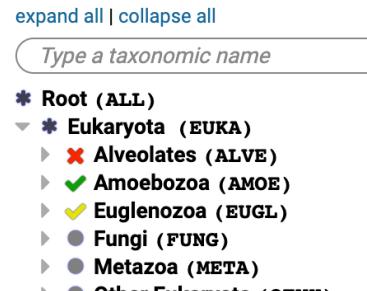
There are two ways to specify a phyletic pattern:

## 1. Using the selectable tree menu

You can click on the grey circle next to a taxon or clade (higher level) to include or exclude it from the search. Each click will cycle through the following options:

- **must be** in group (green check ✓)
- **at least one subtaxon** must be in group (yellow check ✓)
- **must not be** in group (red x)
- **no constraints** (grey circle).

Note: A grey asterisk (\*) is shown when multiple constraints are selected for sub-groups within a clade.



## 2. Using the expression box

This option allows for more precise control (both for the number of taxa in a clade and the number of proteins in those taxa) and uses a set of search terms in the expression box.

For example, the following expression **EUKA>=5T AND hsap>=10** will find ortholog groups in which there are **five or more eukaryotic taxa AND ten or more human sequences**.

In the graphical tree display:

- Click on the > icons to show or hide subtaxa and species.
- Click on the ⓘ icons to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: EUKA>=5T AND hsap>=10

[Get Answer](#)

Key: ⓘ = no constraints | ✓ = must be in group | ✓ = at least one subtaxon must be in group | x = must not be in group | \* = mixture of constraints

The phyletic expression syntax is explained on the **Learn More** tab of the search page.

Configure Search    Learn More

---

### Description

Find Ortholog Groups by phyletic pattern.

Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conservation pattern. Proteins from specific taxa are present or absent. Also, the pattern finds groups with a certain copy number (e.g., >=5).

#### Examples

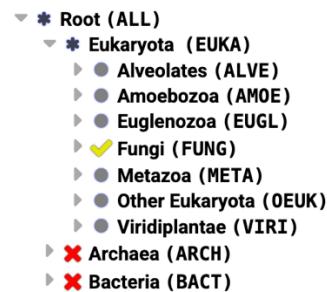
These expressions find ortholog groups in which...

<code>hsap&gt;=5</code>	there are five or more human sequences
<code>hsap+ecol=2T</code>	both human and E. coli are present.
<code>hsap+ecol=1T</code>	only one species of human or E. coli is present.

Let's practice running a few searches:

**A. Using the selectable tree, identify how many Fungal (FUNG) orthogroups do not contain any proteins from bacteria and archaea.**

The yellow check box allows orthogroups that contain a protein in any taxa that is within Fungi. How many orthogroups are found?



**B. What happens to the search result if you use the green check box (every sub-taxon) for Fungi? Why are so few groups found?**

**C. How could you Revise this search to identify orthogroups that are common to many fungi species, but not in prokaryotes? This will require use of the expression box. Try setting the expression to >=10 Fungal taxa (FUNG>=10T).**

270,031 Ortholog Groups    [Revise this search](#)

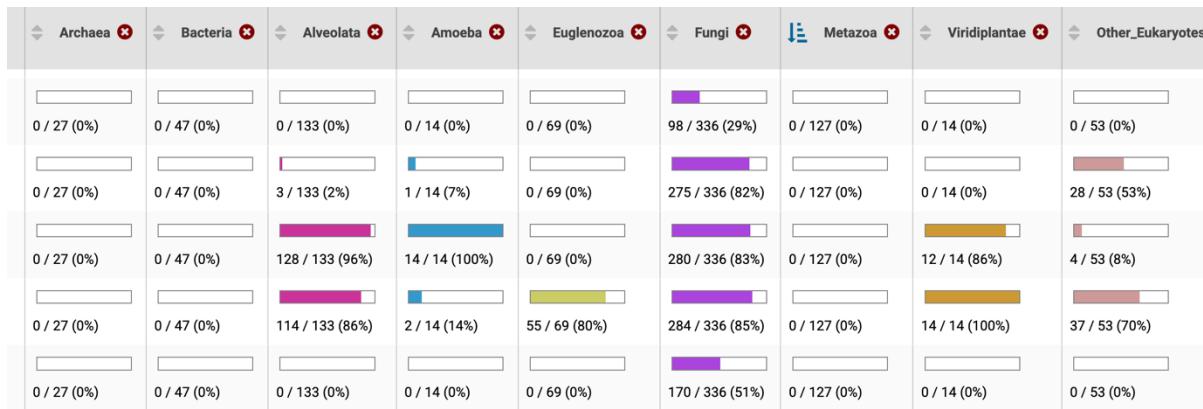
---

Ortholog Group Results

Expression: `FUNG>=10T AND ARCH=0T AND BACT=0T`

Examine results. Scroll to the right in the search results table to see a graphical representation of the taxonomic distribution of proteins in each orthogroup. Some groups are fungal specific while others include plants, animals, etc. How could you modify the search to find only fungal-specific orthogroups?

[Hint: Revise the search and set the red X in the selection tree for each group you want to exclude]



#### D. Design a search for orthogroups that are found only in Fungi, but have at least 2 proteins.

It is wise to avoid single protein orthogroups because many of these are genome assembly or annotation errors. There are two ways to create this search – either using the expression text box, or making a 2-step search strategy.

**Step 1.** Use the selection tree to choose only Fungi.

**Step 2. +Add a step** and use the **Number of Sequences** search for 2 to 10000 proteins.

How many orthogroups are found? A lot of these fungal specific orthogroups are biologically interesting such as MFS proteins (OG7\_0002006) and fungal type zinc finger transcription factor proteins (OG7\_0017974), which might serve as drug targets.

Phyletic  
257,761 Ortholog Groups

+ Add a step

Number of All Proteins

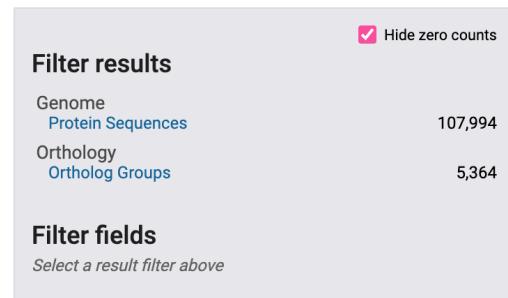
2 to 10000

#### E. Find all fungal proteins that are likely to be phosphatases and that do not have orthologs outside of fungal kingdom.

- a. Use the site search to look for \*phosphatase\* (use asterisks to find any combination of the word “phosphatase”).



How many protein sequences were identified? How many ortholog groups did you identify?



- b. Display the ortholog groups containing the word **phosphatase** in **all ortho group fields** and export the results as a search strategy.

Category	Type	Count
Genome	Protein Sequences	107,994
	Ortholog Groups	5,364

Field	Count
EC Number	22
Keywords	620
Pfam Domains	1,260
Protein Description	3,442

Text  
**3,960 Ortholog Groups**

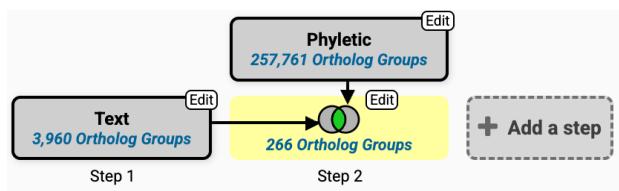
Step 1

+ Add a step

- c. Add a step and run a phyletic pattern search for groups that contain any fungi proteins but do not contain any other organism outside fungi. (**Hint:** make sure everything has a red **x**, except for fungi, which should be a yellow check (contains at least one protein)).

- ▼ \* Root (ALL)
- ▼ \* Eukaryota (EUKA)
  - ▷ ✗ Alveolates (ALVE)
  - ▷ ✗ Amoebozoa (AMOE)
  - ▷ ✗ Euglenozoa (EUGL)
  - ▷ ✓ Fungi (FUNG)
  - ▷ ✗ Metazoa (META)
  - ▷ ✗ Other Eukaryota (OEUK)
  - ▷ ✗ Viridiplantae (VIRI)
- ▷ ✗ Archaea (ARCH)
- ▷ ✗ Bacteria (BACT)

How many groups did the search return?



Strategy URL:

#### Strategies:

- A: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/4764a59520b7e517>
- B: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/77a327acd20b5a2d>
- C: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/2a7a12997022153b>
- D: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/2a7a12997022153b>
- E: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/23d49ae68fa7b642>

# MycoCosm: Secondary Metabolism Clusters (SMURF) Browser

In fungi, secondary metabolite (SM) genes are often organized in chromosomal clusters dedicated to that metabolite's biosynthetic pathway. Each portal's SM Clusters Browser facilitates display and discovery of MycoCosm's SM-annotated genes. JGI has begun to load SM predictions from more than one software tool. The portal annotation option used to be called "Secondary Metabolite Clusters" is now called "SMURF Clusters" to reflect the method we will explore below.

**Scenario:** You have identified a toxic SM produced by *Septoria musiva*, a pathogenic fungus that induces cankers in the poplar tree, but not produced by *Septoria populincola*, which infects a different species of poplar and does not induce cankers. The SM's structure suggests that its biosynthetic pathway may have as its core enzyme a hybrid PKS-NRPS (polyketide synthase-nonribosomal peptide synthetase). Use MycoCosm to find candidate gene clusters for this pathway.

- 1) Go to the MycoCosm *Septoria* PhyloGroup at mycocosm.jgi.doe.gov/Septoria. Both species are represented in the group:

Info • Septoria

SEARCH	BLAST	ANNOTATIONS ▾	MCL CLUSTERS	GEO MAPPING	DOWNLOAD	INFO	HELP!																		
<table border="1"><thead><tr><th>##</th><th>Name</th><th>Assembly</th><th>Length</th><th># Genes</th><th>Published</th></tr></thead><tbody><tr><td>1</td><td><a href="#">Septoria musiva SO2202 v1.0</a></td><td></td><td>29,352,103</td><td>10,233</td><td><a href="#">Ohm RA et al., 2012</a></td></tr><tr><td>2</td><td><a href="#">Septoria populincola v1.0</a></td><td></td><td>33,188,813</td><td>9,739</td><td><a href="#">Ohm RA et al., 2012</a></td></tr></tbody></table>								##	Name	Assembly	Length	# Genes	Published	1	<a href="#">Septoria musiva SO2202 v1.0</a>		29,352,103	10,233	<a href="#">Ohm RA et al., 2012</a>	2	<a href="#">Septoria populincola v1.0</a>		33,188,813	9,739	<a href="#">Ohm RA et al., 2012</a>
##	Name	Assembly	Length	# Genes	Published																				
1	<a href="#">Septoria musiva SO2202 v1.0</a>		29,352,103	10,233	<a href="#">Ohm RA et al., 2012</a>																				
2	<a href="#">Septoria populincola v1.0</a>		33,188,813	9,739	<a href="#">Ohm RA et al., 2012</a>																				

- 2) Click on '*Septoria musiva SO2202 v1.0*' to go to its genome portal:

Home • Septoria musiva SO2202 v1.0

[SEARCH](#) [BLAST](#) [BROWSE](#) [ANNOTATIONS ▾](#) [PATHWAYS](#) [MCL CLUSTERS](#) [SYNTENY](#) [DOWNLOAD](#) [INFO](#) [HOME](#) [STATUS](#) [HELP!](#)



*Septoria musiva* (sexual stage: *Mycosphaerella populincola*) causes leaf spots and cankers on poplars (*Populus spp.* and hybrids). On native North American poplars the pathogen mainly causes leaf spots that can lead to defoliation but generally do not kill the host. But *S. musiva* can also cause cankers on branches and primary stems. These can be lethal and are particularly severe on hybrid poplars in plantations. They often develop on the primary shoots of 2- to 3-year-old trees, leading to restrictions in the movement of water and nutrients and weakening the wood within a few feet of ground level. The weakened trunks collapse easily, greatly reducing the production of biomass. Cankers caused by *S. musiva* can greatly hamper the production of hybrid poplars in the eastern United States and Canada and threaten poplars in western North America.

A major concern with *S. musiva* is with migration to new areas. The pathogen is endemic and appears to have originated on poplars in eastern North America, where it occurs commonly on leaves of the eastern cottonwood, *P. deltoides*. During the past 20 years *S. musiva* has appeared in South America and western Canada, where it is spreading rapidly on native and hybrid poplars causing economic damage as well as threatening native poplars in important riparian zones. It is not yet known in Europe or Asia but has the potential to cause extensive damage if introduced to those areas. Global warming and trade may facilitate the spread of the disease by making northern popular-growing areas more favorable to growth of the fungus.

Availability of a genome sequence for *S. musiva* will help with designing strategies to effectively manage this destructive disease. The genome of the black cottonwood poplar *P. trichocarpa*, another host for *S. musiva*, has been sequenced and provides a rare opportunity to analyze host-pathogen interactions when both host and pathogen have been sequenced. Opportunities for comparative genomics also are available. Other members of the genus *Mycosphaerella* with sequenced genomes include the pine pathogen *Dothistroma septosporum* (aka *Mycosphaerella pini*), the banana pathogen *M. fijiensis* and the wheat pathogen *M. graminicola*. Comparative genomics of these sequences with that of *S. musiva* will help identify the differences involved in pathogenicity to woody vs. herbaceous hosts and could help in our understanding of why *S. musiva* causes cankers on hybrids but not native poplars. Understanding the mechanisms of fungal pathogenicity and of resistance in the host will be essential for protecting our forests and ensuring a stable supply of renewable biomass as the climate warms in the future.

Photo credit: Glen Stanisz, Ph.D., University of Wisconsin-Madison

- 3) Click on "ANNOTATIONS => SMURF CLUSTERS" to go to the portal's SM clusters browser:

- 4) Scroll through the ‘Genomes’ list box and select both ‘*Septoria musiva*’ and ‘*Septoria populincola*’, and only those 2 species. Click the ‘Refresh’ button. Only the SM cluster core gene counts of the 2 *Septoria* spp. are shown, and may be directly compared. *S. musiva* has 2 hybrid core genes (PKS-NRPS genes) while *S. populincola* has none:

SMURF Clusters • Septoria musiva SO2202											
SEARCH	BLAST	BROWSE	ANNOTATIONS	PATHWAYS	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!
<b>Genomes</b>						<b>Cluster Type</b>					
Septoria musiva SO2202 v1.0						all					
Septoria populincola v1.0						<input type="button" value="Refresh"/>					
						DMAT	HYBRID	NRPS	NRPS-Like		
Genome	DMAT	HYBRID	NRPS	NRPS-Like	PKS	PKS-Like	TC	Total			
Septoria musiva SO2202 v1.0	0	2	7	8	9	2	2	30			
Septoria populincola v1.0	0	0	8	7	9	2	3	29			
<b>Total</b>	<b>0</b>	<b>2</b>	<b>15</b>	<b>15</b>	<b>18</b>	<b>4</b>	<b>5</b>	<b>59</b>			

- 5) There are a total of 2 genes in the Hybrid column. Click on the number to show a graphical representation of the 2 gene clusters in *S. musiva*. The ‘Size’ column displays each cluster’s length, and the ‘Genes’ column displays each cluster’s core PKS-NRPS gene (in color) and its accessory, decorator, and other genes (in gray). A core hybrid gene is typically very large, but the total cluster size can be highly variable. To resize the 2 clusters to scale to each other, go to the ‘Scale’ pull-down menu, select ‘Across All Clusters’, and click on the ‘Refresh’ button:

SEARCH BLAST BROWSE ANNOTATIONS PATHWAYS MCL CLUSTERS SYNTENY DOWNLOAD INFO HOME STATUS HELP!

Genomes	Cluster Type	Scale	Clusters Per Page
Septoria musiva SO2202 v1.0 Septoria populincola v1.0	all DMAT <b>HYBRID</b> NRPS NRPS-Like	<input checked="" type="checkbox"/> Per Cluster Per Cluster No Gaps Across All Clusters	0 <input type="button" value="Refresh"/>

Total 2 cluster(s) found. 1

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179	
Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058	

- 6) Each gene in the clusters is represented by an arrow with a single pair of fletching that indicates the gene's 5' to 3' direction. Mouse-over the top cluster's core gene to get more information about the PKS-NRPS hybrid. The listed domains are typical of a hybrid enzyme:

Total 2 cluster(s) found. 1

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179	
Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058	

Contact Us Cite Us Accessibility/Section 508  
 Disclaimer Credits  
 © 1997-2023 The Regents of the University of California.  
 Mycocosm Portal version:17.160 myco-web-3.jgi.lbl.gov Release Date:11-Apr-2023 13:29 PDT

- 7) To get domain information about the other genes in the SM cluster, mouse-over them too. The next gene 3' to the core gene has a p450 domain:

Total 2 cluster(s) found. 1

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Sepmu1.24	HYBRID	scaffold_6:1522811-1553990	31,179	
Sepmu1.25	HYBRID	scaffold_6:1977373-2004431	27,058	

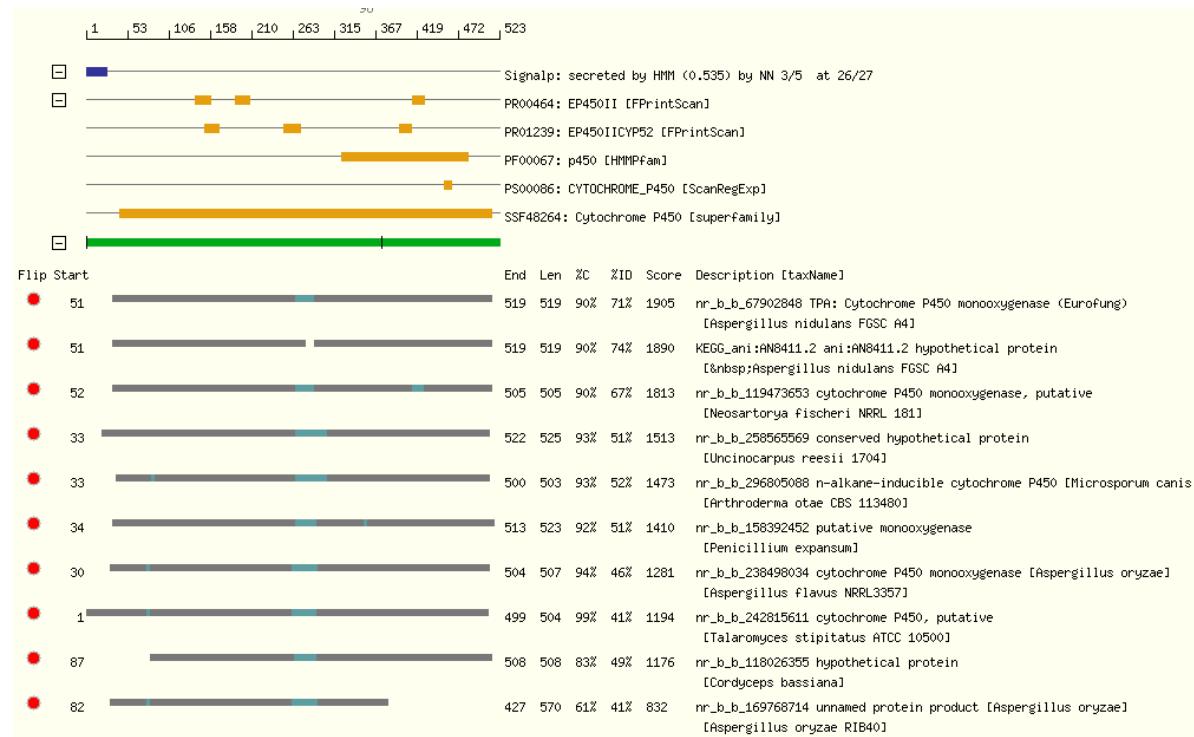
Type: HYBRID  
 ProteinId: 87793  
 Domains: PF00067 p450 1

- 8) To get more detailed information about a gene, click on it directly. Click on the gene with the p450 domain to see its ‘protein page’. Examination of the protein page reveals that:
- The gene is expressed. The blue bars represent UTRs, which can be inferred only from transcriptomic data.
  - The protein has p450 Pfam and other annotations indicative of a cytochrome p450 monooxygenase.
  - The best Blast hit in nr is a cytochrome p450 monooxygenase from *Aspergillus nidulans*, which belongs to a different class of fungi (Eurotiomycetes) from *Septoria* (Dothideomycetes).

**Best BLAST hit**

SEARCH	BLAST	BROWSE	ANNOTATIONS ▾	MCL CLUSTERS	SYNTENY	DOWNLOAD	INFO	HOME	STATUS	HELP!
Name:	estExt_Genewise1.C_6_t30338									
Protein ID:	87793									
Location:	<a href="#">scaffold_6:1535323-1537114</a>									
Strand:	+									
Number of exons:	2									
Description:	<a href="#">gi 67902848 ref XP_681680.1  hypothetical protein AN8411.2 [Aspergillus nidulans FGSC A4] &gt;gi 40747877 gb EAA67033.1  hypothetical protein AN8411.2 [Aspergillus nidulans FGSC A4] &gt;gi 259484346 ipe CBF80485.1  TPA: Cytochrome P450 monooxygenase (Eurofung) [Aspergillus nidulans FGSC A4] (model%: 91, hit%: 90, score: 1905, %id: 71) [Aspergillus nidulans FGSC A4]</a>									
total hits(shown)	683 (10)									
<b>ASPECT</b>	<b>GO Id</b>	<b>GO Desc</b>		<b>Interpro Id</b>	<b>Interpro Desc</b>					
Molecular Function	<a href="#">0016712</a>	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen		<a href="#">IPR002974</a>	Cytochrome P450, E-class, CYP52					
	<a href="#">0004497</a>	monooxygenase activity		<a href="#">IPR002402</a>	Cytochrome P450, E-class, group II					
	<a href="#">0020037</a>	heme binding		<a href="#">IPR001128</a>	Cytochrome P450					
		<a href="#">0005506</a>	iron ion binding		<a href="#">IPR002402</a>	Cytochrome P450, E-class, group II				
Biological Process	<a href="#">0006118</a>	electron transport		<a href="#">IPR002974</a>	Cytochrome P450, E-class, CYP52					
				<a href="#">IPR001128</a>	Cytochrome P450					
					<b>KOG Desc</b>					
<b>KOG GROUP</b>	<b>KOG Id</b>	<b>KOG Class</b>								
Metabolism	<a href="#">KOG0158</a>	<a href="#">Secondary metabolites biosynthesis, transport and catabolism</a>								
<a href="#">View/modify manual annotation</a> <a href="#">View nucleotide and 3-frame translation</a> <a href="#">To Genome Browser</a> <b>NCBI blastp</b> Predicted number of transmembrane domains: 1										
<p><b>Blue: UTRs</b> <b>Red: CDS</b></p>										
<p>Signalp: secreted by HMM (0.535) by NN 3/5 at 26/27</p> <p>PR00464: EP450II [FFprintScan]      PR01239: EP450IICYP52 [FFprintScan]      PF00067: p450 [HMMpFam]      PS00086: CYTOCHROME_P450 [ScanRegExp]      SSF48264: Cytochrome P450 [superfamily]</p> <p><b>InterPro annotations (For example, Pfam domains)</b></p>										

- 9) Based on the annotations and top hits, it seems that this gene is indeed a cytochrome p450 monooxygenase, a class of enzymes that often modify core structures of SM biosynthetic pathways. Similar perusal of the other genes of the cluster says that this cluster is an excellent candidate for synthesis of your SM.



- 10) One explanation for *S. musiva* having this cluster and the congeneric *S. populica* not is that the former acquired the cluster by horizontal gene transfer from a phylogenetically distant source. The ‘best Blast hit’ of the cytochrome p450 enzyme supports this hypothesis. To see if the core enzyme can shed some light, click the web browser back button to go back to the SM CLUSTERS graphic, and click on the same PKS-NRPS core gene we moused over earlier. The protein page is rich in details, including domains and the top 10 hits. All of the hits are high quality and are from Eurotiomycetes. This cluster is an excellent candidate for horizontal gene transfer from the Eurotiomycetes!

## References:

- Dhillon B, Feau N, Aerts AL, Beauseigle S, Bernier L, Copeland A, Foster A, Gill N, Henrissat B, Herath P, LaButti KM, Levasseur A, Lindquist EA, Majoor E, Ohm RA, Pangilinan JL, Pribowo A, Saddler JN, Sakalidis ML, de Vries RP, Grigoriev IV, Goodwin SB, Tanguay P, Hamelin RC. Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. Proc Natl Acad Sci U S A. 2015 Mar 17;112(11):3451-6. doi: 10.1073/pnas.1424293112. Epub 2015 Mar 2. PubMed PMID: 25733908
- Schümann J, Hertweck C. Molecular basis of cytochalasan biosynthesis in fungi: gene cluster analysis and evidence for the involvement of a PKS-NRPS hybrid synthase by RNA silencing. J Am Chem Soc. 2007 Aug 8;129(31):9564-5. Epub 2007 Jul 18. PubMed PMID: 17636916.

# FungiDB & MycoCosm: Secondary Metabolites and clusters

## Learning objectives:

- Explore InterPro search in FungiDB
- Cross-reference the results with MycoCosm data

Fungi generate a wide variety of secondary metabolites. These metabolites can be categorised based on the initial stage of their biosynthesis, particularly the essential “key enzymes” involved: Non-ribosomal peptide synthetases (NRPSs), NRPS-like enzymes, Polyketide synthases (PKSs), PKS-like enzymes, Hybrid PKS-NRPS, Prenyltransferases (DMAT), and Terpene cyclases/synthases (TC).

### • Identify NRPS genes in FungiDB

1. Use the InterPro search to identify NRPS genes in all *Aspergilli*.

NRPS genes have at least three domains:

- AMP-binding (PF00501)
- PP-binding (PF00550)
- Condensation (PF00668)

The screenshot shows the FungiDB search interface for InterPro domains. At the top, there is a search bar labeled "Search for..." with the term "inter" entered. Below the search bar is a section titled "Genes" which includes links to Pathways and interactions, Metabolic Pathway, Substrates/Products, Y2h Protein Interactions, and Protein features and properties, specifically highlighting "InterPro Domain". A detailed description of the InterPro Domain feature is provided, stating it finds genes containing a specified protein domain from the InterPro database, mentioning CATH, CCD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, PRODOM, PROFILE, PROSITE, SF3D, SMART, SUPERFAMILY, and TIGRFAMs.

The main search results are displayed in three sections:

- Identify Genes based on InterPro Domain**: Shows a list of Aspergillus species with checked boxes for PF00501, PF00550, and PF00668 domains. Species listed include A. aculeatus, A. brasiliensis, A. campestris, A. carbonarius, A. clavatus, A. crinitus, A. eucalypticola, A. fijiensis, A. fischeri, A. flavus, A. fumigatus, A. glaucus, A. heteromorphus, A. lentinus, A. lucifer, A. niger, A. nidulans, A. oryzae, A. parasiticus, A. sydowii, A. terreus, and A. versicolor.
- Domain Database**: Shows a dropdown menu set to PFAM.
- Specific Domain(s)**: Displays the results for PF00501: AMP-binding AMP-dependent synthetase/ligase, PF00550: PP-binding Phosphopantetheine binding ACP domain, and PF00668: Condensation Condensation domain.

Below the search results, a flowchart illustrates the NRPS gene assembly process:

```
graph TD; A[PF00501: 2,639 Genes] --> B[PF00550: 2,164 Genes]; B --> C[PF00668: 858 Genes]; C --> D[766 Genes]; D --> E[Add a step]
```

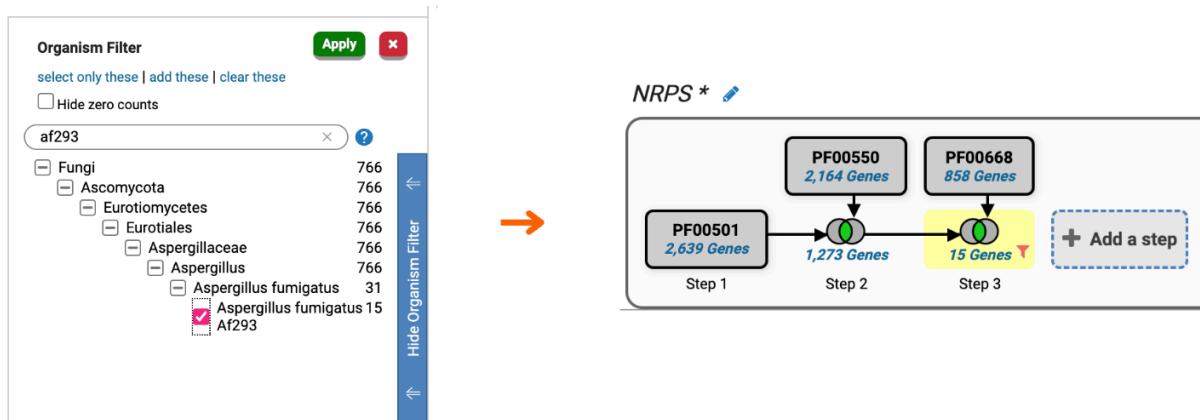
The flowchart is labeled "NRPS" and shows three steps: Step 1 (PF00501, 2,639 Genes), Step 2 (PF00550, 2,173 Genes), and Step 3 (PF00668, 766 Genes). An "Add a step" button is also present.

Strategy URL:

<https://fungidb.org/fungidb/app/workspace/strategies/import/85a1e3a5a603efc6>

- How many genes were identified in *Aspergillus fumigatus* Af293?

Hint: use the organism filter on the left to limit your search results to Af293 genes only.



- Use MycoCosm to explore *Afum* NRPS genes. Access the *A. fumigatus* Af293 portal (<https://mycocosm.jgi.doe.gov/Aspfu1>) and navigate to the SMURF Clusters page (under the “Annotations” tab; this used to be called the “Secondary Metabolism Clusters” page, but was renamed to differentiate between different methodologies). How many genes did you get?

The screenshot shows the 'Annotations' tab in the MycoCosm portal, specifically the SMURF Clusters page. The top navigation bar includes SEARCH, BLAST, BROWSE, ANNOTATIONS (selected), MCL CLUSTERS, SYNTENY, DOWNLOAD, INFO, HOME, and HELP!. The 'Cluster Type' dropdown is set to 'NRPS'. Below the table, a note says 'Total 9 cluster(s) found. 1'. The table has columns: Cluster Id, Cluster Type, Scaffold, Size (bp), and Genes. Each row shows a genomic region with arrows indicating gene direction and color coding for different genes. The last row is a summary.

Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes
Aspfu1.5	NRPS	Chr_3_A_fumigatus_Af293.876157-937897	61,740	[Gene tracks]
Aspfu1.7	NRPS	Chr_3_A_fumigatus_Af293.3423866-3446129	22,263	[Gene tracks]
Aspfu1.10	NRPS	Chr_3_A_fumigatus_Af293.4007787-4023468	15,681	[Gene tracks]
Aspfu1.15	NRPS	Chr_1_A_fumigatus_Af293.2655644-2694887	39,243	[Gene tracks]
Aspfu1.16	NRPS	Chr_1_A_fumigatus_Af293.4662924-4713331	50,407	[Gene tracks]
Aspfu1.18	NRPS	Chr_8_A_fumigatus_Af293.20854-49410	28,556	[Gene tracks]
Aspfu1.28	NRPS	Chr_5_A_fumigatus_Af293.3307809-3342792	34,983	[Gene tracks]
Aspfu1.31	NRPS	Chr_6_A_fumigatus_Af293.2334637-2372302	37,665	[Gene tracks]
Aspfu1.32	NRPS	Chr_6_A_fumigatus_Af293.3004871-3035305	30,434	[Gene tracks]
Cluster Id	Cluster Type	Scaffold	Size (bp)	Genes

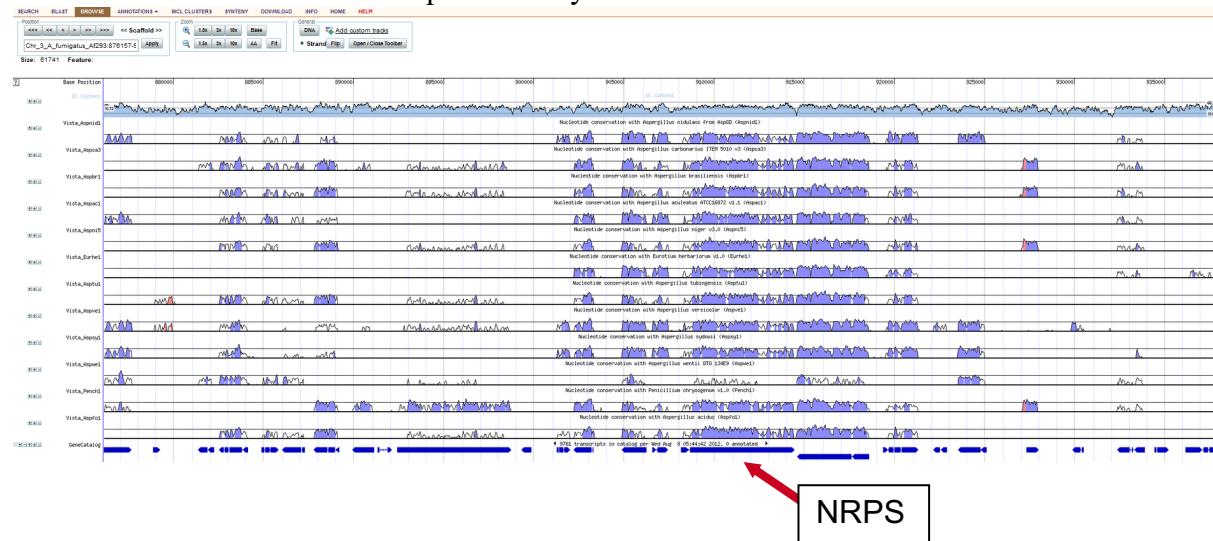
- What do you think may be causing the difference in the predicted gene number?

This view on MycoCosm allows you to analyze backbone and auxiliary proteins across the entire predicted secondary metabolism cluster.

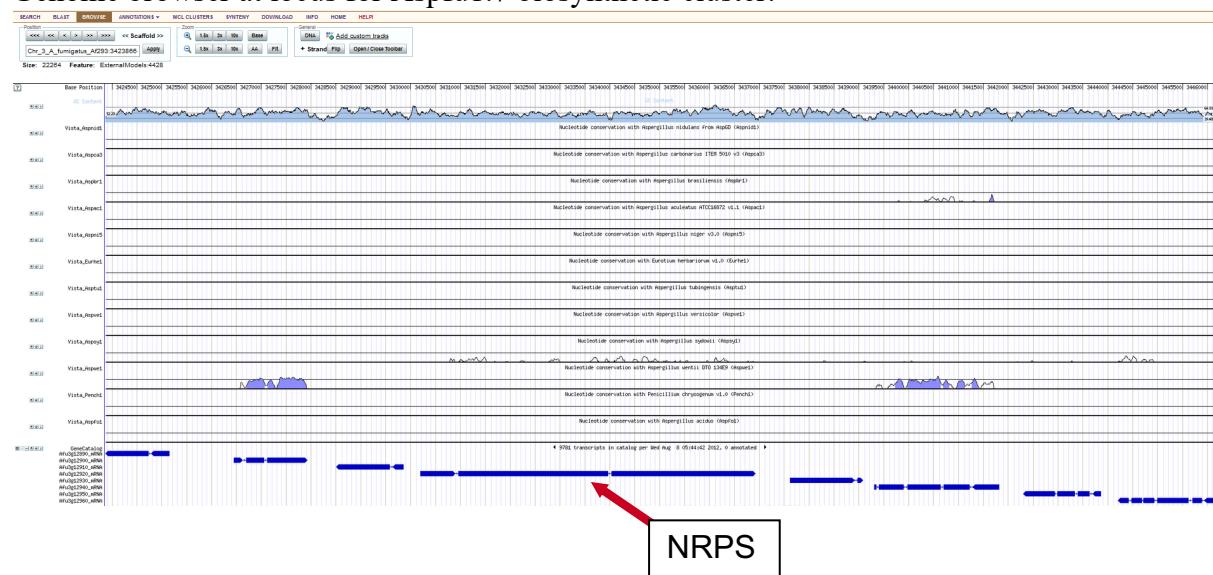
- How conserved are these secondary metabolite clusters across related Aspergilli?
- Click on the scaffold coordinates for Aspfu1.5 and analyze the Vista curve tracks in the genome browser.
- How many related Aspergilli show some synteny with this region?
- **Optional:** Repeat this exercise for the next cluster, Aspfu1.7.

- Answer: Synteny is observed across most Aspergilli for Aspfu1.5, raising the possibility that this SM cluster is widespread across the genus. However, Aspfu1.7 shows no synteny except for a couple of auxiliary genes in *Aspergillus wentii*, suggesting that it is possibly lineage-specific.

## Genome browser at locus for Aspfu1.5 biosynthetic cluster:



## Genome browser at locus for Aspfu1.7 biosynthetic cluster:



## Working with user datasets in FungiDB

FungiDB offers options for uploading user data into private spaces where you can take advantage of the FungiDB infrastructure when analyzing your own data.

User datasets can be uploaded into FungiDB via the My Workspace. The tools to do this are located under My Workspace > My data sets -

<https://fungidb.org/fungidb/app/workspace/datasets/new>

The tools in the **My Data Sets** section allow you to upload:

1. Gene Lists (as a .txt file),

Choose an upload type

2. bigWig files (as a .bw file),

Gene List

Integrate your gene list in FungiDB.

bigWig

Integrate your bigWig data in FungiDB.

3. Normalized RNA-Seq data

(as a zip file containing RNA-seq data with or without the supporting bigwig files).

Normalized RNA-Seq

Integrate your Normalized RNA-Seq data in FungiDB.

Uploading data via Gene List and bigwig options is straightforward. You just need a text file with Gene IDs or a bigwig file as .bw file. For the third option, the normalized RNA-Seq data, you need to upload a collection of files in a format compatible with FungiDB infrastructure. Below, we will go through steps on how to obtain the normalized RNA-Seq data with UseGalaxy using a sample workflow.

**In this workshop, you will work with pre-analyzed and pre-formatted RNA-Seq data. Instructions on accessing these files can be found on the LMS.**

### Data analysis in UseGalaxy.org

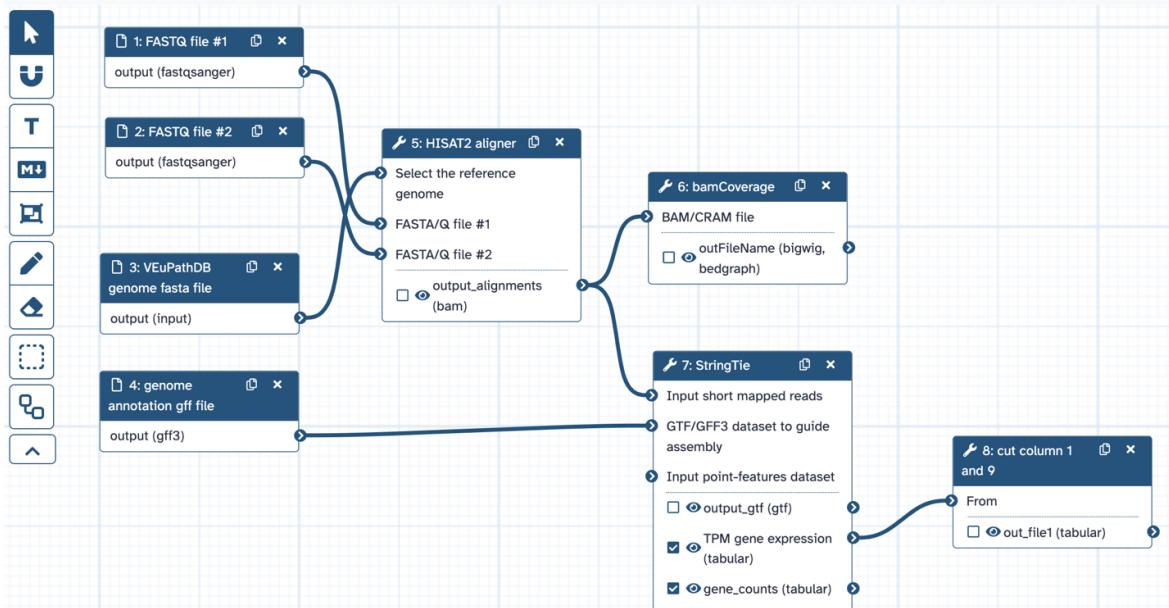
Galaxy (<https://usegalaxy.org/>) is an open-source, web-based platform for data-intensive biomedical research. They offer a multitude of tutorials and thousands of tools for data manipulation and analysis. You can access tutorials and additional resources at this URL: <https://galaxyproject.org/> to learn how to analyze omics data.

Experimental sequencing data or public data in the GenBank Sequence Read Archive (SRA) is usually available as raw reads in FASTQ format. To analyze this data on VEuPathDB, the reads must be aligned to a reference genome, and per-gene expression values counted according to a map of gene annotations in gff format.

A Galaxy workflow resembles a recipe or a sequential method for data analysis. This link directs you to a publicly shared workflow that utilizes a VEuPathDB reference genome to analyze RNA-Seq data and yield results compatible with FungiDB tools:

<https://usegalaxy.org/u/stuart/w/rnaseq-with-hisat2-stringtie--tpm>.

This workflow takes paired-end fastq files as input, aligns to the reference genome with HISAT2, makes a bigwig coverage map with DeepTools bamCoverage, counts expression per gene with StringTie, and cuts the geneID and TPM normalized expression values from the StringTie output.



## Components of the sample RNA-Seq workflow.

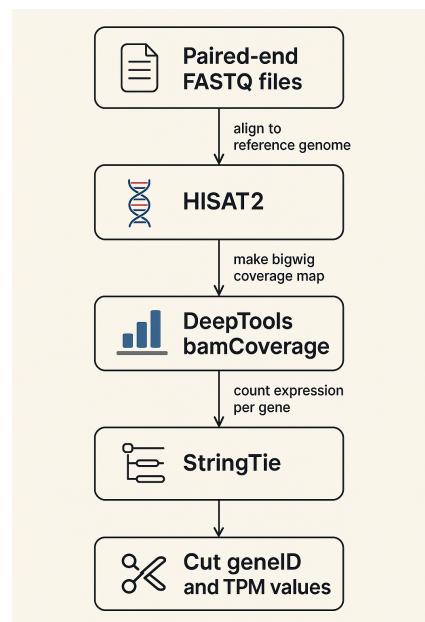
**Paired-end fastq files (input):** These are files that contain raw RNA sequencing data. "Paired-end" means the RNA was sequenced from both ends, giving more accurate information.

**FASTQ** is a file format used to store both the DNA/RNA sequences and the quality of each sequence. **HISAT2** is a fast tool that lines up (or aligns) the RNA reads from your FASTQ file to a known reference genome (like a map of an organism's DNA). This helps identify where each RNA piece came from in the genome.

**bamCoverage** is part of a toolkit called DeepTools. It turns the alignment data (from HISAT2) into a BigWig file, which is a visual summary showing where in the genome there is high or low RNA expression. This is useful for creating graphs and seeing patterns in gene activity.

**StringTie** is a tool that looks at the aligned RNA data and calculates how much each gene is being expressed (i.e., how active it is in the cell). It reconstructs transcripts and estimates their abundance. The output from StringTie includes a lot of information. This step extracts just the gene ID and the TPM (Transcripts Per Million) values.

**TPM** is a way of standardizing gene expression data so you can compare across genes and samples more easily.



To perform data analysis (e.g. RNA-Seq) on UseGalaxy.org, you must supply a reference genome (sequence (.fasta) and annotation (.gff) files). You can do this via several options, including:

1. Using the dropdown menu in Galaxy (available for some tools and selected genomes). Make sure to pay attention to the genome version.

The screenshot shows the 'Tool Parameters' section for the HISAT2 tool. Under 'Source for the reference genome', there is a link to 'Use a built-in genome'. Below it, a note says 'Built-in references were created using default options'. A dropdown menu labeled 'Select a reference genome \*' is open, showing 'Select Value' at the top. Two options are listed: 'Aspergillus niger (GCF\_000002855.4\_ASM285v2)' and 'Aspergillus niger ATCC 1015 (GCA\_000230395.2\_ASPPNI\_v3.0)', with the latter being highlighted in a blue box.

2. Providing a direct URL pointing to the fasta and gff files in VEuPathDB (the easiest way to make sure you are using the latest FungiDB). To use the URL, select the Paste/Fetch data from the Tools section in Galaxy.

The screenshot shows the Galaxy 'Upload from Disk or Web' interface. On the left is a sidebar with various tools and history sections. The main area has tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based', with 'Regular' selected. It displays a message: 'You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.' Below this is a text input field containing two URLs: 'https://fungidb.org/common/downloads/release-68/CaurisB8441/fasta/data/FungiDB-68\_CaurisB8441\_Genome.fasta' and 'https://fungidb.org/common/downloads/release-68/CaurisB8441/gff/data/FungiDB-68\_CaurisB8441.gff'. At the bottom are buttons for 'Choose local file', 'Choose remote files', 'Paste/Fetch data' (which is checked), 'Start', 'Pause', 'Reset', and 'Close'.

3. Manually uploading both files.

FungiDB genome files can be found under the Data > [Download data files](#) menu. To find relevant files, choose the organism from the taxonomic checklist and the current release, then check boxes for the **Genome fasta** and the Annotation and Curation **Full GFF** file.

The screenshot shows the FungiDB 'Download Data Files' interface. At the top, there's a navigation bar with links like Site search, My Strategies, Searches, Tools, My Workspace, Data, About, Help, Subscriptions, Contact Us, and a 'My Organism Pref' button. Below the navigation is a search bar and a main content area titled 'Download Data Files'. It includes a note about using the tool to download genome scale files such as genome fasta or GFF files. A 'selected out of 331' message is displayed above a tree view of organisms. The tree shows 'Auris' expanded, with 'Fungi' and 'Ascomycota' collapsed. Under 'Auris', 'Metchnikowiales' is expanded, showing several Candida auris strains, with 'Candida auris strain B8441 [Reference]' checked. Below this is a 'Data File properties' section with tabs for Release, Contents, Category, and File Format. The 'File Format' tab is selected, showing checkboxes for fasta (checked), gff (checked), txt, gaf, gzt, and smd. A table at the bottom lists 6 rows of selected data files, each with columns for File, Organism, Release, Category, Contents, and File Format. The first row is highlighted in blue.

File	Organism	Release	Category	Contents	File Format
FungDB-68_CaurisB8441.gff	Candida auris strain B8441	68	Annotation and Curation	Full GFF	gff
FungDB-68_CaurisB8441_AnnotatedCDSS.fasta	Candida auris strain B8441	68	Sequence	CDS	fasta
FungDB-68_CaurisB8441_AnnotatedProteins.fasta	Candida auris strain B8441	68	Sequence	Protein	fasta
FungDB-68_CaurisB8441_AnnotatedTranscripts.fasta	Candida auris strain B8441	68	Sequence	Transcript	fasta
FungDB-68_CaurisB8441_Genome.fasta	Candida auris strain B8441	68	Sequence	Genome	fasta
FungDB-68_CaurisB8441_Orf50.gff	Candida auris strain B8441	68	Sequence	ORF	gff

## Getting raw sequence data into Galaxy.

Above are just three methods to input your data into Galaxy, though many more options are available.

For example, raw sequence data in FASTQ format can be transferred directly from SRA. Just enter the SRA accession numbers in Galaxy Tools > Get Data > **Faster Download and Extract Reads in FASTQ format from NCBI SRA**.

Note: When uploading files manually, it is essential to set their format, as this is important for the steps downstream. For example, for the RNA-Seq workflow, set the file type as fastqsanger.gz; this is what HISAT2 will expect as input. The 2 FASTQ files for paired-end sequencing of a single sample will automatically become data collections (a list with 1 fastqsanger.gz pair).

When the Reference and RNA-Seq FASTQ files have been loaded into Galaxy, they will become selectable within the workflow. For example, to deploy an RNA-Seq workflow, you will need to select your pair-end data (FASTQ files), reference genome, and annotation gff file and click on the Run Workflow button to deploy data analysis.

The screenshot shows the Galaxy interface. On the left, there's a sidebar with icons for Upload, Tools (which is selected), Workflows, Workflow Invocations, Visualization, Histories, and History Multiview. The main area is titled 'All Tools' and shows a search bar with 'search tools'. Below the search bar is a list of available tools:

- Download and Extract Reads in BAM format from NCBI SRA
- Download and Extract Reads in FASTQ format from NCBI SRA
- Download run data from EBI Metagenomics database
- EBI Search to obtain search results on resources and services hosted at the EBI
- EBI SRA ENA SRA
- EGA Download Client
- Faster Download and Extract Reads in FASTQ format from NCBI SRA

After the workflow is complete, TPM counts are contained within the output from the “Cut on Collection” and bigWig files are generated by the BamCoverage tool.

### Uploading data in FungiDB.

We will work with two data types for *Candida auris*:

- **Study 1:** RNA-Seq PRJEB60034 - *C. auris* was treated with tunicamycin (low and high drug concentrations). Note, reference genome - *C. auris* B11221.
- **Study 2:** A gene list from Lara-Aguilar et al. 2021 – *C. auris* was treated with high doses of caspofungin. Note, reference genome – *C. auris* B8441.  
<https://doi.org/10.1080/21505594.2021.1927609>

**Homework: Please download the Study 1 and Study 2 files from LMS and complete Parts A and B of this exercise before the session on Zoom.**

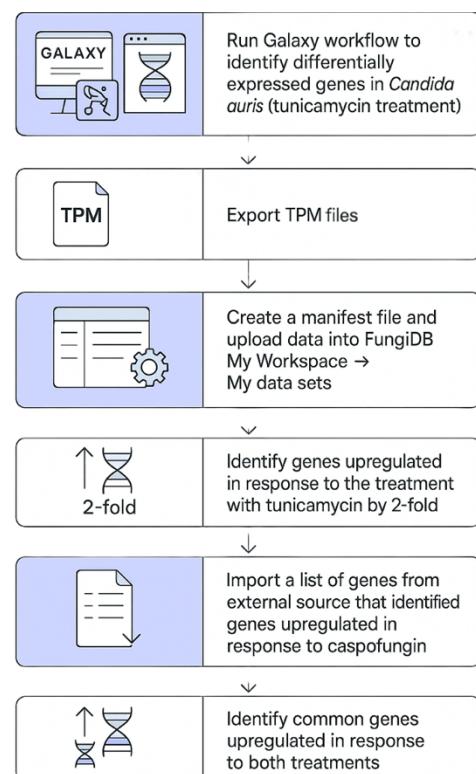
In this exercise, we will identify genes upregulated in *Candida auris* in response to different drug treatments (tunicamycin (Study 1) and caspofungin (Study 2)). The outline of the in-silico experiment is on the left. The data analysis for Study 1 was performed using the Galaxy workflow described above.

Identifying genes commonly upregulated in response to different drugs provides valuable insights into shared stress response pathways that may underlie antifungal resistance and pathogenic adaptability.

Caspofungin targets the fungal cell wall, while tunicamycin induces ER stress by inhibiting protein glycosylation.

Genes activated by both treatments may reveal core mechanisms of survival under therapeutic pressure. These genes could serve as potential targets for combination therapy, markers of stress adaptation, or indicators of virulence.

To do this, you will use the RNA-Seq analysis output from Galaxy and tools integrated under the private My Workspace section in FungiDB.



#### A. Upload Study 1 data to FungiDB - Create a zip file collection for the “Normalized RNA-Seq” analysis tool.

- **Download TPM files for control, low (low\_c) and high (high\_c) drug concentration samples.** The files will be posted on the LMS. Look for the folder called Study1\_RNASeq\_B11221.
- **Create a manifest file (.txt document).**
  - The file should include 3 columns: sample name, file name (.txt file only) and strand information (Only 'unstranded' is currently supported.)

```

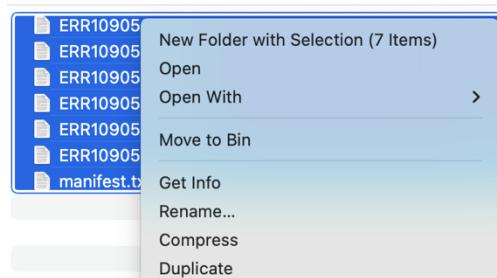
manifest.txt
1 control_rep1 ERR10905080.txt unstranded
2 control_rep2 ERR10905079.txt unstranded
3 low_c_rep1 ERR10905084.txt unstranded
4 low_c_rep2 ERR10905083.txt unstranded
5 high_c_rep1 ERR10905085.txt unstranded
6 high_c_rep2 ERR10905086.txt unstranded
  
```

The manifest.txt file contains a list of six entries, each with a sample name, a corresponding file name, and the strand information 'unstranded'. The entries are numbered 1 through 6.

Note: all files being uploaded must be listed in the manifest file

- **Compress all files.**

Select all files you want to upload, right-click and select “compress”.



Once you have compressed all files, you are ready to upload the data for the “Normalized RNA-Seq” data tool.

- **Upload your dataset to FungiDB.**
- Navigate to the My Workspace > My data sets  
<https://fungidb.org/fungidb/app/workspace/datasets/new/rnaseq>

- Click on the “New upload” tab,

My Data Sets

All [New upload](#) Help

Choose an upload type

Gene List  
Integrate your gene list in FungiDB.

bigWig  
Integrate your bigWig data in FungiDB.

Normalized RNA-Seq  
Integrate your Normalized RNA-Seq data in FungiDB.

- Click on the “Normalized RNA-Seq” button,

- Fill in information about the dataset

Note: Set the reference genome to *C. auris* B11221,

- Attach the compressed files from the previous step and click on the “Upload Data Set” button.

Upload My Normalized RNA-Seq Data Set

**⚠ Before uploading your dataset, please ensure your data is formatted according to the instructions listed in the “Help” tab.**

Name\*

Summary\*

Description

Reference Genome\*

Upload File\*  untitled folder.zip  
File must be less than 10GB

Upload URL

Additional instructions are here: <https://fungidb.org/fungidb/app/workspace/datasets/help>.

## B. Upload Study 2 gene list to FungiDB

- Navigate to the My Workspace > My data sets

<https://fungidb.org/fungidb/app/workspace/datasets/new/rnaseq>

- Click on the “New upload” tab,

My Data Sets

All [New upload](#) Help

- Click on the “Gene List” button,

Choose an upload type



Gene List

Integrate your gene list in FungiDB.



bigWig

Integrate your bigWig data in FungiDB.



Normalized RNA-Seq

Integrate your Normalized RNA-Seq data in FungiDB.

- Fill in information about the dataset

Note: Set the reference genome to *C. auris* B8441

- Attach the “Study2\_list\_B8441.txt” file you downloaded from LMS.
- Click on the “Upload Data Set” button.
- Check that the dataset was installed successfully

Status: This data set is installed and ready to use in FungiDB.

## C. This following section will be completed in class during Zoom breakout sessions.

In our breakout sessions on Zoom, we will develop search strategies utilizing Study 1 and 2 data.

- Explore the newly integrated and private Study 1 dataset using the FungiDB interface and features.
  - a. Navigate to the My Data Sets area in My Workspace.
  - b. Click on the *C. auris* RNA-Seq data.
  - c. Notice that the dataset has been installed in your workspace and it can now be explored via the “RNA-Seq user dataset (fold change) query. Click on the link to deploy the search.

[All My Data Sets](#)

### My Data Set: *Cauris B11221*

Status: This data set is installed and ready to use in FungiDB.

Available searches: [RNA-Seq user dataset \(fold change\)](#)

Owner: Me

Visibility: This data set is only visible to the owner and those they have shared it with.

Summary: RNA-Seq

Description: tunicamycin low and high concentrations

Created: 09/05/2025, 15:06:06

Data set size: 8.64 MB

ID: 1KsLsTfsh3D

Data type: RNA-seq (rnaseq 1.0)

- d. Find genes upregulated **1 or more fold** during the treatment with tunamycin (low concentration) when compared to control (reference sample). This search should be familiar to you as you have used to explore the integrated dataset in FungiDB before.

Cauris B11221

For the **Experiment** unstranded  
return protein coding Genes  
that are up-regulated  
with a **Fold change >=** 2  
between each gene's average expression value

in the following **Reference Samples**

- low\_c\_rep1  
 low\_c\_rep2  
 control\_rep1  
 control\_rep2  
 high\_c\_rep1  
 high\_c\_rep2

[select all](#) | [clear all](#)

and its average expression value

in the following **Comparison Samples**

- low\_c\_rep1  
 low\_c\_rep2  
 control\_rep1  
 control\_rep2  
 high\_c\_rep1  
 high\_c\_rep2

[select all](#) | [clear all](#)

- e. Name your strategy as “low\_c”

- f. Find genes that are **uniquely upregulated** in cells treated with high doses of tunicamycin

- i. Duplicate your original strategy first and modify it to use the high\_c data,

that are up-regulated

with a **Fold change >=** 1

between each gene's average expression value

in the following **Reference Samples**

- low\_c\_rep1  
 low\_c\_rep2  
 control\_rep1  
 control\_rep2  
 high\_c\_rep1  
 high\_c\_rep2

[select all](#) | [clear all](#)

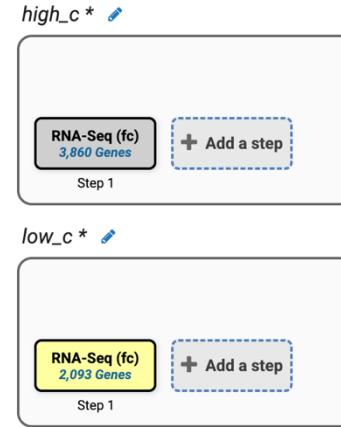
and its average expression value

in the following **Comparison Samples**

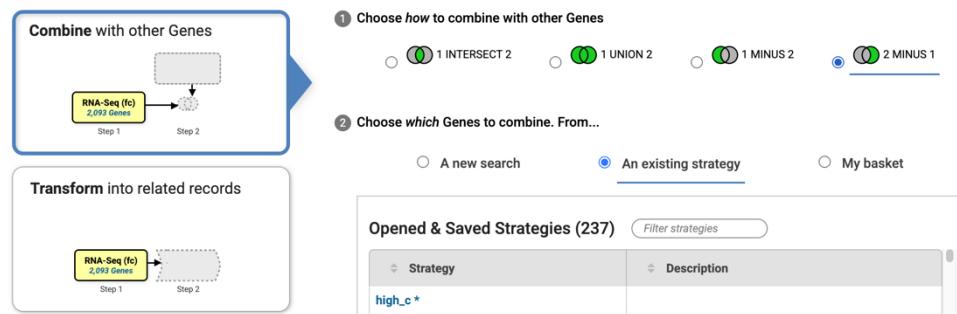
- low\_c\_rep1  
 low\_c\_rep2  
 control\_rep1  
 control\_rep2  
 high\_c\_rep1  
 high\_c\_rep2

[select all](#) | [clear all](#)

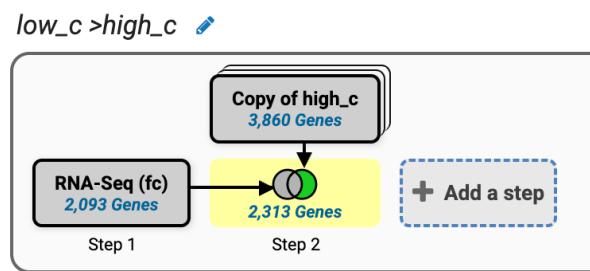
- ii. Name your new strategy as “high\_c”,



- iii. Return to the “low\_c” strategy, click on the “Add Step” and select the “high\_c” strategy and appropriate Boolean operator.



- g. How many genes did the search return?



Next, incorporate Study 2 gene list (caspofungin) into the current search strategy to look for upregulated genes in response to both antifungal drugs.

- Deploy the Gene List search.

- Navigate to the My Data Sets
- Click on the gene list dataset from Study 2
- Deploy the Gene List Dataset search

## My Data Set: Cauris B8441

**Status:**  This data set is installed and ready to use in FungiDB.

**Available searches:** [My Gene List Dataset](#)

**Owner:** Me

**Visibility:** This data set is only visible to the owner and those they have shared it with.

- Save the search strategy (e.g. caspofungin).

caspofungin \* 



- Revise your previous strategy to return genes upregulated in response to both low and high concentrations of tunicamycin treatments (use the Boolean operators to set the appropriate search parameter)

**Revise as a boolean operation**

 1 INTERSECT 2

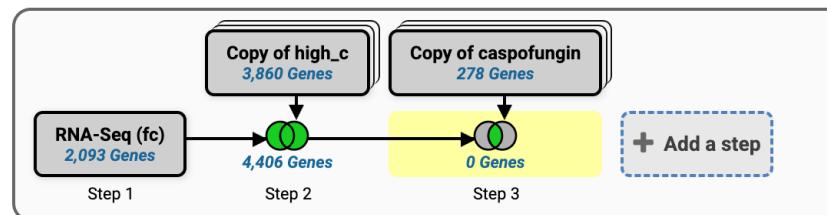
 1 UNION 2

 1 MINUS 2

 2 MINUS 1

- Click on the “Add a step” and cross-reference your results with caspofungin gene list.

low\_c >high\_c>Study 2 \* 



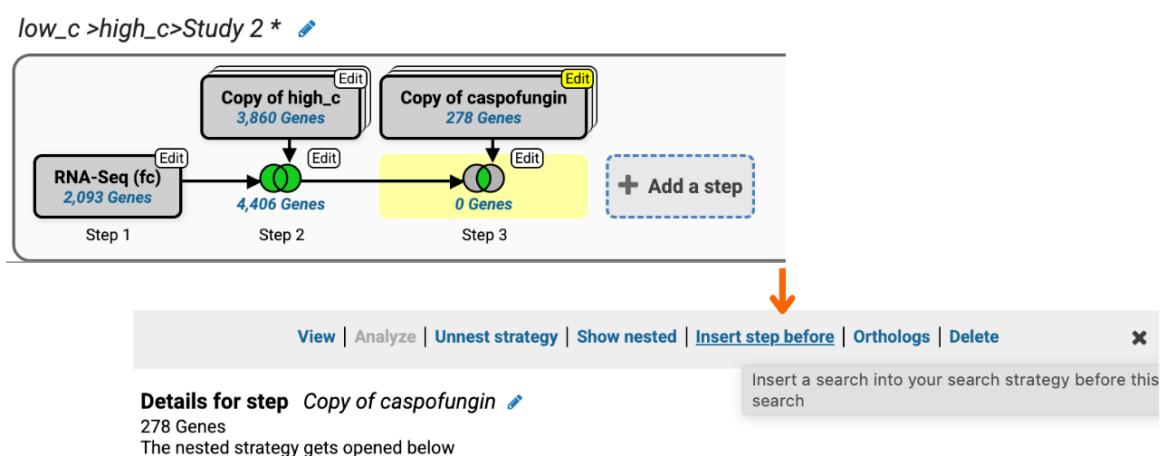
Do you think the results make sense?

Hint: Steps 1 & 2 are using gene IDs for the Ca B11221 genome, while Step 3 uses IDs for the Ca B8441.

How can you ensure that all steps are using the same reference genome?

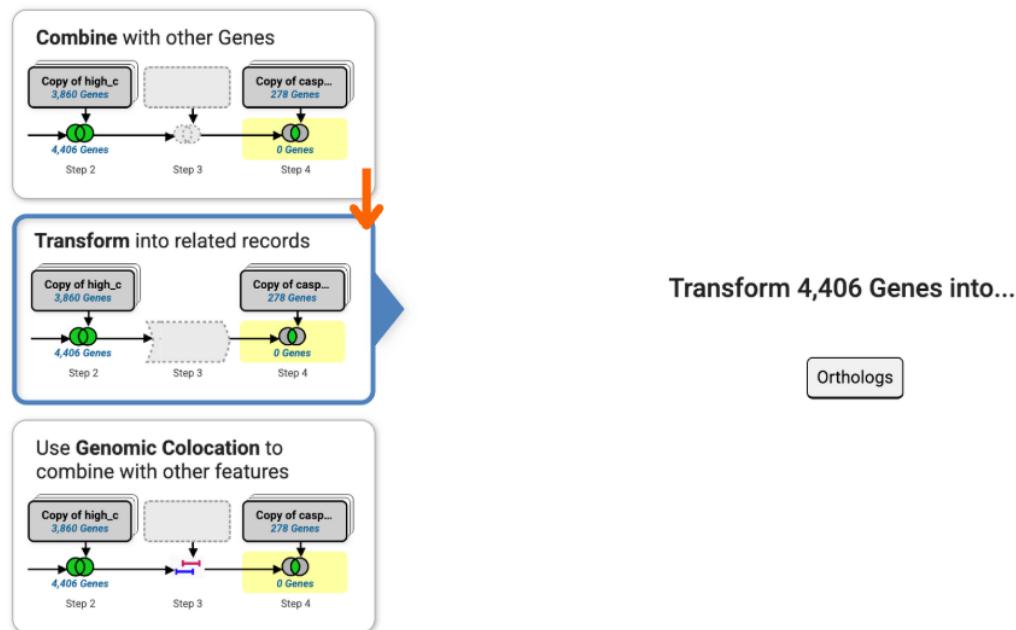
Hint: convert the results in Step 2 into B8441 Gene IDs to cross-reference both studies properly.

- Hover over the last step and click on the “Edit” option and click on the “Insert a step before”.



- Select the “Transform into related records” > “Orthologs” option.

[←](#) Add a step to your search strategy [?](#)

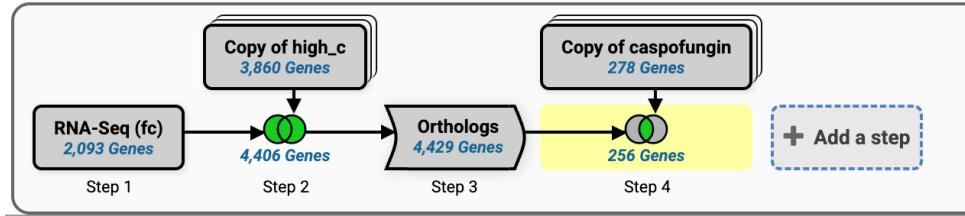


- Choose the reference genome – B8441 and click the “Run Step” button.

The screenshot shows the search configuration interface with the following sections:

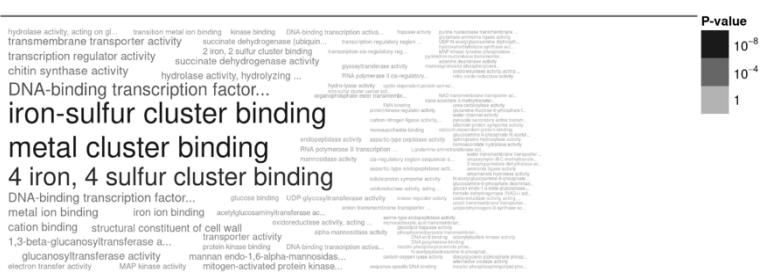
- Configure Search**: Includes "Reset values to default", "Learn More", and "View Data Sets Used".
- Organism**: Shows "1 selected, out of 274" and a dropdown menu with "B8441" selected. The dropdown menu includes:
  - Fungi
  - Ascomycota
  - Saccharomycetes
  - Metschnikowiaeae
  - [Candida] auris
  - Candida auris strain B8441 [Reference]** (with a red arrow pointing to it)
- Syntenic Orthologs Only?**: A dropdown menu set to "no".
- Run Step**: A button at the bottom right.

*low\_c > high\_c > Study 2 \** 



You can explore this data further in several ways, including GO Enrichment analysis, AlphaFold prediction data for hypothetical proteins, signal peptide or transmembrane searches, and more!

- Perform a GO enrichment analysis (Molecular function).



The enrichment of genes with functions such as metal cluster binding, DNA-binding transcription factor activity, chitin synthase activity, and transmembrane transporter activity in

response to tunicamycin and caspofungin suggests that *Candida auris* activates a coordinated stress response in response to antifungal drugs. This possibly includes transcriptional reprogramming to regulate stress-adaptive genes, metabolic adjustments involving iron-sulfur proteins and metal ion binding, and cell wall remodeling through increased chitin synthesis to compensate for antifungal-induced damage. Upregulation of transporters and ion-binding proteins further supports adaptation by facilitating nutrient uptake, detoxification, and maintenance of cellular homeostasis under drug-induced stress.

## Learning objectives:

- Examine gene models in JBrowse
- Assess gene models based on RNA-Seq or other types of data (e.g. intron evidence).
- Determine if a gene model is accurate or if alternate models are possible
- 

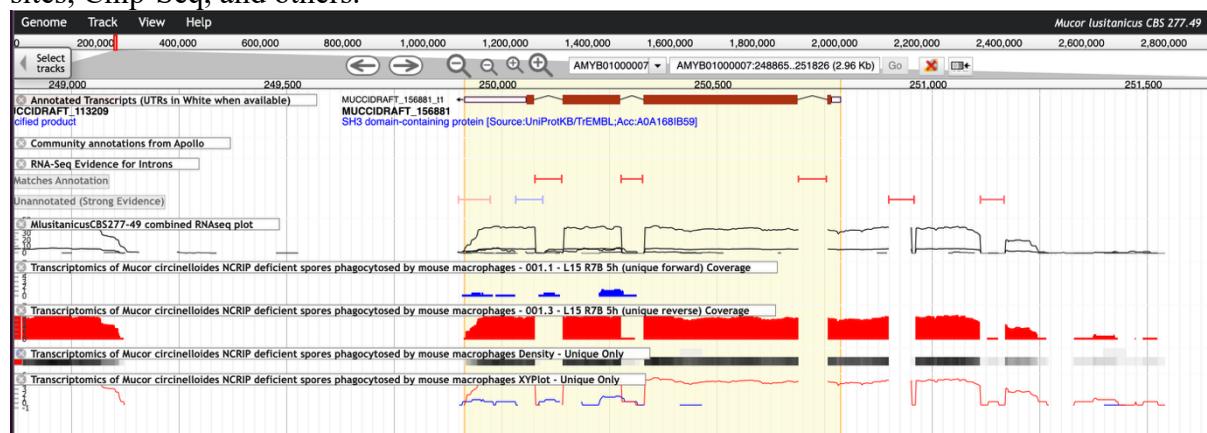
In previous exercises, you spent some time learning about gene pages and examining genes in the context of the JBrowse genome browser. It is important to recognize that gene models (structural annotation) are often open to interpretation, especially with respect to:

- transcript initiation and termination sites (5' and 3' untranslated regions, or UTRs)
- alternative processing events ... if you sequence deep enough, virtually *all* genes (in organisms that process transcripts) display alternative splicing, even for single exon genes.
- the potential significance of non-coding RNAs

Even actively curated genomes for well-established model organisms do not fully reflect all available knowledge about stage-specific splicing, as new information is constantly emerging! In addition, many gene models were computationally derived using methods that may not have relied on experimental evidence supporting intron/exon boundaries (e.g., RNA-Seq data).

In this exercise, we will explore several lines of evidence to interpret gene models and assess their accuracy and completeness. You will apply your newfound skills to examine a couple of genes and discuss your findings within the group.

The screenshot below shows a sample of data tracks that can be turned on in JBrowse. Stranded RNA-Seq data and RNA-Seq evidence for introns would be two useful tracks to begin with when evaluating gene models for the correct annotation of exons, introns, and UTRs. Depending on the species, other data types may be available as well - transcript start sites, Chip-Seq, and others.



- Take a look at several genes from the list below and activate several tracks in JBrowse that can help you evaluate gene models.
- Do you agree with the current annotated model? Would you have any modifications?

*Aspergillus nidulans* FGSC A4

AN10121

AN4483

AN8437

AN11226

AN12338

*Aspergillus fumigatus* Af293

Afu7g04610

Afu7g05330

Afu7g03670

Afu7g03700

Afu8g04280

Afu8g04320

Afu8g04350

Afu8g04420

*Neurospora crassa* OR74A

NCU05356

NCU06817

NCU06787

NCU01605

*Mucor lusitanicus* CBS 277.49

MUCCIDRAFT\_156881

MUCCIDRAFT\_127064

MUCCIDRAFT\_149865

MUCCIDRAFT\_149755

MUCCIDRAFT\_150170